Universidade Federal de Santa Catarina
Programa de Pós-Graduação em Engenharia de Produção

DESCOBERTA DE CONHECIMENTO RELEVANTE EM BANCO DE DADOS SOBRE CIÊNCIA E TECNOLOGIA

Wesley Romão

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para a obtenção do título de Doutor em Engenharia de Produção

Orientador: Prof. Dr. Roberto C. S. Pacheco

Co-orientador: Prof. PhD. Alex Alves Freitas

Florianópolis, SC.
Fevereiro, 2002.

Wesley Romão

DESCOBERTA DE CONHECIMENTO RELEVANTE EM BANCO DE DADOS SOBRE CIÊNCIA E TECNOLOGIA

Esta tese foi julgada e aprovada para a obtenção do título de
Doutor em Engenharia de Produção
no Programa de Pós-Graduação em Engenharia de Produção
da Universidade Federal de Santa Catarina

Florianópolis, 26 de Fevereiro de 2002.

Prof. Ricardo Miranda Barcia, Ph.D.
Coordenador do Curso

BANCA EXAMINADORA

Prof. Roberto C. S. Pacheco, Dr.
Orientador

Prof. Alex Alves Freitas, Ph.D.
Co-orientador e Membro externo

Prof. Aran Tcholakian Morales, Dr.

Prof. Júlio César Nievola, Dr.
Membro externo

Prof. Alejandro Martins, Dr.

Prof. José Leomar Todesco, Dr.

Agradecimentos

À Rosicler, minha querida esposa, pelo imenso e constante apoio que viabilizou a realização deste trabalho, e às nossas filhas gêmeas Larissa e Isabela, nascidas durante o período do doutorado, pela dupla alegria proporcionada.

Ao Prof. PhD. Ricardo Miranda Barcia, coordenador do Programa de Pós-Graduação em Engenharia de Produção, por ter concedido a oportunidade de ingressar neste programa.

Ao meu orientador, Prof. Dr. Roberto Carlos dos Santos Pacheco, pelo seu acompanhamento, entusiasmo e competência que sempre transmitiram segurança e tranqüilidade.

Ao Prof. PhD. Alex Alves Freitas, co-orientador externo, pelo seu conhecimento transmitido sobre o tema, o tempo e dedicação dispensados nas revisões e reuniões.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pelo ininterrupto apoio financeiro.

Aos professores colegas de trabalho no Departamento de Informática da Universidade Estadual de Maringá - UEM, pelo apoio concedido.

A Carlos Pittaluga, amizade conquistada na fase de realização das disciplinas, pelo incentivo, companheirismo e ajuda na escolha do tema.

Ao Prof. Dr. José Mazzucco Jr. maior amizade conquistada em Florianópolis juntamente com sua família que nos acolheu durante todos estes anos.

Ao Prof. Dr. Gilberto Cezar Pavanelli, Pró-Reitor de Pesquisa e Pós-Graduação da UEM, ao Prof. Dr. Flavio Bortolozzi, Pró-Reitor de Pesquisa e Pós-Graduação da PUC-PR, e o Prof. Dr. Manoel Camillo Penna Neto, Coordenador de Pesquisas também da PUC-PR, pelas entrevistas concedidas.

À minha mãe que, apesar de estar vivendo uma sobrevida em seus quase noventa anos, deixou sua marca de amor que me ajuda a superar todo medo.

À Dirce e Deir, que me ensinaram desde pequeno o bom caminho que deveria andar e até hoje me apoiam com amor e companheirismo.

Aos meus sobrinhos: Edemilson e Nivaldo, pelo constante incentivo e orações.

Aos amigos, Cleber Fuentes e Suely Maldonado, pelas constantes orações.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho.

SUMÁRIO

Lista de Figuras

Pág.

Lista de Quadros

Pág.

Lista de Tabelas

Pág.

Siglas

AE – Algoritmos Evolucionários

AG – Algoritmos Genéticos

AGD – Algoritmo Genético para Descoberta de Regras Difusas

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CCT – Conselho Nacional de C&T

CGEE – Centro de Gestão e Estudos Estratégicos

CNEN – Comissão de Energia Atômica

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

C&T – Ciência e Tecnologia

CRM – Customer Relationship Management

CS – Classifier Systems

EMBRAPA – Empresa Brasileira de Pesquisas Agropecuárias

FAP – Fundação de Amparo à Pesquisa

FAPERGS – Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul

FAPESP – Fundação de Amparo a Pesquisa do Estado de São Paulo

FINEP – Agência Financiadora de Estudos e Projetos

FIOCRUZ – Fundação Oswaldo Cruz

FUNCITEC – Fundação de Ciência e Tecnologia

GOCNAE – Grupo de Organização da Comissão Nacional de Atividades Espaciais

IA – Inteligência Artificial

IG – Impressões Gerais

INPA – Instituto Nacional de Pesquisa da Amazônia

INPE – Instituto Nacional de Pesquisas Espaciais

KDD – Knowledge Discovery in Databases

LS – Learning System

MCT – Ministério da Ciência e Tecnologia

MD – Mineração de Dados

MEC – Ministério da Educação

OCDE – Organização para a Cooperação e o Desenvolvimento Econômico

P&D – Pesquisa e Desenvolvimento

PIB – Produto Interno Bruto

PG – Programação Genética

PLI – Programação Lógica Indutiva

PPA – Plano Plurianual

PPGEP – Programa de Pós-Graduação em Engenharia de Produção

PUC-PR – Pontifícia Universidade Católica do Paraná

RNA – Redes Neurais Artificiais

SBPC – Sociedade Brasileira para o Progresso da Ciência

SCI – Science Citation Index

SCT – Secretaria de C&T

SGBD – Sistemas Gerenciadores de Banco de Dados

UFSC – Universidade Federal de Santa Catarina

UEM – Universidade Estadual de Maringá

WEKA - Waikato Environment for Knowledge Analysis

RESUMO

Com a estruturação das fundações de amparo a pesquisa da região sul do Brasil (e.g., Fundação Araucária no Paraná), a adequada gestão do fomento à ciência e tecnologia (C&T) nos estados do sul vem se tornando um assunto de interesse nas definições das políticas tecnológicas. No cenário nacional, as agências federais de fomento a C&T possuem bancos de dados, contendo informações relevantes à tomada de decisão em C&T, que abre perspectivas para a descoberta de conhecimento oculto, potencialmente relevante ao planejamento de C&T.

O processo geral de descoberta de conhecimento em banco de dados é composto por diversas etapas destacando-se a etapa de Mineração de Dados (MD). Existem diversas tarefas de MD, mas a tarefa de classificação é a mais conhecida e pode ser realizada por algoritmos convencionais (*e.g.*, estatísticos) ou por métodos de inteligência artificial (*e.g.*, redes neurais, algoritmos evolucionários, etc.). Nesta tese o objetivo é resolver uma generalização da tarefa de classificação, conhecida como modelagem de dependência, para extrair conhecimento na forma de regras de previsão que seja surpreendente no contexto de C&T.

Propõe-se um novo algoritmo genético capaz de descobrir regras de previsão difusas. O objetivo é descobrir conhecimento relevante (surpreendente, novo) oculto em banco de dados de C&T adaptando-se uma técnica que usa Impressões Gerais (conhecimento subjetivo) fornecidas pelo usuário, que é um assunto pouco explorado, além de se utilizar lógica difusa, para validação das regras, e termos lingüísticos difusos para representar as variáveis contínuas.

Um protótipo foi implementado, tendo como estudo de caso a região sul do Brasil e banco de dados fornecido pelo CNPq. A qualidade do protótipo foi avaliada diante dos dados devidamente preparados (pré-processamento) tendo como referencial o algoritmo J4.8, uma versão do algoritmo C4.5. O protótipo apresentou eficiência aproximadamente equivalente ao J4.8 quanto à taxa de acerto, mas forneceu conhecimento mais compreensível, fruto do uso de regras com poucas condições e termos lingüísticos difusos.

Os resultados experimentais, em forma de regras de produção difusas, foram apresentados a usuários potenciais, através de entrevistas, que avaliaram o conhecimento novo obtido. Os entrevistados classificaram 45% das regras como muito relevantes, 31% de médio interesse e 24% de baixo interesse. A avaliação subjetiva, considerada satisfatória, foi próxima do grau de interesse fornecido pelo protótipo, calculado como contradição às impressões gerais fornecidas pelos usuários, confirmando a utilidade e relevância do novo algoritmo implementado.

ABSTRACT

With the creation and expansion of the research-support foundations in the states of the South of Brazil (e.g. Fundação Araucária at Paraná state), Science and Technology (S&T) management has become a relevant point for the determination of a technology policy. There are some databases, containing data derived from operational processes in federal research-support agencies, which contain information relevant to S&T decision-makers.

The Knowledge Discovery in Databases (KDD) process consists of many stages, out of which Data Mining (DM) is the main one. Among many DM tasks, the literature emphasizes classification, which can be addressed by conventional algorithms (e.g., statistics) or by artificial intelligence techniques (e.g., neural networks, evolutionary algorithms, etc.). The goal of this thesis is to solve a generalization of classification task, called dependence modeling, in order to extract knowledge in the form of prediction rules which are interesting in the context of S&T management.

This thesis shows the viability of designing a genetic-fuzzy algorithm (based on the combination of genetic algorithms and fuzzy sets) to solve the dependence-modeling task and discover interesting (surprising, new) knowledge in S&T databases. This is done by adapting a technique little exploited in the literature, which is based on user-defined general impressions (subjective knowledge).

The south region of Brazil is the case study addressed in this thesis. The databases of the CNPq ("National Council of Scientific and Technological Development") were used to test a prototype genetic-fuzzy system to discover interesting rules. This prototype was evaluated by comparing it with the J4.8 algorithm, a modified version of the well-known C4.5 algorithm. The predictive accuracy obtained by the prototype was similar to the one obtained by J4.8, but in general the former discovers rules with fewer conditions, in addition to working with natural linguistic terms, which leads to the discovery of more comprehensible knowledge.

The rules discovered by the system were shown to potential users in interviews, in order to evaluate those rules. The users considered 45% of the rules as having a high degree of interestingness, 31% as having a medium degree of interestingness and 24% as having a low degree of interestingness.

1 INTRODUÇÃO

A Engenharia de Produção é uma área de pesquisa multidisciplinar que tradicionalmente envolve disciplinas da produção civil, elétrica e mecânica. Ultimamente tem envolvido a disciplina de produção de software, a qual possui estreita relação com as demais disciplinas das engenharias.

Problemas de produção existem em qualquer empresa pública ou privada. Deve-se reconhecer que nas instituições de pesquisa também existem dificuldades que exigem a aplicação de técnicas para solução de problemas e otimização de sua produção. Um problema relevante nos tempos atuais é a gestão eficiente de agências de fomento a C&T e das instituições responsáveis pelo desenvolvimento científico e tecnológico da nação.

A grande quantidade de informações nos bancos de dados informatizados destas instituições pode esconder conhecimentos valiosos e úteis para a tomada de decisão. O aumento no volume dos dados, associado à crescente demanda por conhecimento novo voltado para decisões estratégicas, tem provocado o interesse crescente em descobrir novos conhecimentos em banco de dados, em especial sobre a produção em C&T.

As limitações financeiras impostas ao setor de C&T, em especial no apoio à pesquisa, torna fundamental que os processos de tomada de decisão neste segmento sejam executados com base em conhecimento estratégico obtido a partir de uma base íntegra de dados.

Portanto, a formação de bancos de dados consistentes é o primeiro passo para se obter informações estratégicas para uma gestão eficiente nas agências de C&T. Neste segmento, bancos de dados contém informações sobre pesquisadores, grupos, projetos e instituições de pesquisa. O desafio é transformar estas informações em subsídios para esta mesma comunidade.

Alguns bancos de dados sobre C&T já existem no Brasil. O CNPq possui a Plataforma Lattes, que inclui o Diretório dos Grupos de Pesquisa e o banco de currículos de pesquisadores, a Plataforma Coleta da CAPES, e outros que estão se consolidando (Capítulo 2).

A existência destes bancos de dados em C&T permite conhecer como se

organiza o parque científico e tecnológico nacional, o que é estratégico para uma gestão segura do setor, permitindo, inclusive, traçar diretrizes políticas realistas. Neste contexto, o desafio que se apresenta pode ser simplificado como a resolução de duas questões básicas: *Como extrair conhecimento destes dados?* e; Como obter conhecimento que seja estratégico para tomada de decisões?

É importante distinguir entre as duas questões acima para estabelecer critérios que irão nortear a definição de técnicas adequadas ao objetivo traçado. Muitas técnicas já existentes permitem atender apenas a primeira questão, mas é necessário estabelecer técnicas que atendam a ambas.

O processo de extração de conhecimento a partir de dados é ilustrado pelo triângulo da Figura 1.

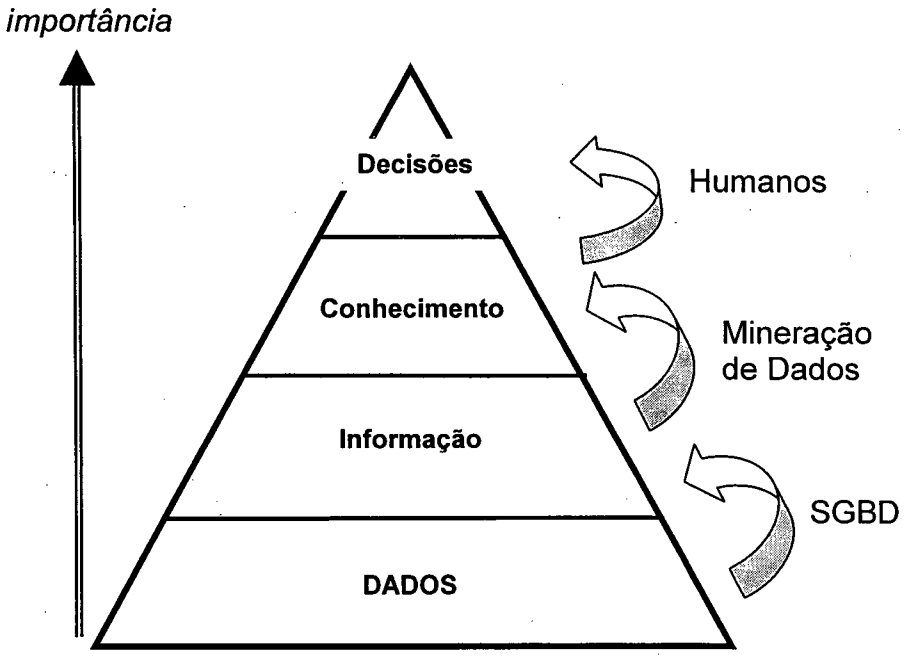


Figura 1: Obtenção de conhecimento para tomada de decisões.

Na base do triângulo estão os dados, os quais tomam o maior volume da memória do computador, e oferecem pouca utilidade estratégica na hora de se tomar decisões. A partir dos dados é possível obter muita informação através de aplicativos desenvolvidos para fins específicos ou através das ferramentas dos Sistemas Gerenciadores de Banco de Dados (SGBD) que exigem

conhecimento das mesmas por parte do analista para se obter o máximo proveito da montanha de dados disponíveis e em crescimento.

A partir das informações ou dos próprios dados é possível extrair um tipo de informação mais completa, o conhecimento, normalmente mais resumido e em menor quantidade, mas de maior inteligibilidade para se tomar decisões. Uma definição mais precisa sobre dados, informação e conhecimento encontra-se na seção 3.2.

Finalmente, no topo do triângulo da Figura 1, aparecem as decisões realizadas pelo homem com base no conhecimento obtido pelas ferramentas de Mineração de Dados (MD). A aplicação de algoritmos específicos deve garantir que o tipo e forma do conhecimento obtido estejam adequados ao processo de tomada de decisões rápidas e inteligentes.

O produto principal de qualquer ferramenta de apoio à decisão é o conhecimento que ela pode fornecer. Existem inúmeras técnicas (*e.g.*: algoritmo Apriori) capazes de extrair conhecimento em banco de dados, mas em geral este conhecimento ainda é de grande volume, dificultando a tomada de decisões. Para viabilizar decisões eficientes, devem ser implementadas ferramentas de apoio à tomada de decisão capazes de extrair conhecimento novo e surpreendente a partir de banco de dados.

A existência de bancos de dados sobre C&T já é por si só uma grande motivação para se desejar implementar ferramentas capazes de extrair conhecimento novo a partir destas bases de dados. Para implementar ferramentas como essas é necessário conhecer resultados de pesquisa de uma área conhecida como KDD (Knowledge Discovery in Database – Descoberta de Conhecimento em Banco de Dados).

KDD é descrito (Capítulo 3) como um processo genérico de descoberta de conhecimento que inclui pré-processamento dos dados, Mineração de Dados (MD) e pós-processamento do conhecimento obtido, onde Mineração de Dados é a etapa mais relevante neste contexto e exige a sua investigação em profundidade. MD pode ser empregada para resolver diversas tarefas (*e.g.*, associação, classificação, etc.).

Nesta pesquisa, o primeiro estudo empírico realizado envolveu a tarefa de

descoberta de regras de associação, ou seja, aplicou-se um algoritmo padrão de extração de regras de associação (Apriori) à base de dados do *Diretório dos Grupos de Pesquisa no Brasil* (CNPq, 1999), versão 3.0. Os resultados desse experimento (ver anexo A) forneceram muitas regras a respeito dos padrões e relacionamentos entre atributos existentes nos dados, mas não conhecimento novo para realizar planejamento em C&T (Romão *et al.*, 1999b).

Para descobrir conhecimento novo e relevante é necessário investigar técnicas voltadas à descoberta de regras de previsão. Para descobrir regras desse tipo geralmente é necessário resolver a tarefa de classificação ou modelagem de dependência – ou outra tarefa relacionada à aprendizagem de máquina - mas não a tarefa de associação padrão (Freitas, 2000a). Logo, no escopo desta tese é dada atenção especial às tarefas de classificação e modelagem de dependência, assuntos tratados no Capítulo 4.

Os algoritmos de MD em geral são projetados para descobrir conhecimento exato e compreensível. A compreensibilidade sintática é facilitada com o uso de algoritmos de indução (*e.g.*: árvores de decisão) que descobre regras do tipo:

SE alguns valores de atributos previsores ocorrem em um registro

ENTÃO prever o valor de algum atributo meta para aquele registro

Apesar de já existirem inúmeros algoritmos de mineração de dados disponíveis na literatura (Fayyad, 1996b; Romão *et al.*, 1999b; Gonçalves, 2000), um dos desafios está em adaptar estas técnicas tradicionais para serem viáveis diante dos bancos de dados de grande volume que existem na atualidade.

Muitas das técnicas tradicionais de mineração de dados (seção 4.2.2) têm sido aplicadas com sucesso e outras esbarram em limitações, tanto no desempenho como na qualidade do conhecimento gerado. Pesquisas recentes têm demonstrado que técnicas da área de IA, tais como Algoritmos Genéticos (Bojarczuk *et al.*, 2000; Freitas, 2000b; Dhar *et al.*, 2000; Kim *et al.*, 2000; Freitas, 2001b) e Conjuntos Difusos, podem ser utilizadas com sucesso (Delgado, 1999; Fertig, 1999; Voznika & Mendes, 2000).

Para realização da maioria das tarefas de MD, os algoritmos de indução de

regras incluem, como uma das suas principais etapas, um processo de busca. Nesta etapa aparece um desafio: a interação entre atributos. Para resolver este problema existe uma abordagem baseada em algoritmos genéticos (Romão *et al.,* 1999b), cuja principal virtude é a capacidade de considerar a interação entre os atributos envolvidos, além de realizar busca global no universo de valores possíveis destes atributos, evitando a convergência para máximo/mínimo local. Uma revisão sobre Algoritmos Genéticos (AG) é apresentada no Capítulo 5.

A área de algoritmos evolucionários (Freitas, 2001b), onde se incluem os Algoritmos Genéticos (AG's), vem despertando interesse crescente dos meios científicos e industriais devido, principalmente, ao fato dos computadores estarem se tornando cada vez mais velozes e com grande capacidade de armazenamento de informações, o que era uma restrição à aplicação destas técnicas no passado.

Entretanto, conforme escrito acima, não basta extrair todo e qualquer tipo de conhecimento. Para que o conhecimento obtido seja estratégico para tomada de decisão, ele deve ser correto, compreensível e também novo (surpreendente) para o usuário. Esta última característica implica na extração de *conhecimento relevante*[1] que acrescente algo de novo ao conhecimento do decisor. Para isto propõe-se uma abordagem baseada em impressões gerais (seção 4.5.4) do usuário sobre o domínio da aplicação (descrita no Capítulo 6).

Para atender a exigência de conhecimento correto (válido) pode-se utilizar um critério de validação das regras implicitamente dentro do algoritmo de mineração de dados, descrito na seção 6.8. A compreensão pode ser facilitada utilizando representação do conhecimento na forma de regras de produção com poucas condições e utilizando termos lingüísticos difusos. O principal desafio nesta tese é obter conhecimento que seja relevante, novo e surpreendente.

---

[1] A frase "*conhecimento relevante*" foi a melhor tradução encontrada por este autor para a expressão "interesting knowledge", empregada nas publicações em inglês, significando a importância do conhecimento extraído para o usuário.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo principal nesta tese é desenvolver um algoritmo para obter conhecimento relevante que auxilie gestores das agências a efetuar um planejamento que aumente a eficiência das instituições de C&T.

A abrangência desta tese contempla desde o estudo da área de aplicação (C&T), obtenção de dados reais e preparação destes dados (pré-processamento), até a escolha e adaptação de técnicas de inteligência artificial (AG e conjuntos difusos) para extração de conhecimento novo e relevante (mineração de dados), baseado no conhecimento prévio do usuário (impressões gerais), que seja relevante à tomada de decisão em C&T. O foco do algoritmo de MD é resolver a tarefa de modelagem de dependência, onde qualquer atributo relevante pode ser utilizado como um atributo meta a ser previsto, ou seja, diferentes regras podem ter diferentes atributos em seus conseqüentes.

Para alcançar este objetivo geral, surgiram os seguintes desafios:

- Conhecer a área de aplicação (C&T);
- Definir um estudo de caso baseado em dados reais sobre C&T;
- Descobrir regras com caráter preditivo;
- Utilizar conhecimento do usuário no processo de busca;
- Não usar pós-processamento para extrair as regras, ou seja, descobrir regras diretamente com o algoritmo de MD;
- Implementar um novo algoritmo híbrido;
- Descobrir conhecimento relevante para o usuário;

Portanto, esta tese propõe um algoritmo híbrido de mineração de dados, baseado na combinação de algoritmos genéticos e conjuntos difusos, para extrair conhecimento correto, compreensível e relevante à tomada de decisão em gestão de C&T na região sul do Brasil. O caráter "relevante" do conhecimento extraído é baseado em impressões gerais do usuário. Os

detalhes deste sistema são apresentados no Capítulo 6.

1.1.2 Objetivos Específicos

Para alcançar o objetivo geral, e resolver os desafios inerentes a ele, foi necessário realizar um levantamento bibliográfico tanto sobre o domínio da aplicação como sobre as técnicas de mineração de dados, efetuar implementação de um sistema de computador, obter resultados experimentais e realizar entrevistas.

Isto exigiu o estabelecimento dos seguintes objetivos específicos:

- Estudar a área de aplicação procurando identificar que tipo de conhecimento é estratégico para realizar planejamento em C&T;
- Obter dados reais sobre C&T;
- Preparar estes dados para viabilizar a aplicação de técnicas de mineração de dados;
- Escolher uma tarefa de mineração de dados adequada para solução do problema;
- Realizar um levantamento bibliográfico sobre os principais métodos de mineração de dados para solução desta tarefa e verificar suas limitações de aplicabilidade e qualidade do conhecimento gerado;
- Considerar a interação entre atributos no algoritmo de mineração de dados;
- Utilizar conjuntos difusos para evitar uma rígida discretização das variáveis contínuas;
- Realizar um levantamento sobre sistemas híbridos genético-difusos para descoberta de regras de previsão;
- Propor um algoritmo capaz de realizar a tarefa de modelagem de dependência e extrair conhecimento novo e relevante;
- implementar o algoritmo proposto e obter resultados experimentais (avaliar o classificador);
- Avaliar o interesse nas regras descobertas através de entrevistas com usuários potenciais.

1.2 Justificativas

Boa parte das técnicas de descoberta de conhecimento tem sido aplicada para explorar relações de compra e venda em empresas, visando aumentar a comercialização de produtos. No entanto, mais recentemente, tem se intensificado sua aplicação no apoio a gestores de diversos segmentos em que há disponibilidade de banco de dados.

A introdução do código de barras na maioria dos produtos, a popularização e redução dos preços dos computadores e a automação de muitas empresas e estatais têm resultado em grandes volumes de dados, tornando insuficientes os métodos tradicionais de análise.

Existe uma necessidade significante por uma nova geração de técnicas e ferramentas com habilidades para assessorar humanos a analisar montanhas de dados de forma inteligente e automática através do fornecimento de conhecimento resumido e estratégico (Fayyad *et al.,* 1996a).

Muitas empresas e instituições governamentais estão iniciando a exploração de seus dados, através da construção de Data Warehouse (Dias *et al.*, 1998) e ferramentas de extração de conhecimento, com o objetivo de reduzir custos e otimizar a qualidade de seus produtos e serviços.

Técnicas de mineração de dados podem auxiliar na descoberta automática ou semi-automática de conhecimento que, via de regra, permite a previsão de valores de atributos com maior probabilidade de acerto. Não se tratam de previsões do tipo mágico ou aleatório, mas de previsões seguras baseadas em informações reais, as quais estão disponíveis nos bancos de dados das agências, mas que podem estar sendo desperdiçadas.

Estas técnicas podem ser úteis para responder perguntas do usuário (e encontrar outras respostas para perguntas que ainda não foram formuladas e que sejam relevantes) e auxiliar no processo de decisão em C&T.

Existem muitas técnicas capazes de descobrir conhecimento correto e compreensível, mas a descoberta de conhecimento realmente relevante permanece um desafio em MD. Há algumas abordagens subjetivas para descoberta de conhecimento relevante (Silberschatz & Tuzhilin, 1996; Liu *et al.*,1997), mas em geral na forma de pós-processamento, ou seja, algum

algoritmo de MD deve ser executado primeiro, gerando um conjunto com grande número de regras, para depois outro algoritmo extrair as regras relevantes a partir deste conjunto.

Nesta tese propõe-se um novo algoritmo genético, onde os indivíduos representam regras com termos lingüísticos difusos, projetado para descobrir regras de previsão surpreendentes, onde o processo de busca incorpora uma preferência por regras relevantes na função de Fitness (a função de avaliação de regras, como será visto posteriormente). Isto possibilita a geração direta de regras de interesse, evitando a realização de um pós-processamento.

O algoritmo proposto incorpora comparação difusa das regras com os registros no banco de dados, para validação das regras, e comparação das regras com impressões gerais do usuário sobre o domínio da aplicação como forma de guiar a busca apenas por regras de interesse do usuário.

Para comprovar a eficácia do algoritmo proposto efetuou-se a sua implementação para extrair conhecimento de dados reais sobre C&T.

1.3 Organização do Trabalho

Esta tese está organizada em oito capítulos, incluindo o presente. Estes capítulos objetivam apresentar uma revisão bibliográfica sobre KDD, Mineração de Dados, algoritmos genéticos, alguns tópicos sobre conjuntos difusos, e algoritmos genéticos em combinação híbrida com conjuntos difusos para resolver a tarefa de modelagem de dependência. Além disso, apresenta-se a especificação de um novo algoritmo genético para extração de regras, alguns detalhes de sua implementação e resultados de experimentos computacionais. No final são descritos resultados de avaliação das regras realizada através de entrevistas e as conclusões.

Em suma, a organização da tese apresenta-se da seguinte maneira:

Capítulo 2: Ciência & Tecnologia no Brasil - é apresentada uma pequena introdução sobre a ciência e tecnologia no Brasil, demonstrando a sua evolução. É dado destaque aos indicadores de C&T como um meio para visualizar a realidade, incluindo a análise da importância dos indicadores regionais;

Capítulo 3: Descoberta de Conhecimento – apresenta uma visão geral sobre o processo de descoberta de conhecimento, identificando cada uma de suas etapas, enfatizando a etapa de mineração de dados;

Capítulo 4: *Descoberta de Regras de Previsão* - é apresentada uma revisão sobre algoritmos de classificação e algumas considerações sobre a tarefa de modelagem de dependência, seguida da descrição de algumas técnicas subjetivas para descoberta de conhecimento relevante;

Capítulo 5: Sistemas Híbridos Genético-Difusos – apresenta-se uma revisão geral sobre AG's que descobrem regras de previsão, destacando a forma de codificação do indivíduo, os operadores e a função de aptidão que sejam adequados para a aplicação. É dado enfoque especial à combinação híbrida de AG e conjuntos difusos na tarefa de classificação;

Capítulo 6: O Algoritmo Genético Proposto – Propõe-se um algoritmo genético, híbrido com conjuntos difusos, para descoberta de conhecimento "relevante" baseado em impressões gerais do usuário, com ênfase na extração de conhecimento novo (surpreendente) à tomada de decisão em C&T a partir de bancos de dados do CNPq, tendo como estudo de caso a região sul do país;

Capítulo 7: *Experimentos e Resultados* – Neste capítulo apresenta-se o seguinte: uma descrição geral dos experimentos realizados, a metodologia utilizada na seleção de atributos; a "fuzzificação" dos atributos contínuos; as impressões gerais fornecidas pelos usuários; resultados completos de dois experimentos incluindo, além das regras descobertas, a taxa de acerto, em cada classe, do algoritmo novo comparada com a taxa de acerto do algoritmo J4.8; uma análise comparativa entre os resultados dos dois experimentos e resultados de avaliação subjetiva das regras (obtidos através de entrevistas);

Capítulo 8: *Conclusões e Sugestões* – Apresentam-se análise e discussão dos resultados alcançados durante a elaboração desta tese

seguidas de sugestões para pesquisas futuras.

Anexos: Nesta seção, dividida em três partes, primeiramente apresenta-se uma revisão sobre a tarefa de associação e resultados experimentais (preliminares) obtidos aplicando o algoritmo Apriori sobre dados do Diretório 3.0. Em seguida apresentam-se resultados de dois experimentos realizados com o algoritmo genético proposto. Finalmente, na última parte, apresenta-se uma avaliação subjetiva das regras obtida através de entrevistas com três usuários.

2 CIÊNCIA E TECNOLOGIA NO BRASIL

2.1 Introdução

A Ciência pode ser definida como um sistema de produção de informação, em particular informação na forma de publicações (e patentes), considerando publicação como qualquer "informação registrada em formatos permanentes e disponíveis para o uso comum" (Spinak, 1998).

A Tecnologia é definida como a aplicação de conhecimento da ciência à produção em geral. Na perspectiva de Szmrecsányi (1987), ambas, ciência e tecnologia se tornaram produtos corriqueiros da economia e da sociedade, frutos de uma determinada divisão do trabalho, dentro da qual os cientistas e tecnólogos desenvolvem suas atividades específicas em estreita e permanente interação com outros agentes produtivos da agricultura, da indústria e dos serviços.

Depois do pós-guerra, em muitos países tomou-se consciência da importância da atividade científica para impulsionar a produção de bens e serviços levando as sociedades industriais a aumentarem as fatias destinadas à pesquisa nas universidades e nas empresas industriais (Brisolla, 1998).

No Brasil, segundo Brisolla (1998), as políticas em Ciência e Tecnologia (C&T) vêm sendo relegadas ao segundo plano por sucessivos governos. Mas o que seriam políticas?

> *"Políticas são orientações de ordem geral e têm como função subsidiar os administradores na tomada de decisão. As políticas constituem o feixe de idéias que corporificam a decisão. Representam a síntese das grandes opções. O elenco de políticas é o próprio enunciado da decisão, desdobrando-se em níveis de planejamento para ações"* (Silveira, p.21, 1996).

Portanto, quais seriam os planos para a C&T no Brasil?

No âmbito federal, já existe uma definição da política de C&T. O entrave

está na execução destas políticas, em função de sucessivos problemas orçamentários e da dificuldade em se conseguir executar o planejado.

Há macro diagnósticos precisos sobre os setores estratégicos e carentes de investimentos no país (Coutinho & Ferraz, 1994) e mesmo sobre a situação da C&T nacional (MCT, 1994, 1996). Porém, entre o ato de planejar e o ato de executar há uma dicotomia muito grande. Basta uma rápida inspeção nos orçamentos realizados pelo CNPq e pela CAPES para constatar o descompasso entre planejamento e execução.

No âmbito estadual, tornaram-se comuns cortes orçamentários, extinção de secretarias de C&T, quebra de compromissos no repasse de recursos garantidos por leis estaduais, etc. Entretanto, não há nação desenvolvida que tenha negligenciado o investimento em C&T. Isto é ponto pacífico e a literatura é vasta em enumerar exemplos de sucesso econômico e social que foram precedidos pelo progresso científico e tecnológico (Schwartzman *et al.*, 1996).

É desejável expandir os investimentos em C&T, principalmente em Pesquisa e Desenvolvimento (P&D). Mas para evitar que os escassos recursos destinados a C&T sejam utilizados em descompasso com o que é planejado e evitar o desperdício, é necessário conhecer em detalhes a realidade da infra-estrutura e do potencial de pesquisa do país. O governo deve realizar melhorias na qualidade do gasto público, efetuar a modernização gerencial e enfrentar o desafio de realizar mais com menos recursos.

Fazer mais com menos é um dos principais desafios da atualidade para nações em desenvolvimento como o Brasil. Conforme consta no Plano Plurianual do governo (PPA 2000-2003, 1999), este objetivo busca condições para que o Estado cumpra suas funções com maior racionalização na alocação de recursos com base em: gerenciamento, definição de prioridades, atividades estratégicas e de coordenação integrada das ações governamentais.

A própria atividade em C&T faz parte do PPA como diretriz estratégica, intitulada "*ampliar a capacidade de inovação*", resumida nos seguintes parágrafos:

> *"Estimular a expansão das atividades empresariais em Pesquisa e Desenvolvimento, em articulação com órgãos públicos de apoio ao setor,*

universidades e laboratórios".

"Orientar o modelo de gestão das instituições de pesquisa e desenvolvimento e das universidades para melhorar o desempenho quanto ao atendimento da demanda tecnológica" (PPA 2000-2003, 1999).

Fazendo parte do PPA, estas diretrizes determinam (ou pelo menos viabilizam) que a atividade em C&T seja organizada por uma política global coerente e de abrangência nacional.

Apesar das dificuldades citadas anteriormente, o sistema brasileiro de Ciência e Tecnologia tem experimentado avanços significativos. A consolidação de bases de dados importantes contendo informações sobre o seu universo de pesquisa e de seus recursos humanos envolvidos com pesquisa revela a magnitude e crescimento da C&T no Brasil. São destaques o *Diretório dos Grupos de Pesquisa no Brasil* (CNPq, 1999a) e o *Sistema de Currículos Lattes* (CNPq, 1999b), desenvolvidos no âmbito do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e o Sistema Coleta da CAPES sobre a pós-graduação brasileira (CAPES, 1999). A partir deles é possível traçar um panorama bastante confiável da C&T das Universidades de qualquer região do Brasil.

Portanto, muitos avanços já foram alcançados. O sistema LATTES, por exemplo, já está integrando os dados curriculares dos pesquisadores de várias agências (FINEP, CAPES, CNPq) e deverá ser adotado por outras agências. Esta integração facilita e viabiliza a obtenção dos dados necessários para o planejamento de ações em C&T.

Neste capítulo é feita uma breve revisão bibliográfica sobre C&T no Brasil como subsídio à aplicação de um algoritmo proposto para extração de conhecimento novo em C&T.

Na seção 2.2 apresenta-se um resumo da evolução do sistema de C&T brasileiro seguido, na seção 2.3, de uma análise sobre a importância de se realizar planejamento neste contexto.

Na seção 2.4 descreve-se uma das principais fontes de informação utilizada para se realizar planejamento, os indicadores. Na seção 2.5 destacam-se os

indicadores regionais e mostra-se um seguimento que carece de informações de caráter regional, as agências estaduais de fomento à C&T.

Na seção 2.6 descreve-se a principal fonte de informações sobre C&T disponível no Brasil, a Plataforma Lattes.

Finalmente, na seção 2.7, apresenta-se as motivações para um estudo de caso visando utilizar os dados sobre C&T disponíveis para obter novos conhecimentos úteis ao planejamento e gestão de C&T na região sul do Brasil, seguido das considerações finais deste capítulo na seção 2.8.

2.2 Evolução do Sistema de C&T no Brasil

O sistema de C&T brasileiro se desenvolveu pautado principalmente na fundação de diversos conselhos e órgãos públicos de custeio à pesquisa. Uma forma de ver a evolução da C&T no Brasil é através da retrospectiva cronológica de criação destes órgãos, conforme segue:

- 1916 → fundação da Academia Brasileira de Ciências.
- 1949 → fundação da SBPC (Sociedade Brasileira para o Progresso da Ciência).
- 1951 → criação do Conselho Nacional de Pesquisas (CNPq) destinado a financiar projetos de pesquisa individuais através do fornecimento de bolsas de estudos e subsídios, cujo nome foi transformado posteriormente para *Conselho Nacional de Desenvolvimento Científico e Tecnológico*; Neste mesmo ano foi criada a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) como Campanha e instituída como Fundação em 1992.
- 1954 → criação do INPA (Instituto Nacional de Pesquisa da Amazônia).
- 1956 → criação do CNEN (Comissão de Energia Atômica), uma autarquia federal vinculada ao Ministério de Ciência e Tecnologia, para estabelecer normas e regulamentos em radioproteção e segurança nuclear, licenciar, fiscalizar e controlar a atividade nuclear no Brasil.
- 1961 → criação do Grupo de Organização da Comissão Nacional de

Atividades Espaciais (GOCNAE), oficializado como INPE (Instituto Nacional de Pesquisas Espaciais) em 1971.

- 1970 → criação da FINEP (Agência Financiadora de Estudos e Projetos). A década de 70 foi marcada por uma onda de dificuldades orçamentárias decorrentes do choque do petróleo e aumento das taxas de juros no mercado internacional. Com isto, houve queda dos recursos destinados ao fomento de C&T em cerca de 65%.
- 1985 → criação do MCT (Ministério de Ciência e Tecnologia) que passa a comandar CNPq, FINEP, INPE e INPA. No final da década de 80, durante o governo Collor, houve um aprofundamento da crise fiscal e uma virtual falência do estado brasileiro que provocou mais escassez de recursos para a pesquisa.
- 1990 → O MCT é transformado em Secretaria de C&T (SCT).
- 1992 → O MCT volta, em substituição a SCT.
- 1996 → criação do CCT (Conselho Nacional de C&T) para regular as atividades do setor e viabilizar um plano ou um programa de ciência e tecnologia. Nesta década iniciou uma crise na principal base institucional de C&T do país que são as universidades públicas (Guimarães, 1994).
- 2001 → criação do Centro de Gestão e Estudos Estratégicos (CGEE), com a finalidade de gerir os fundos setoriais.

O financiamento da pesquisa nacional se origina principalmente no MCT, seguido do Ministério da Educação através da CAPES, o Ministério da Saúde através da FIOCRUZ (Fundação Oswaldo Cruz) e Ministério da Agricultura através da EMBRAPA (Empresa Brasileira de Pesquisas Agropecuárias).

Além do custeio pelo poder público federal, alguns estados também custeiam pesquisas em institutos e universidades estaduais através da manutenção destas instituições e através de fundações de amparo à pesquisa (*e.g.,* FAPESP) que financiam projetos de pesquisa individuais.

Conforme descrito, em ordem cronológica, apesar da grande estrutura

nacional de fomento à Ciência e Tecnologia, os recursos para tal estão sempre em declínio. Um levantamento realizado pelo MCT (*In:* Krieger, 1999) revela que os gastos totais em C&T no Brasil, em 1996, chegaram a aproximadamente R$ 15 bilhões, equivalentes a 1,1% do PIB. Mas em 1999 o orçamento total do MCT não passou de R$ 1.7 bilhão.

Em termos comparativos, todo o investimento em C&T feito no conjunto dos países da América Latina e do Caribe é apenas uma fração do que é investido nos Estados Unidos (Spinak, 1998).

Em 2000 surgiu um pequeno incremento financeiro para alguns setores da C&T: conforme publicado no jornal da ciência (JC-Email, 2000a), foram criados quatro de treze fundos setoriais destinados a financiar pesquisas em C&T com recursos privados. Além destes quatro fundos (energia, mineral, espacial e transporte), já estava em vigor o fundo do petróleo. Segundo divulgado na nota, durante o ano 2001 o conjunto dos fundos deverá acrescentar ao todo cerca de R$ 1 bilhão, correspondendo a um acréscimo de mais de 58% nos recursos destinados a C&T.

Apesar de que até hoje a maior parte dos investimentos em pesquisa era custeada por recursos públicos, os fundos setoriais são formados por percentuais do faturamento de empresas privatizadas ou por contribuições pela exploração de recursos naturais. Esta alternativa foi viabilizada pelo fato de, em sua maior parte, os recursos serem oriundos de receitas já previstas e cobradas pelo governo das empresas, não implicando em aumento tributário.

Além dos fundos mencionados, em dezembro de 2000 foi aprovado um novo fundo, chamado *verde-amarelo*, para viabilizar intercâmbios entre universidades e empresas. Este fundo é destinado a criar mais uma fonte de recursos e ampliar a participação das empresas em inovações tecnológicas mediante projetos comuns com as Universidades.

> "*A pesquisa cooperativa é um poderoso instrumento de desenvolvimento e difusão de tecnologia, utilizado por países como Estados Unidos, Coréia, Canadá, França e Japão. A interação desses dois pólos do processo de desenvolvimento de inovações, em torno de uma ação estratégica e orientada para solução de grandes*

problemas nacionais, transforma esse programa no nervo central da estratégia dos Fundos Setoriais" (MCT, 2000).

Independente da fonte e do montante dos recursos, de fato a evolução da C&T está diretamente relacionada com o crescimento das universidades públicas (e dos institutos de pesquisa), tanto em nível de graduação onde são formados os futuros pesquisadores, como em nível de pós-graduação, onde se produz a maior parte das pesquisas, apesar da maior parte dos estudantes estarem nas universidades privadas.

Krieger (1999) revela que 60% dos estudantes universitários (cerca de 1 milhão) estudam em universidades privadas. Os 40% restantes (aproximadamente 700.000) estudam em universidades públicas onde a maior parte da pesquisa acadêmica é realizada e apenas 20% estão matriculados em cursos da área de tecnologia.

O sistema de pós-graduação, que está diretamente ligado à atividade de pesquisa e à produção científica imediata, está em plena evolução e já é composto de mais de 2000 programas de Pós-Graduação incluindo mestrados e doutorados.

Os dados obtidos por Brito Cruz (CPCI-1/CCT/1997 (*apud* Krieger, 1999)), confirmam que o pessoal envolvido com atividades de C&T no Brasil está mais concentrado em instituições públicas.

Quadro 1: Pessoal envolvido com C&T no Brasil em 1996.

Professores Universitários	56.760
Federais	32.652
Estaduais	17.062
Privadas	7.046
Estudantes de Pós-Graduação	62.613
Profissionais em Institutos	26.142
Federais	7.632
Estaduais	4.704
Privados	8.765
Outros	5.041
TOTAL	**145.515**

Os dados do Quadro 1 indicam que as pessoas possivelmente envolvidas em pesquisa estão distribuídas em 39% de professores universitários, 43% de estudantes de pós-graduação e 18% de pesquisadores lotados em institutos de pesquisa.

Todo este contingente de pessoas e o dispêndio financeiro dedicados à pesquisa dão algum subsídio para enfrentar muitos problemas da C&T nacional. Para a C&T no Brasil prosseguir na sua evolução, Krieger (1999) destaca os seguintes desafios:

- promover a educação generalizada;
- aumentar a quantidade e qualidade do pessoal envolvido em C&T;
- aumentar o intercâmbio entre a universidade e o setor produtivo;
- aumentar os investimentos em C&T com relação ao PIB;
- aumentar a contribuição da indústria nos investimentos de C&T;
- promover projetos estratégicos com impacto socioeconômico;
- obter o desenvolvimento sustentado e a preservação ambiental.

Para enfrentar estes desafios é necessário conhecer a fundo o potencial de pesquisa nacional e fazer planejamentos.

2.3 Planejamento em C&T

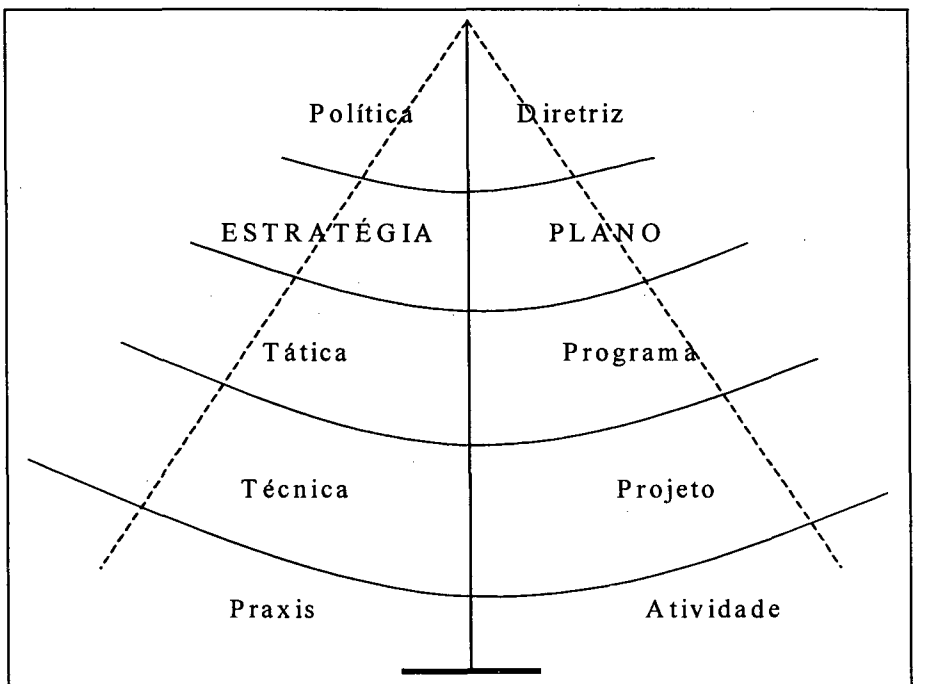
É inconcebível qualquer país almejar o seu desenvolvimento sem que seus governantes percebam a necessidade de planejar sua política científica e tecnológica. A situação em que se encontra a nação (e, por conseqüência, as agências de fomento em C&T) exige meio eficiente de planejamento e o estabelecimento de políticas.

As constantes mudanças no cenário econômico nacional, e a própria dinâmica das organizações, têm exigido a adoção de uma administração estratégica, onde uma das principais etapas é o planejamento. Uma abordagem que leva em consideração a mudança organizacional é o **planejamento estratégico**.

> *"O planejamento estratégico é um processo iterativo da análise das oportunidades e ameaças e de pontos fortes e fracos visando à busca de uma equação para a definição de objetivos apropriados ao*

ajustamento da organização às condições ambientais de mudança". "Consiste na utilização de um arcabouço de técnicas direcionadas para a elaboração de uma análise ambiental interna e externa da organização, definição da missão formulação de objetivos estratégicos, quebra e fixação de novos paradigmas, definição do perfil de negócio e área de negócio, grupos de clientes e produtos ou serviços, formulação de políticas e diretrizes e detalhamento destas em projetos e ações estratégicas" (Silveira, p.19, 1996).

Para chegar à definição de planejamento estratégico acima, Silveira revela a sistemática de importação de termos de outras áreas para a administração e mostra um paralelismo entre a terminologia militar (estratégia, tática, técnica e práxis) e a terminologia administrativa (plano, programa, projeto e atividade), reproduzida na Figura 2.



Fonte: Hausen, 1978; In: Silveira, 1996.

Figura 2: Estratégia empresarial.

Para adaptar estes conceitos no contexto do planejamento em C&T e viabilizar a sua aplicação, será necessário conhecer, além das políticas

públicas na área, fatos relevantes (e surpreendentes) a respeito de C&T, a fim de viabilizar a análise de oportunidades e efetuar um plano estratégico. Para isto, é necessário utilizar técnicas capazes de fornecer conhecimento estratégico. Guimarães explicita bem esta necessidade.

> *"No plano das atividades de articulação (planejamento de C&T), as políticas inexistem ou, no mínimo, não conseguem apresentar resultados visíveis, seja em termos de atuação conjunta da esfera federal com os 'sistemas' estaduais de C&T, seja no âmbito das próprias tarefas de coordenação do MCT"* (Guimarães, 1994).

O planejamento em C&T não pode ser realizado através de decisões de gabinete, de cima para baixo, resultando em investimentos aleatórios e sem levar em consideração, de forma ampla e sistemática, as prioridades e necessidades de desenvolvimento da nação. O decisor necessita de informações precisas e seguras para tomada de decisão, exigindo métodos e meios modernos para obtenção destas informações.

Uma das dificuldades na formulação da política em C&T é sua característica de atividade horizontal, que perpassa todas as funções governamentais (Brisolla, 1998).

Em um país que investe algo em torno de 0,7% do PIB em C&T e cujos recursos provêm predominantemente dos cofres públicos, sempre sujeitos a descontinuidades, o componente planejamento é fundamental para a aplicação adequada dos recursos.

Em agências de promoção de C&T, como o CNPq, CAPES, FINEP e Fundações Estaduais, este planejamento tem como objetivo criar ações de indução do desenvolvimento científico e tecnológico. Para isto é necessário utilizar conhecimento adequado que em boa parte está disponível na forma de indicadores.

2.4 Indicadores em C&T

Para realizar qualquer tipo de análise de informações, seja manual ou automatizada, para qualquer finalidade específica, é necessário fazer uso de

informações que representem, de forma concisa, a realidade que se pretende analisar. Isto pode ser viabilizado pelo uso de indicadores (*e.g.,* indicadores econômicos). Esta seção dedica-se ao aprofundamento deste tema e discute as limitações dos indicadores no âmbito regional.

Na área de C&T também se utilizam indicadores específicos, criados para permitir estudos sobre a atividade científica e tecnológica e auxiliar em tarefas tais como: avaliação da pesquisa, planejamento de política científica, etc. Os indicadores de C&T podem ser úteis para fazer a distribuição dos recursos para pesquisa entre os vários objetivos socioeconômicos e as disciplinas científicas, entre as especialidades, dentro de cada disciplina, e entre os diversos centros e institutos de pesquisa.

Os indicadores científicos surgem da medição dos insumos (recursos humanos, financiamento público e privado, etc.) e dos resultados (produção bibliográficas, patentes, etc.) das instituições científicas. Logo, a existência de bancos de dados sobre C&T é extremamente importante para a medição dos insumos e resultados (produtos) que são bases dos indicadores científicos.

Entretanto, há dificuldade em construir indicadores que reflitam com segurança a realidade que se pretende representar e em estabelecer uma relação de causa-efeito entre a atividade científica e tecnológica e o impacto socioeconômico que a própria ciência provoca (Brisolla, 1998).

Para analisar a C&T de um segmento ou região, é necessário conhecer alguns conceitos quantitativos criados no contexto da atividade científica, como a cientometria e a bibliometria, além da informetria. Macias-Chapula (1998) enfatiza que "em tudo o que se refere à ciência, os indicadores bibliométricos e cientométricos tornaram-se essenciais".

Estes conceitos são apresentados em Taque-Sutcliffe (In: Macias-Chapula, 1998), conforme segue:

Bibliometria: é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. Seus resultados são usados para elaborar previsões e apoiar a tomada de decisões.

Cientometria: é o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. É aplicada no desenvolvimento de políticas

científicas. Sobrepõe-se a bibliometria.

Informetria: é o estudo dos aspectos quantitativos da informação em qualquer formato, não apenas dos cientistas.

Através destes conceitos é possível abstrair a realidade e estabelecer parâmetros numéricos capazes de resumir informações generalizadas sobre investimentos, produção e tendências no campo da ciência e tecnologia. Estes parâmetros são conhecidos como **indicadores** de C&T.

Utilizada para avaliar a literatura (*e.g., Science Citation Index* – SCI), a **bibliometria** é um meio de situar a produção de uma instituição em relação a seu país e cientistas em relação às suas próprias comunidades. No âmbito da gestão de C&T, utiliza-se muito informações bibliométricas, classificadas como um subconjunto da cientometria (veja Figura 3). Combinados a outros indicadores, os estudos bibliométricos podem ajudar na tomada de decisões e no gerenciamento da pesquisa (Macias-Chapula, 1998).

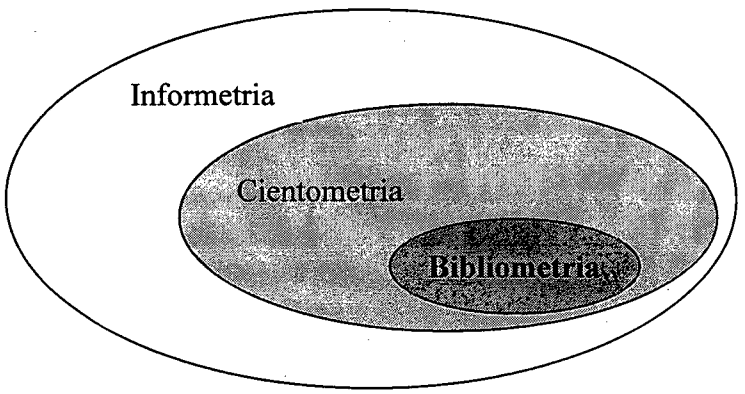


Figura 3: Abrangência dos conceitos de indicadores.

Segundo Spinak (1998), a bibliometria é uma disciplina multidisciplinar que analisa um dos aspectos mais relevantes e objetivos da comunidade científica, a comunicação impressa, compreendendo:

- aplicação de análises estatísticas para estudar as características do uso e criação de documentos;
- estudo quantitativo da produção de artigos;
- aplicação de métodos matemáticos e estatísticos no estudo do uso de livros nas bibliotecas;

- estudo quantitativo das unidades físicas publicadas.

Os indicadores cientométricos podem ser divididos em dois grupos :

- indicadores de publicação: medem a quantidade e impacto das publicações científicas;
- indicadores de citação: medem a quantidade e impacto dos vínculos ou relações entre as publicações científicas.

Velho (1989) define **cientometria** como a área que compreende todos os tipos de análises quantitativas da ciência que se baseiam em fontes de arquivo, sem observação direta da atividade de pesquisa, e que são devotadas aos produtos ou resultados dos processos científicos.

A cientometria aplica técnicas bibliométricas à ciência e inclui, além da bibliometria, outras informações da ciência (*e.g.,* políticas científicas).

Algumas vezes é necessário obter informações de caráter mais geral (*e.g.:,* desemprego no país, PIB) que são classificadas como informetria a qual inclui a disciplina de cientometria, conforme mostra a Figura 3.

Os temas que interessam a cientometria incluem o crescimento quantitativo da ciência, o desenvolvimento das disciplinas, a relação entre ciência e tecnologia, a obsolescência dos paradigmas científicos, a estrutura de comunicação entre os cientistas, produtividade e criatividade dos investigadores, as relações entre o desenvolvimento científico e o crescimento econômico, etc. (Spinak, 1998).

Spinak faz comparações concluindo que a bibliometria trata com as várias medições da literatura, dos artigos e outros meios de comunicação, enquanto que a cientometria trata com a produtividade e utilidade científica, mas ambas podem ser aplicadas a:

- identificar as tendências e o crescimento do conhecimento nas diferentes disciplinas;
- estimar a cobertura das revistas secundárias;
- identificar usuários, autores e tendências em diferentes disciplinas;
- prever as tendências de publicação;
- identificar as revistas do núcleo de cada disciplina;

- formular políticas de aquisição baseadas em previsões;
- estabelecer normas para padronização;
- prever a produtividade de editores, autores, organizações, países, etc. (Sengupta, 1992 (*apud* Spinak, 1998)).

A informetria incorpora muitas informações que estão fora dos limites tanto da bibliometria como da cientometria. Na informetria, desenvolvem-se modelos matemáticos, que servem como base prática para tomada de decisões, cujo valor é sua capacidade de sintetizar, em poucos parâmetros, as características de muitos grupos de dados (Macias-Chapula, 1998).

Segundo Macias-Chapula (1998), a cientometria e a bibliometria, aplicadas à política científica e à indústria da informação, concentram-se em:

- aspectos estatísticos e freqüência de citação de frases;
- relação autor-produtividade medidas;
- características das publicações;
- análise de citação;
- avaliação;
- uso da informação em base de dados;
- avaliação da obsolescência da literatura;
- crescimento de literaturas;
- medida da informação.

Retomando a discussão sobre indicadores, na visão de Kondo (1998), eles são úteis em muitas situações, tais como:

- compreender a contribuição do progresso técnico ao crescimento econômico;
- responder perguntas sobre políticas (e.g.: Qual é o nível de qualidade da pesquisa entre as Universidades da região sul?);
- auxiliar a gestão da C&T regional (e.g.: negociar orçamentos);
- apoiar atividades tais como: prestar assessoria a secretários estaduais, prestação de contas das atividades de C&T, etc.

A comunidade internacional, em especial a OCDE (Organização para a Cooperação e o Desenvolvimento Econômico), desenvolveu metodologias

para elaborar indicadores que podem ser resumidas em três manuais de referência: Manual de Frascati, Manual de Oslo e Manual de Canberra.

O Manual de Frascati (OCDE, 1993) teve sua primeira edição em 1963. O Manual de Oslo (OCDE, 1997) oferece metodologias para a recompilação de dados que permitam interpretar a inovação em C&T. O Manual de Canberra (OCDE, 1995) proporciona diversas metodologias para avaliar os recursos humanos dedicados a C&T, mas não menciona os métodos cientométricos.

A maioria dos indicadores é baseada em insumos e produtos. Muitas propostas (*e.g.,* Cozzens, 1994, In: Brisolla, 1998) têm como idéia básica estimar o impacto da pesquisa através do produto da pesquisa.

Brisolla (1998) questiona se os indicadores já disponíveis realmente desempenham seu papel de "indicar" o sentido do desenvolvimento científico e tecnológico, considerando o caminho não linear (e, portanto, complexo) percorrido pelo processo de transformação de uma invenção científica em uma inovação ou produto. Como não é possível medir diretamente o impacto da pesquisa, a alternativa tem sido medir indiretamente através da avaliação de seus produtos. Outra dificuldade é medir o resultado socioeconômico de um sistema de pesquisa no longo prazo. Deve-se perceber em que medida o resultado diretamente almejado pela pesquisa é atingido em termos de formação de pessoal e de produtos científicos, como publicações e patentes.

Indicadores quantitativos são muito criticados devido ao seu caráter empresarial e à sua limitação para medir idéias. No caso de publicações, elas podem oferecer contribuições diferentes ao conhecimento científico. Os escritos de Macias-Chapula, transcritos abaixo, revelam esta preocupação:

> *"Os produtos da ciência não são objetos, mas idéias, meios de comunicação e reações às idéias de outros. Enquanto os cientistas e o dinheiro investido em pesquisa estiverem inter-relacionados, mais difícil será medir a ciência como um corpo de idéias e fenômenos, ou compreender sua relação com o sistema econômico e social"* (Macias-Chapula, p.135, 1998).

Os indicadores tomam sentido quando baseados em uma abordagem comparativa e podem apresentar plenos significados quando comparados com

valores de outros grupos. Macias-Chapula (1998) apresenta os seguintes indicadores mais conhecidos:

- número de trabalhos;
- co-autoria (mede o nível de cooperação entre grupos);
- número de patentes;
- mapas dos campos científicos e dos países (para localizar as posições relativas na cooperação global);
- número de citações de publicações e patentes.

Macias-Chapula (1998) adverte sobre a dificuldade em se compreender o significado da citação, devido à sua dependência da realidade social, e apresenta os seguintes problemas:

- influências não citadas;
- citação tendenciosa ou preconcebida;
- influências informais não citadas;
- autocitação;
- diferentes tipos;
- variações nas medidas;
- limitações técnicas.

Estes indicadores, criados pela OCDE, têm se mostrado adequados para analisar a produção de C&T dos países centrais, da ciência *mainstream*, porém apresentam sérios problemas de índole epistemológica e instrumental na análise da produção dos países menos desenvolvidos (Spinak, 1995, In: Spinak, 1998).

Kondo (1998) considera que os indicadores estratégicos em C&T para países menos desenvolvidos devem considerar a especificidade desses países, não devendo meramente replicar os indicadores utilizados pela OCDE, que foca apenas a eficiência econômica. O autor propõe um novo marco conceitual que destaca a importância do equilíbrio entre a eficiência econômica e o bem-estar social.

Na perspectiva de Kondo, deve-se trabalhar para criar marcos alternativos que respondam mais adequadamente às necessidades específicas de cada

país. Ele considera o dilema das sociedades democráticas onde o crescimento econômico, por exemplo, traz a destruição ambiental.

Kondo traz à tona o papel do governo na transformação da sociedade: os bens públicos precisam ser produzidos de maneira eficiente, o que é diferente da produção eficiente de bens privados. Se um bem público (*e.g.,* meio ambiente) é mal gerido, surgem desigualdades significativas na sociedade.

Portanto, a escolha de indicadores e consequentemente o processo de tomada de decisão devem levar em consideração razões sólidas, tais como: necessidades sociais do país, efeitos em cascata, melhorar um bem público com uma ação particular que poderá melhorar o bem-estar da sociedade.

2.5 Indicadores Regionais

Os indicadores existentes muitas vezes são insuficientes para apoio à decisão em agências regionais, exigindo a obtenção de novos indicadores ou outras formas de conhecimento úteis à gestão de C&T regional.

Estudos realizados no Observatoire de Sciences et Techniques, utilizando indicadores regionais, produziram resultados interessantes sobre a estratégia dos cientistas e empresários para combinar as fontes locais de recursos para pesquisa com as fontes regionais (Gusmão (*apud* Brisolla, 1998)).

Outra experiência com indicadores regionais vem da FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo) que produziu um conjunto de informações relativas a C&T no estado de São Paulo com o objetivo de subsidiar as políticas públicas naquele estado. Os indicadores produzidos permitiram traçar um panorama do financiamento e da execução das atividades científicas no estado e do espaço que elas ocupam dentro do cenário nacional e poderão subsidiar o governo federal na sua interação com a política estadual (Brisola, 1998).

No campo do desenvolvimento científico e tecnológico, um dos aspectos mais positivos proporcionados pela constituição de 1988 foi o estímulo à criação de agências estaduais de fomento à pesquisa - FAP (Fundação de Amparo à Pesquisa).

A idéia era criar sistemas estaduais que aproveitassem a bem sucedida experiência da FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo), criada através da lei estadual n° 5.918 de 18/10/60, e que viessem a preparar os estados para o aumento da transferência de recursos proporcionada pela descentralização fiscal (Guimarães, 1994).

A criação de FAP's em outros estados é uma alternativa para diminuir as disparidades regionais. Na região sudeste concentra-se 54% dos pesquisadores brasileiros, 73% dos doutores, 70% dos grupos de pesquisa e 92% dos doutorandos. O CNPq destinou 60% das bolsas e fomento à pesquisa a esta região em 2000 (CNPq, 2001).

Na região sul do país a FAPERGS, Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul, está consolidada. Santa Catarina constituiu o FUNCITEC (Lei n.º 7.958, de 5 de junho de 1990), atualmente em processo de consolidação e ainda de pouco impacto na atividade de pesquisa do estado. Mais recentemente foi criada no Paraná uma nova FAP (através do Decreto Lei n° 12.020 de 09/01/98), denominada Fundação Araucária. Contudo, em contraste ao incipiente fomento estadual direto à Ciência e Tecnologia nesta região, contrapõe-se à existência de grandes universidades e centros de pesquisas, tanto públicas como privadas.

Entre os desafios às novas agências estaduais estão as necessidades de conhecer o potencial científico e tecnológico do estado, mapear demandas no setor produtivo e fomentar intercâmbio Universidade-empresa nas áreas com carência de desenvolvimento. É neste cenário que a constituição de bases de dados e sistemas de informações pode servir de subsídio à atividade de indução de desenvolvimento sustentável no âmbito regional ou estadual. O parque tecnológico constituído na esfera federal pode ser subsídio significativo no esforço inicial de constituição deste conhecimento.

Entretanto, devido às limitações dos indicadores e à própria ausência de indicadores regionais, torna-se preeminente a criação de um sistema informatizado para extração de novos conhecimentos, para apoio à decisão em C&T, com enfoque no planejamento regional.

A preocupação em se extrair conhecimentos relativos à região sul é

consonante com a estratégia nacional de respeitar as dimensões continentais do país quando do planejamento de seu desenvolvimento.

Vislumbra-se que tanto os gestores das FAP's como dos centros de pesquisa, quer estatais ou não, certamente serão usuários potenciais de sistemas voltados à gestão de C&T.

Para obter conhecimento sobre C&T são necessárias bases de dados sobre o parque científico e tecnológico do Brasil. O CNPq construiu um importante banco de dados denominado Plataforma Lattes, que pode fornecer conhecimento regional.

2.6 Plataforma Lattes do CNPq

Os novos sistemas computacionais integrados em rede têm permitido a armazenagem de montanhas de dados que cada vez mais desperta o interesse em se explorar estas bases. Nas agências de fomento à pesquisa não é diferente: novos sistemas, baseados em banco de dados sobre C&T, têm sido viabilizados, destacando-se o sistema Coleta da CAPES e a Plataforma Lattes do CNPq.

A Plataforma Lattes é um conjunto de sistemas computacionais que visa compatibilizar e integrar as informações coletadas em diferentes momentos de interação do CNPq com seus usuários. Inclui, entre outros, um sistema de currículos de pesquisadores e o Diretório dos Grupos de Pesquisa no Brasil.

O Currículo Lattes possui informações de pesquisadores de todo o Brasil, compreendendo bolsistas de iniciação científica, bolsistas de mestrado e doutorado, orientadores credenciados e outros clientes do CNPq, fornecendo uma contribuição direta à cientometria. O Diretório possui informações dos grupos de pesquisa cadastrados no CNPq, onde cada membro de grupo pode estar incluído no banco de currículos Lattes.

2.6.1 Diretório dos Grupos de Pesquisa no Brasil

O Diretório dos Grupos de Pesquisa no Brasil, coordenado pelo CNPq (CNPq, 1999a), é uma base de dados implementada desde 1992. Originou-se

em 1991, a partir de uma proposta de elaboração de um Almanaque de Pesquisa no CNPq, bem como no levantamento de grupos de pesquisa realizado pelo Fórum Nacional de Pró-Reitores de Pesquisa. O intuito era organizar o "Programa de Laboratórios Associados" encomendado em 1990 pela então Secretaria de Ciência e Tecnologia (atual Ministério de Ciência e Tecnologia) (Guimarães, 1994). O objetivo básico do projeto era oferecer um suporte de informações atualizadas sobre as atividades de pesquisa, pretendendo obter, periodicamente, a configuração dos recursos humanos e a organização da produção científica e tecnológica brasileiras.

Os grupos de pesquisa inventariados estão localizados em universidades, instituições isoladas de ensino superior, institutos de pesquisa científica, institutos tecnológicos, laboratórios de pesquisa e desenvolvimento de empresas estatais ou ex-estatais, e em algumas organizações não-governamentais com atuação em pesquisa.

Hoje, o Diretório tem o claro objetivo de ser uma plataforma de informação básica sobre o parque científico e tecnológico brasileiro (CNPq, 1999a). O esforço empreendido pelo Brasil após a Segunda Grande Guerra, gerou o maior parque de C&T da América Latina. Entretanto, ainda há carência de informação organizada a respeito, o que enfraquece e dificulta a tomada de decisão sobre os desígnios da C&T nacional. Tal fato transforma o Diretório em instrumento essencial para a gestão de C&T (Martins & Galvão, 1994). O Diretório possui três finalidades importantes:

- fortalecer o intercâmbio entre pesquisadores brasileiros, bem como entre estes e pesquisadores estrangeiros;
- preservar a memória da atividade de pesquisa; e
- constituir-se como ferramenta estratégica para as atividades de planejamento do CNPq.

A terceira finalidade é de vital importância para a tomada de decisão no âmbito do CNPq, quer em nível estratégico ou no âmbito gerencial (ex.: na formulação de políticas de investimentos em C&T) (CNPq, 1998).

2.6.1.1 Organização do Diretório

As informações constantes na base dizem respeito aos recursos humanos constituintes dos grupos, às linhas de pesquisa em andamento, às especialidades do conhecimento, aos setores de atividade envolvidos, aos cursos de mestrado e doutorado com os quais o grupo interage e à produção científica e tecnológica nos três anos imediatamente anteriores à época da coleta dos dados. Dessa forma, cada grupo é situado no espaço e no tempo.

O Diretório tem como unidade básica de análise o *grupo de pesquisa*. Um grupo de pesquisa é caracterizado por um ou dois pesquisadores líderes, podendo ter ou não outros pesquisadores, técnicos, estudantes de graduação ou pós-graduação e estagiários. Todos devem trabalhar em uma ou mais linhas de pesquisa, compartilhando instalações, equipamentos e demais recursos. Devido às peculiaridades de cada área do conhecimento, não há uma estrutura rigorosa. Assim, um grupo pode contar apenas com um pesquisador, trabalhando individualmente com seus estudantes. Também não é obrigatório que os integrantes de um grupo pertençam a uma única instituição. Do mesmo modo, a definição de "pesquisador" também é flexível. O único requisito é que o integrante do grupo tenha pelo menos concluído a graduação.

A base de dados do Diretório está organizada em quatro unidades de análise independentes, incluindo os grupos, mas que interagem:

- Grupos de pesquisa;
- Pesquisadores;
- Linhas de pesquisa e;
- Produção científica, tecnológica e artística.

A primeira versão do Diretório teve seu trabalho de campo realizado no segundo semestre de 1993, englobando a produção científica de 21.541 pesquisadores no triênio 1990-92 (http://www.cnpq.br/gpesq.html).

As informações para a segunda versão foram colhidas no segundo semestre de 1995, cobrindo a produção de 26.799 pesquisadores no biênio 1993-94. As informações podem ser acessadas através da página do CNPq na Internet http://www.cnpq.br/gpesq2.html.

A terceira versão teve seu trabalho de campo realizado no final de 1997 e

foi disponibilizada ao público em geral no segundo trimestre de 1998 através do endereço http://www.cnpq.br/gpesq3, correspondendo à produção de 34.060 pesquisadores do período de 1º de janeiro de 1995 a 30 de junho de 1997, atuando em 8.632 grupos de pesquisa em 181 instituições.

O Diretório 4.0, a sua versão mais recente, é um projeto completo que inclui desde a captura dos dados até a análise do censo da pesquisa e ciência brasileira e sua divulgação. Para tal, o projeto foi organizado em 3 fases:

- captura dos dados sobre os grupos de pesquisa;
- complemento e formação da base integrada;
- dispor os resultados do censo na Internet.

A informação contida no Diretório dos Grupos de Pesquisa no Brasil em sua versão 4.0 foi organizada em Data Warehouse, construído com o objetivo de fornecer uma rápida recuperação das informações. Este Data Warehouse se apoia em seis conjuntos básicos de informação, compostos pelos dados referentes aos Grupos de Pesquisa, aos Pesquisadores, aos Estudantes, ao Pessoal Técnico, às Linhas de Pesquisa e de Produção Científica, Tecnológica e Artística.

Os dados da versão 4.0, colhidos no primeiro semestre de 2000, estão disponíveis na Internet no endereço: http://www.cnpq.br/dgp.html e incluem informações referentes ao triênio 1997-99 de 48.781 pesquisadores de 11.760 grupos em 224 instituições. Comparando com a versão anterior, houve um aumento de 33,6% na quantidade de grupos cadastrados, conforme mostra a Tabela 1, indicando o fortalecimento do Diretório como censo da pesquisa na medida em que o mapeamento aumenta significativamente.

Tabela 1: Diretórios dos grupos de pesquisa.

Diretório	Ano	Instituições
1	1993	99
2	1995	158
3	1997	181
4	2000	224

2.7 Motivações para um Estudo de Caso

As necessidades descritas nas seções anteriores justificam a realização de um estudo de caso para obter determinações capazes de responder aos anseios dos planejadores e executores de políticas regionais.

Apesar de já existirem diversos bancos de dados sobre C&T no Brasil é necessário a partir destes construir bases de dados específicas para cada região do país. Precisa-se descobrir que tipo de conhecimento regional é útil para alcançar o desenvolvimento científico e tecnológico e tentar resolver problemas específicos de determinada região.

Portanto, considerando o desafio de se obter conhecimento novo para uso regional, decidiu-se adotar a região sul do Brasil como estudo de caso.

O principal desafio deste estudo de caso é responder as seguintes perguntas: Técnicas de mineração de dados são viáveis para extrair conhecimento em banco de dados sobre C&T? Como obter conhecimento que seja de interesse à gestão de C&T na região sul?

A primeira pergunta foi respondida parcialmente com a aplicação de uma técnica para extração de regras de associação (algoritmo Apriori) ao Diretório 3.0 do CNPq (Veja ANEXO A). Naquele experimento concluiu-se que as técnicas de mineração de dados são viáveis e promissoras para extrair conhecimento sobre C&T. Gonçalves (2000) também chegou a esta conclusão.

Para responder a segunda pergunta (obter conhecimento que seja relevante às instituições de pesquisa da região sul) é necessário, entre outros, que as agências de fomento à pesquisa desta região adotem métodos de apoio à decisão em C&T e que utilizem técnicas computacionais inteligentes. Estas técnicas podem auxiliar os decisores das agências de fomento a ter acesso às informações estratégicas que justifiquem a escolha de prioridades na pesquisa, que os auxilie a planejar o fomento com base em prioridades e permita gerir, de forma controlada, os recursos destinados à rede de pesquisa do sul do país.

No passado, as decisões eram baseadas apenas no bom senso e experiências passadas do decisor. Atualmente o CNPq possui um banco de dados (Diretório 4.0) contendo registros de 11.760 grupos de pesquisa

espalhados pelo Brasil, compreendendo aproximadamente 50.000 pesquisadores incluindo professores e alunos envolvidos em projetos de pesquisa (veja seção 2.6.1). Além disso, conta também com o Currículo Lattes, contendo dados dos pesquisadores de todo o Brasil, que viabilizam a extração de conhecimento relevante para tomada de decisões. Neste estudo de caso, estes dados foram utilizados conforme ilustra a Figura 4, onde o conhecimento extraído seria correspondente apenas à região sul.

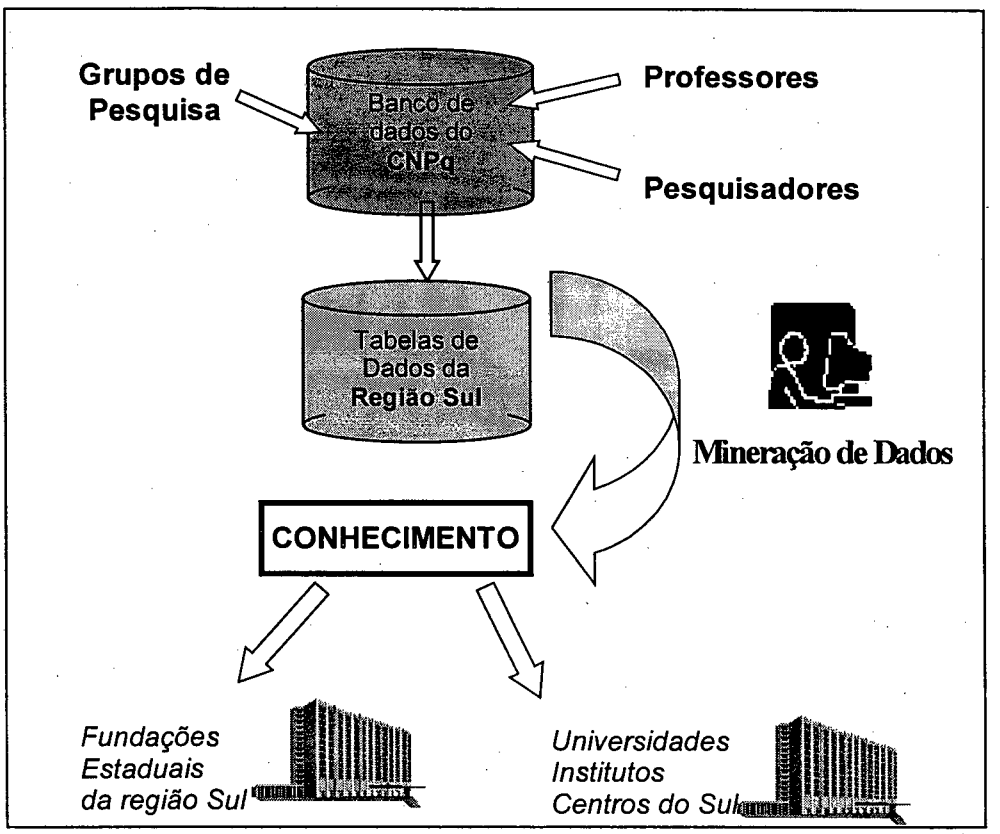


Figura 4: Fluxo de dados para o estudo de caso.

Em suma, com este estudo de caso o objetivo é fornecer informações para orientação de metas a longo prazo aos gestores das agências da região sul e dar subsídios para realizar um planejamento mais eficiente otimizando os recursos existentes. Este estudo de caso é desenvolvido no capitulo 7.

2.8 Considerações Finais

Neste capítulo, foi abordada a importância dos indicadores de C&T como apoio aos gestores das agências encarregadas do planejamento e execução da política científica e tecnológica no Brasil.

Foi dado destaque aos indicadores regionais cuja utilidade é justificada pela existência das fundações de amparo à pesquisa, além das pró-reitorias de pesquisa de universidades locais e outras instituições de pesquisa. As limitações dos indicadores, discutida na seção 2.5, a inexistência de indicadores regionais, bem como a demanda por novos conhecimentos úteis ao planejamento de C&T regional, indicaram a necessidade de se desenvolver um sistema capaz de extrair novos conhecimentos neste seguimento, tendo como estudo de caso a região sul do Brasil.

Um exemplo de conhecimento relevante obtido a partir de base de dados educacionais encontra-se em uma edição da revista Veja:

> *"Um levantamento realizado pelo MEC mostra que existe uma relação direta entre o desempenho dos estudantes no Provão e o número de professores doutores da universidade"*. Veja, 1999.

Quadro 2: Onde tem doutor tem bom aluno (MEC, 1999).

	CAMPEÃ	VICE-CAMPEÃS			
Universidade	USP	UFRGS	UnB	UFPR	UFSM
Doutor/alunos	1/8	1/19	1/20	1/31	1/38

O Quadro 2 mostra o ranking das universidades campeãs do Provão desde 1996. Este exemplo trouxe um indicador novo que alguns podem considerar evidente, mas na ausência dele, muitos poderiam argumentar que não há justificativa para se investir mais na formação de doutores. Entretanto, que outras variáveis explicam o fato destas universidades estarem à frente no ranking das melhores instituições de ensino superior do país? Esta pergunta poderia ser melhor respondida através da mineração dos dados disponíveis e outras respostas surpreendentes poderiam ser obtidas para perguntas que talvez ainda não tenham sido formuladas.

A análise descrita neste capítulo revela que o rápido crescimento dos

bancos de dados em C&T tem inviabilizado o acompanhamento pelos analistas destes dados. Isto indica a demanda crescente por métodos (semi-) automáticos para extração de conhecimento a partir destes dados.

É muito útil saber alguma coisa a respeito do domínio: quais são os campos importantes, quais os relacionamentos possíveis, o que é uma função útil para o usuário, que padrões já são conhecidos e assim por diante. Para isto, o estudo apresentado neste capítulo foi necessário para adquirir uma visão geral da área de aplicação (Ciência e Tecnologia).

O estudo de caso a ser desenvolvido no Capítulo 7 é viabilizado pela disponibilidade de dados no CNPq e por usuários potenciais que podem ser entrevistados. No entanto, exige estudos sobre temas relacionados à descoberta de conhecimento em banco de dados, assunto do próximo capítulo.

3 DESCOBERTA DE CONHECIMENTO

> *"O objetivo é a* extração *de conhecimento de alto nível a partir de dados de baixo nível disponíveis em grandes bancos de dados"* (Fayyad *et al.*, 1996b).

3.1 Introdução

O termo "Descoberta de Conhecimento em Banco de Dados" (KDD – Knowledge Discovery in Databases) surgiu no primeiro *workshop* de KDD em 1989, para enfatizar que o produto final do processo de descoberta em banco de dados era o "conhecimento" (Fayyad *et al.*, 1996b).

Nos anos subseqüentes foram promovidos outros *workshops*, sendo que o último (quinto) foi realizado em 1994. Em 1995 foi realizada a Primeira Conferência Internacional sobre este tema. No ano seguinte, em 1996, realizou-se a Segunda Conferência Internacional, intitulada KDD-96, evento que tem se repetido anualmente reunindo os principais pesquisadores da área e agrupadas inúmeras publicações importantes deste seguimento que têm contribuído com os rumos da pesquisa em KDD.

KDD é uma área interdisciplinar específica que surgiu em resposta à necessidade de novas abordagens e soluções para viabilizar a análise de grandes bancos de dados. Particularmente, KDD tem obtido sucesso na área de marketing, onde a análise de banco de dados de clientes revela padrões de comportamento e preferências que facilitam a definição de estratégias de vendas. A empresa *American Express*, por exemplo, fez aumentar as vendas de cartão de crédito em 15 a 20% com a utilização de marketing auxiliado por técnicas de KDD (Berry (*apud* Fayyad *et al.*, 1996a)).

A viabilidade de aplicação de KDD depende de aspectos práticos e técnicos. O aspecto prático inclui considerações sobre o impacto que a aplicação irá provocar, medido por critérios tais como rendimento, redução de custos, melhora na qualidade dos produtos ou economia de tempo na instituição. Em aplicações científicas, o impacto pode ser medido por novidade

e qualidade do conhecimento descoberto bem como pelo aumento da automação de processos de análises manuais.

O aspecto técnico se refere à disponibilidade de dados suficientes, ou seja, a complexidade do problema pode exigir grande quantidade de atributos e casos (ou "registros" de banco de dados). Por outro lado, muitos atributos podem ser irrelevantes para o problema tratado. Em ambos, o conhecimento do domínio da aplicação, tais como: campos mais importantes, qual o relacionamento entre eles, qual a utilidade para o usuário, que padrões já são conhecidos, etc., poderá contribuir para redução tanto da busca na tarefa de MD quanto nas demais etapas do processo de KDD.

Apesar das informações resumidas e significativas para tomada de decisão serem de volume menor, geralmente elas não estão disponíveis e exigem a sua extração a partir de grandes quantidades de dados que crescem com o tamanho e a idade das instituições, dificultando o processo de extração de conhecimento. Além disso, muitas vezes o usuário não sabe sequer formular uma questão desejada.

A aplicação de KDD muitas vezes se depara também com os seguintes desafios: bancos de dados enormes ou poucos dados; muitas dimensões; mudança nos dados; dados com ruído ou perda de dados; interação complexa entre atributos, etc.

Neste contexto, o desafio que se apresenta para as organizações pode ser simplificado como a resolução de duas questões básicas:

1. Como organizar os dados?
2. Como extrair conhecimento dos dados organizados?

A primeira questão pode ser equacionada através da construção de *Data Warehouse*. Esta tecnologia permite armazenar informações, anteriormente dispersas, através da identificação, compreensão, integração e agregação dos dados, de forma a posicioná-los nos locais mais apropriados visando a atender à estratégia organizacional das empresas (Brackett,1996).

Entretanto, em resposta à segunda questão, para extrair conhecimento de um sistema de *Data Warehouse,* são necessárias ferramentas de exploração, hoje conhecidas como *Mineração de Dados* (MD), que podem incorporar

técnicas estatísticas e/ou de Inteligência Artificial (IA), capazes de fornecer respostas a várias questões ou mesmo de descobrir novos conhecimentos em grandes bancos de dados. *MD* é especialmente útil em casos onde não se conhece a pergunta, mas, mesmo assim, existe a necessidade de respostas.

3.2 Dados, Informação e Conhecimento

Seja qual for a área de atuação de uma organização, a grande quantidade de informações acumuladas nos bancos de dados informatizados desta organização pode esconder conhecimentos valiosos e úteis para a tomada de decisões. O aumento acentuado no volume dos dados, associado à crescente demanda por conhecimento novo para decisões estratégicas, tem provocado o interesse crescente em descobrir conhecimento em banco de dados. Portanto, é importante uma distinção mais precisa entre dados, informação e conhecimento.

No mundo científico, dados representam observações coletadas sobre algum fenômeno em estudo e o desafio é como explicar melhor o que foi observado. Nos negócios, os dados capturam informações sobre os mercados, concorrentes e clientes. Em sistemas de manufatura, dados capturam oportunidades de melhorar o desempenho, otimização e meios de melhorar processos e resolver problemas (Fayyad, 1997).

No coração do problema de inferência a partir de dados está o campo da estatística. Tanto no lado da validação de hipóteses como no lado da análise exploratória dos dados as técnicas estatísticas são de importância fundamental. No entanto, as abordagens tradicionais das áreas de estatística e reconhecimento de padrões podem entrar em colapso diante de montanhas de dados.

Nos segmentos responsáveis pelo fomento a C&T, dados representam informações sobre pesquisadores, grupos, projetos e instituições de pesquisa. O desafio é transformar estas informações em subsídios para esta mesma comunidade, ou seja, como descobrir conhecimento, a partir destes dados, que auxilie a tomada de decisão no planejamento de C&T. Logo, é importante distinguir informação de conhecimento.

Informação pode ser definida como dados organizados, geralmente em forma de tabelas, que pode ser obtida diretamente a partir do banco de dados através de alguma ferramenta desenvolvida especificamente para este fim e com o auxílio de um gerenciador de banco de dados. Por exemplo, na página do CNPq na Internet há muita informação disponível sobre os grupos de pesquisa do Brasil.

Conhecimento é um tipo de informação mais complexa. Segundo Schreiber *et al.* (2000), conhecimento é "*todo o conjunto de dados e informações que as pessoas utilizam na prática para executar ações, a fim de realizar tarefas e criar nova informação*".

No entanto, segundo Schreiber e seus colegas, as fronteiras da definição entre dados, informação e conhecimento não são bem definidas e dependem muito do contexto considerado.

Existem diversas formas de representar o conhecimento em um sistema de aprendizagem, sendo mais comum através da relação entre atributos do banco de dados. No contexto da tarefa de classificação, o conhecimento descoberto em geral é expresso como um conjunto de regras de produção (veja seção 4.2.1) do tipo SE... ENTÃO..., tipo de representação do conhecimento bastante intuitivo para o usuário (Carvalho & Freitas, 2000).

Apesar da sua disponibilidade e importância, conhecimento nem sempre pode ser obtido diretamente ou com ferramentas comuns, tornando-se muitas vezes uma tarefa difícil. Para facilitar esta tarefa há um processo consagrado na literatura para descoberta de conhecimento, assunto da próxima seção.

3.3 O Processo de Descoberta de Conhecimento

> *"KDD é o processo não trivial de identificação, a partir de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis"* (Fayyad, 1996b).

Na definição de Fayyad, KDD é descrito como um processo geral de descoberta de conhecimento composto por várias etapas, incluindo: preparação dos dados, busca de padrões, avaliação do conhecimento e

refinamentos. O termo não trivial significa que envolve algum mecanismo de busca ou inferência, e não qualquer processamento de dados direto de uma quantidade pré-definida.

Nessa definição, um conjunto de dados representa fatos enquanto que os padrões podem ser interpretados como uma expressão em alguma linguagem capaz de descrever um subconjunto de dados ou um modelo aplicável a este subconjunto. Os padrões descobertos devem ser válidos diante de novos dados com algum grau de certeza. Estes padrões podem ser considerados conhecimento dependendo de sua natureza.

Os padrões devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para tomada de decisão.

Um conceito importante, chamado de *interestingness* ("grau de interesse") (Piatetsky-Shapiro & Matheus, 1994) é usualmente utilizado como uma medida geral do valor de um padrão, podendo combinar validade, novidade, simplicidade (compreensibilidade) e utilidade.

Para descobrir conhecimento que seja relevante, é importante estabelecer metas bem definidas. Segundo Fayyad *et al.* (1996b), no processo de descoberta de conhecimento as metas são definidas em função dos objetivos na utilização do sistema, podendo ser de dois tipos básicos: *verificação* ou *descoberta*.

Quando a meta é do tipo *verificação*, o sistema está limitado a verificar hipóteses definidas pelo usuário, enquanto que na *descoberta* o sistema encontra novos padrões de forma autônoma. A meta do tipo *descoberta* pode ser subdividida em: *previsão* e *descrição*, conforme a Figura 5.

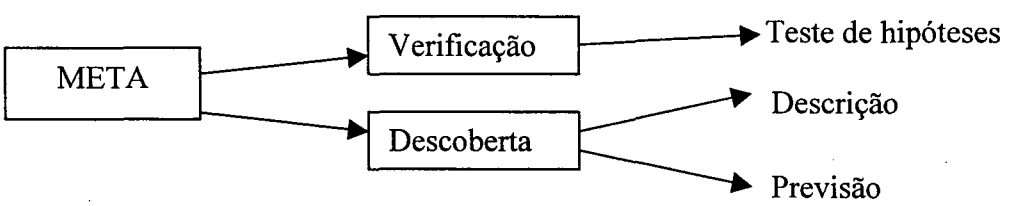


Figura 5: Tipos de metas no processo de KDD.

A *descrição* procura encontrar padrões, interpretáveis pelos usuários, que descrevam os dados. A *previsão* parte de diversas variáveis para prever outras variáveis ou valores desconhecidos (Fayyad *et al.*, 1996a).

Na *previsão*, o sistema irá encontrar padrões com o propósito de estimar o comportamento futuro de algumas entidades, enquanto que na *descrição* o sistema deverá encontrar padrões com o propósito de apresentá-los ao usuário em uma forma compreensível pelo homem. As fronteiras entre previsão e descrição não são bem definidas, mas em KDD a descrição tende a ser mais importante do que a previsão (Fayyad *et al.*, 1996b).

As metas de previsão e descrição são alcançadas através de alguma das seguintes tarefas de MD: classificação, regressão, agrupamento, sumarização, modelagem de dependência e identificação de mudanças e desvios, sendo a tarefa de classificação a mais empregada.

Na modelagem preditiva para classificação ou regressão podem ser utilizadas, dentre inúmeros outras formas de representação do conhecimento, árvores de decisão e regras.

Retomando a explanação sobre o processo de KDD, apesar da mineração de dados ser a etapa principal, o processo de descoberta de conhecimento em banco de dados não se resume a minerar os dados. Exige-se a construção de mais dois estágios: pré-processamento e pós-processamento, conforme ilustra a Figura 6.

Fayyad classifica o processo geral de KDD nas seguintes etapas:

- desenvolver um entendimento do domínio da aplicação, identificar o tipo de conhecimento que interessa, e identificar a meta do processo de KDD a partir do ponto de vista do usuário;
- realizar pré-processamento incluindo operações básicas, tais como: seleção de atributos relevantes, remoção de ruído, tratamento da ausência de valores de atributos e conversão de dados categóricos ou contínuos;
- reduzir os dados em função do objetivo da tarefa;
- escolher a tarefa de MD baseado no objetivo do processo de KDD;
- escolher o algoritmo de MD apropriado;

- realizar a mineração dos dados propriamente dita;
- interpretar os padrões descobertos, podendo retornar para um dos passos anteriores;
- consolidar o conhecimento descoberto, incluindo a conferência e a solução de possíveis conflitos com conhecimentos anteriores.

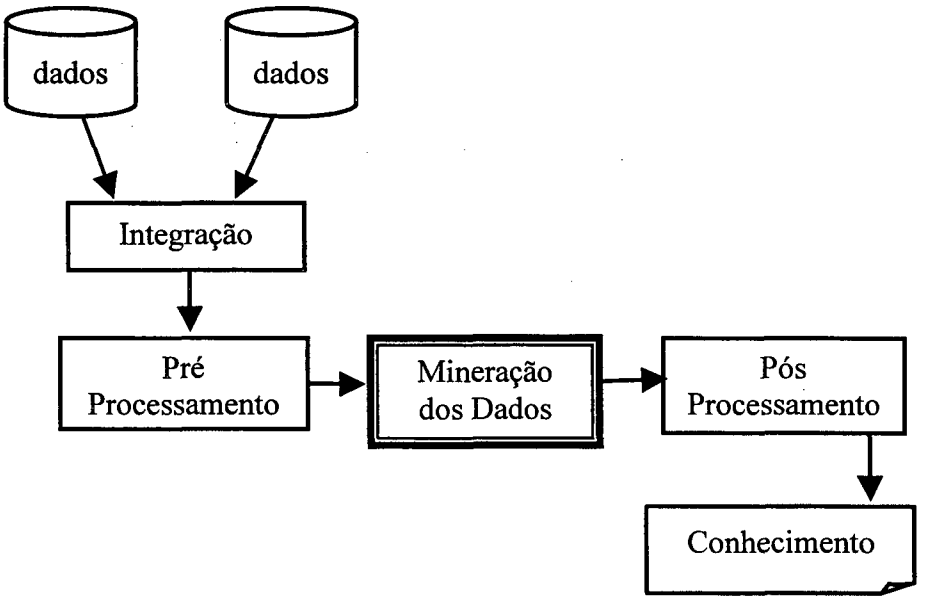


Figura 6: Etapas para descoberta de conhecimento.

Portanto, o processo de KDD utiliza banco de dados para realizar: seleção de atributos e transformações necessárias sobre os dados (pré-processamento); aplicação de métodos (algoritmos) de MD para extrair padrões dos dados; e avaliação do produto da MD para identificar os padrões julgados como "conhecimento" (pós-processamento).

3.3.1 Características dos Dados

Uma das principais hipóteses de Rendell & Cho (1990) é que, apesar de haver diversos tipos de algoritmos de indução de regras, as características destes têm menos efeito do que os dados e o caráter dos conceitos.

Algumas características nos dados de treinamento, tais como número de registros, tipo dos dados e quantidade de erros, podem afetar a exatidão da

aprendizagem de forma significativa.

Há outras características também importantes para a aprendizagem, incluindo: número de atributos, tamanho da classe de conceitos e concentração dos conceitos no espaço de exemplos possíveis (número de picos separados). Experimentos realizados por Rendell & Cho (1990) revelaram que algumas destas características afetam drasticamente a exatidão da aprendizagem de conceitos.

Conceito identifica uma característica relevante que pode ser classificada em função dos valores dos atributos e da freqüência dos mesmos. Exemplo: classificar um tipo de pesquisador em função dele ter ou não potencial para coordenar determinados tipos de projetos.

Algumas vezes, as características dos dados interagem entre si de uma forma não intuitiva. Por exemplo, ruído nos dados pode degradar a exatidão de classificação de forma diferente dependendo do tamanho do conceito. Comparada com os efeitos de algumas características dos dados, a escolha do algoritmo de aprendizagem se torna menos importante.

Há questões que precisam ser consideradas, tais como:

- Quais características dos dados afetam mais a mineração de dados?
- Como estas características interagem entre si?
- Os dados podem ser caracterizados de forma a melhorar a descoberta de conceitos?

Rendell & Cho (1990) verificaram que é adequado considerar um conceito como uma função ou superfície, dentro do espaço de valores possíveis de atributos, para avaliar seus efeitos na descoberta de conhecimento. Eles enfocam duas características: tamanho e concentração do conceito, onde **tamanho** representa a proporção de exemplos positivos (um exemplo é classificado como positivo quando sua classe é aquela que o usuário está diretamente interessado – por exemplo, o paciente está doente) e **concentração** caracteriza a distribuição de exemplos positivos dentro do espaço de valores possíveis, onde:

alta concentração → poucas regiões de conceitos;

baixa concentração → conceitos distribuídos em muitas regiões.

Segundo os autores, uma aplicação com maior concentração dos conceitos tende a apresentar melhor desempenho.

O número de atributos também afeta a descoberta de conhecimento, mas de forma linear. Se o número de atributos é muito grande pode-se utilizar técnicas de seleção ou substituição de atributos (Apte *et al.*, 1993; Bala *et al.*, 1995; Cherkauer & Shavlik, 1997; Freitas & Lavington, 1998; Kim *et al.*,2000; Páircéir *et al.*, 2000; Becher *et al.*, 2000).

Escala de atributos: os atributos podem ser desordenados (nominal), parcialmente ordenados (árvore estruturada), ordenados ou de valores inteiros ou reais (intervalo). O espaço de valores prováveis de um conceito usando n atributos é n-dimensional. Logo, a estrutura do espaço depende das escalas dos atributos.

Ruído nos dados: tanto atributos errados como erro nos valores das classes afetam a exatidão. Sistemas de indução (tipo de raciocínio em que de fatos particulares se tira uma conclusão genérica) podem aprender mesmo na presença de ruído, tanto nas classes como nos atributos, apesar de que erro de classe provoca prejuízos mais significativos. Isto porque o erro de classe não só provoca a destruição de informação, mas inverte a mesma. Com um atributo errado, o sistema ainda pode classificar corretamente. Portanto, erro na classe provoca distorções no resultado, enquanto que erro em atributos pode ou não causar distorções.

Tamanho do conceito: é a proporção de exemplos positivos. Pode afetar a exatidão da classificação, especialmente quando há ruído nos dados.

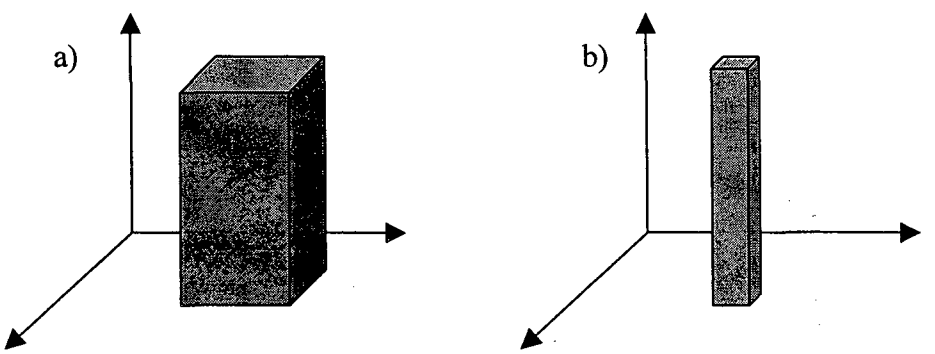


Figura 7: Variação do tamanho relativo do conceito (positivos).

Pela Figura 7 é fácil observar que há maior probabilidade de um ruído interferir no resultado de (a) do que de (b) devido à sua maior área. No entanto, (b) representa dados típicos encontrados em alguns problemas. Exemplo: há poucos pesquisadores da área de letras com alto índice de publicações de livros internacionais.

Concentração dos conceitos: é a quantidade de picos de conceitos e interfere drasticamente na descoberta de conhecimento. No caso de valores inteiros ou reais, um pico é uma vizinhança de S (espaço) que tem a concentração maior do que a média. Vizinhança é um termo com significado apenas para escalar ordenados. Para outras escalas (nominal) um pico é um ponto até que alguma ordem seja aprendida.

Em particular, os autores mostram que:

Aumento no número de picos → Aumenta interação entre atributos.

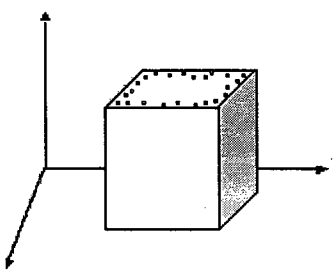


Figura 8: Concentração de um conceito (Rendell & Cho 1990).

A concentração de conceitos afeta mais a exatidão porque conceitos dispersos dificultam a aprendizagem em algoritmos convencionais. Dados concentrados nas bordas da classe (Figura 8) podem aumentar a velocidade de aprendizagem e a exatidão. Algoritmos de indução convencionais têm seu desempenho reduzido quando o conceito é muito disperso.

Porquanto, em geral aumentar a concentração de conceitos melhora a exatidão. Logo, pode-se transformar o espaço de valores possíveis dos exemplos (*instance space*) para diminuir os picos. Uma forma de fazer isto é converter atributos nominais em inteiros ordenados que, em alguns casos, facilita a descoberta. Exemplo: o atributo NÍVEL_DE_FORMAÇÃO, que indica a formação do pesquisador, pode ter seu domínio convertido em inteiros

ordenados ou até ser transformado em atributo difuso.

3.3.2 Pré-Processamento

Um dos principais obstáculos para MD são dados de má qualidade. Quando os dados são precários o produto de qualquer tarefa de MD também é precário.

Muitos algoritmos não processam dados com ausência de valores de atributos, outros não trabalham com valores contínuos, outros não aceitam dados categóricos ou binários. Para resolver estes problemas é necessário efetuar um pré-processamento, que pode ser realizado manualmente ou de forma automática.

Existem dois tipos básicos de erros em dados: sistemáticos e não-sistemáticos (ou ruído). Os erros sistemáticos (*e.g.,* falha na calibração de equipamento) são introduzidos de forma previsível e são potencialmente detectáveis e corrigíveis.

Os erros não-sistemáticos (*noise*) são introduzidos de forma imprevisível e são muito difíceis de detectar e corrigir. Exemplos: usuário fornece informação equivocada; erros de digitação; inconsistência nos valores (*e.g.,* NACIONALIDADE = B/E ou 1/2).

Outra situação é a falta de alguns valores de atributos. Neste caso a causa pode estar na coleta dos dados, remoção de dados devido à inconsistência ou o próprio significado do atributo ser incompreensível. No caso de ausência de valores de atributos, existem as seguintes alternativas:

- remover registros com valores faltando;
- prever o valor faltando com base nos valores de outros atributos;
- lidar com os valores faltantes dentro do algoritmo de MD;
- substituir os valores ausentes pela moda (valor mais freqüente, no caso de atributos categóricos) ou pela média ou mediana (no caso de valores contínuos).

Para substituir dados de treinamento, a moda (média, mediana) é calculada com base no conjunto de treinamento apenas. Para os dados de teste, eles são calculados com base em toda a população (King *et al.*, 1995).

Outro problema básico em KDD é a privacidade dos dados. A OCDE

(Organization for Economic Cooperation and Development) (O'Leary, 1995, In: Fayyad *et al.*, 1996a) sugere que dados sobre indivíduos específicos não devem ser analisados sem os seus consentimentos e, caso autorizem, coletados apenas para um propósito especificado. Entretanto, nada impede a utilização de registros de transações ou cadastros para se procurar padrões gerais (não de indivíduos). Uma alternativa é remover ou substituir campos de identificação de indivíduos durante a etapa de pré-processamento.

Em muitos casos (*e.g.*, dados sobre C&T) o objetivo pode ser a descoberta de padrões a respeito de grupos, não de indivíduos, o que não viola as restrições de privacidade individual.

A redundância é outro problema de qualidade dos dados. Atributos diferentes podem conter praticamente a mesma informação. Por exemplo, data de nascimento e idade. Redundância pode existir por diversas razões, tais como: objetivando aumentar a performance do sistema ou devido à integração de dados de fontes diferentes.

Em alguns domínios de aplicação existe o problema de dados dependentes do tempo (*e.g.,* mercado acionário) que mudam dinamicamente. Neste caso, mesmo quando a taxa de mudança em função do tempo é pequena, tais como o estado civil em formulários e informações em páginas da Internet, atualizações no conteúdo dos dados podem não ser captadas.

Quanto à natureza dos dados, muitos algoritmos não processam certos tipos. Por exemplo, os algoritmos Discrim (linear discriminant) e Quadra (quadratic discriminant) não trabalham com valores categóricos (King *et al.*, 1995). Para resolver este problema, os valores categóricos devem ser substituídos por N atributos binários, onde N é o número de categorias. Por outro lado, o algoritmo CASTLE não trabalha com atributos contínuos, logo os dados necessitam ser discretizados.

Portanto, para a eficiente aplicação das técnicas de MD é necessário antes realizar uma preparação dos dados, conhecida como pré-processamento, que inclui as seguintes etapas:

- Integração dos dados: remover inconsistências nos nomes ou em valores de atributos de diferentes origens;

- Limpeza dos dados: detectar e corrigir erros nos dados, substituir valores perdidos, etc.;
- Conversão de dados nominais, ou em forma de códigos, para números inteiros;
- Redução do domínio (valores possíveis) para reduzir a distribuição dos valores no espaço de valores originalmente possíveis;
- Construir ou derivar novos atributos;
- Discretização: transformar atributos contínuos em categóricos, quando o algoritmo de MD não trabalha com atributos contínuos ou para melhorar a compreensão do conhecimento descoberto;
- Seleção de atributos: escolher atributos relevantes para a tarefa em questão (Wang & Sundaresh, 1998). Por exemplo, o atributo "*nome do pesquisador*" não deve ser escolhido para não gerar regras de aplicação pessoal (*e.g.*: nome = "um_nome_específico" $\Rightarrow$ produção = "ruim").

Dentre as etapas mencionadas acima, um dos maiores desafios é a seleção de atributos relevantes para realizar uma tarefa de MD específica. Existem duas abordagens principais utilizadas para este fim: *Processo Envoltório* (*Wrapper*) (Kohavi & John, 1998) e *Processo por Filtro*. Em geral, as técnicas tipo envoltório tendem a ser mais efetivas, ou seja, resultam em uma menor taxa de erro de classificação se comparadas com as técnicas do tipo filtro, mas estas últimas normalmente são mais eficientes, uma vez que consomem menor tempo de processamento.

As técnicas para seleção de atributos do tipo envoltório (Bala *et al.*, 1995) exigem inúmeras execuções do algoritmo de MD, razão pela qual geralmente consomem mais tempo de processamento.

Yang & Honavar (1998) utilizaram algoritmos genéticos para encontrar um subconjunto quase ótimo de atributos relevantes para descoberta de conhecimento.

3.3.3 Mineração de Dados

Este é um tema de pesquisa cujas aplicações são virtualmente ilimitadas. Pode-se aplicar Mineração de Dados a qualquer tipo de área (financeira, comercial, medicina, ciências, etc.), desde que se tenham dados disponíveis. De fato as pessoas estão se afogando em dados, mas sedentos de conhecimento, o problema é como extrair conhecimento novo a partir de uma enorme quantidade de dados.

Segundo uma publicação da revista Times (2000), Mineração de Dados é um dos 10 "hottest jobs" para o futuro juntamente com programadores de genes e outros serviços de tecnologia altamente avançada.

"Mineração de dados" é um termo mais utilizado por profissionais da área de estatística, analistas de dados e pela comunidade que desenvolve sistemas de informações gerenciais, enquanto KDD tem sido mais utilizado por pesquisadores em IA e aprendizagem de máquina (Fayyad *et al.*, 1996a).

Fayyad diferencia os termos MD e KDD destacando que o componente de MD se refere apenas ao meio pelo qual padrões são extraídos e enumerados a partir dos dados, enquanto que KDD envolve a avaliação e interpretação dos padrões para decidir o que é conhecimento e o que não é, incluindo a escolha do esquema de codificação, pré-processamento, amostragem e projeções realizadas antes da etapa de MD, bem como o pós–processamento naturalmente realizado depois da etapa de MD.

Segundo Fayyad (1997), tarefas realizadas através de técnicas oriundas das áreas de estatística, reconhecimento de padrões, RNA (Rede Neural Artificial), aprendizagem de máquina e banco de dados podem ser enquadradas na fase de MD. Outros campos relacionados são otimização (de busca), computação paralela e de alto desempenho, modelagem de conhecimento, gerência de incertezas e visualização de dados.

Técnicas de MD utilizam dados históricos para aprendizagem objetivando realizar alguma tarefa específica. Esta tarefa tem como meta responder alguma pergunta particular de interesse do usuário. Portanto, é necessário informar qual problema se deseja resolver.

Como ilustração, considere um banco de dados contendo registros de

clientes e mercadorias vendidas. Uma consulta ao banco de dados para a extração da informação poderia ser: "Quantos computadores foram vendidos para o cliente X na data dd/mm/aa?".

Esta seria uma operação comum da baixa administração da empresa. Entretanto, as técnicas de *Mineração de Dados* visam atender as aplicações de níveis administrativos mais elevados, tais como: marketing (mala direta direcionada), planejamento de estoque, abertura de novas filiais e outras decisões estratégicas. Seguindo a ilustração acima, uma técnica de *MD* poderia extrair conhecimento do tipo "SE (idade = '[25 a 35] anos') E (profissão = 'advogado') ENTÃO (compra = 'computador')" com uma freqüência, por exemplo, de 90%.

Como resultado, o conhecimento obtido poderia ser usado para responder a uma provável pergunta do setor de marketing: Quais os clientes que têm alta probabilidade de comprar computadores?

Além da área de negócios, MD tem sido também utilizada na área científica (*e.g.,* biologia molecular, modelagem de mudanças climáticas globais, etc.). Exemplos de aplicações podem ser obtidas em Fayyad & Uthurusamy (1995).

Para encontrar respostas, ou extrair conhecimento relevante, existem diversas técnicas de *MD* disponíveis na literatura (Chen *et al.,* 1996; Cheung *et al.,* 1996). As principais podem ser agrupadas em:

- Indução e/ou Extração de Regras;
- Redes Neurais;
- Algoritmos Evolucionários;
- Técnicas estatísticas (classificadores e redes Bayesianas, etc.); e
- Conjuntos Difusos.

Essas técnicas podem ser aplicadas a diversas tarefas de mineração de dados, tais como: extração de regras de associação, classificação, previsão em geral, determinação e análise de agrupamento, etc.

Seja qual for a tarefa a ser realizada, a aplicação cega de métodos de MD (chamada na literatura de estatística de "dragagem de dados") pode se tornar uma atividade perigosa e conduzir facilmente para a descoberta de padrões sem sentido (Fayyad *et al.*, 1996b).

Para a escolha da técnica mais adequada é estratégico saber alguma coisa a respeito do domínio da aplicação de MD: quais são os atributos importantes, quais os relacionamentos possíveis, o que é uma função útil para o usuário, que padrões já são conhecidos e assim por diante.

> *"Não há um método de Mineração de Dados 'universal' e a escolha de um algoritmo particular para uma aplicação particular é de certa forma uma arte". (Fayyad et al., p. 86, 1996b).*

Segundo Fayyad *et al.* (1996b), os algoritmos de MD diferem primariamente nos critérios utilizados para avaliar o modelo e/ou no método de busca utilizado. Ele adverte que não há critérios estabelecidos para se decidir quais métodos devem ser usados em dada circunstância e que muitas abordagens são aproximações heurísticas para evitar o alto custo de processamento que seria necessário para se encontrar soluções ótimas.

Fayyad identifica três componentes primários em algoritmos de MD:

a) **Representação do modelo:** é a linguagem utilizada para descrever os padrões a serem descobertos;
b) **critério de avaliação do modelo:** afirmação quantitativa (ou função de aptidão) da qualidade que um padrão específico possui (um modelo e seus parâmetros) em alcançar as metas do processo de KDD. Modelos preditivos muitas vezes são julgados pela exatidão de previsão medida utilizando algum conjunto de dados de teste. Modelos descritivos podem ser avaliados pela novidade, utilidade e facilidade de compreensão do modelo obtido, além da exatidão;
c) **método de busca:** é constituído por dois componentes (busca de parâmetros e busca do modelo). Após a escolha da representação e do critério de avaliação do modelo, o problema de MD fica reduzido à tarefa de otimização (encontrar os parâmetros/modelos que satisfaçam o critério de avaliação).

Na busca, o algoritmo deve procurar os parâmetros que otimizem o critério de avaliação do modelo. A busca do modelo ocorre em um processo iterativo externo ao método de busca dos parâmetros.

3.3.4 Pós-Processamento

O pós-processamento é utilizado principalmente para avaliar o processo de descoberta, melhorar a compreensão e/ou selecionar conhecimento descoberto que seja mais relevante. Quando são geradas muitas regras, é importante remover algumas regras e/ou condições para facilitar a compreensão do conhecimento extraído.

3.3.4.1 Avaliação do Processo de Descoberta

Existem diversas abordagens para avaliar o processo de descoberta de conhecimento, incluindo-se: exatidão dos resultados (*e.g.*, alguma medida da taxa de acerto), eficiência (tempo de processamento), facilidade de compreensão do conhecimento extraído, etc.. A maior parte da literatura utiliza exatidão (taxa de erro) como principal meio para avaliar as técnicas de KDD (Freitas, 1997a), principalmente no contexto da tarefa de classificação.

Os dados utilizados para efetuar a extração de conhecimento são divididos em dois grupos exclusivos: conjunto de treinamento e conjunto de teste. O algoritmo deve descobrir regras acessando apenas os dados de treinamento.

Uma vez que o processo de treinamento tenha terminado e o algoritmo tenha encontrado um conjunto de regras de classificação, a *performance* para estas regras é medida através da aplicação destas regras sobre os dados de teste, caracterizando uma forma de aprendizagem supervisionada.

A exatidão é representada pela proporção de classificações realizadas corretamente. Ela pode ser medida sobre os dados de treinamento e sobre os dados de teste, sendo que uma alta exatidão neste último é mais difícil de ser alcançada, devido ao fato de se utilizar dados que não foram considerados anteriormente durante o treinamento. Logo, a exatidão perante os dados de teste é considerada muito mais importante.

Domingos (1998, p.37) apresentou um estudo no qual se comparam duas abordagens para avaliar processos de descoberta (modelos) baseado em exatidão e simplicidade, chamadas de "N*avalhas de Occam*".

Na primeira abordagem, intitulada "primeira navalha de *Occam*", Domingos

afirma que "dados dois modelos com a mesma taxa de erro de generalização (calculado sobre os dados de teste), deve-se preferir o mais simples porque simplicidade é desejável por si só". Na segunda abordagem, ou segunda navalha, ele afirma que "dados dois modelos com a mesma taxa de erro (utilizando apenas os dados de treinamento), o modelo mais simples deve ser preferido porque provavelmente terá o menor erro de generalização (com os dados de teste)" .

Domingos (1998, p.39) afirma que "para todo domínio onde um modelo simples é mais exato do que um modelo mais complexo, existe um domínio onde o contrário é verdadeiro, ou seja, o modelo complexo é mais exato, implicando que não há argumento que justifique a preferência universal ou por modelos simples ou por modelos complexos".

Na sua pesquisa, Domingos conclui que a primeira navalha é consenso e a melhor forma de aplicá-la é, dado dois modelos com o mesmo erro de generalização, deve-se preferir o mais compreensível.

Domingos considera mais promissores "sistemas que primeiro encontram um modelo mais exato possível e depois extraem, a partir deste, um modelo mais compreensível de complexidade variada".

Retomando o primeiro parágrafo desta seção, a eficiência (ou velocidade de processamento) se refere ao tempo consumido para aprender e ao tempo consumido para classificar um novo exemplo. Dependendo da aplicação, a eficiência pode ser mais relevante do que a exatidão, como em alguns sistemas de tempo real.

Em algumas situações o tempo de aprendizagem é de extrema importância, especialmente quando o sistema está inserido em um ambiente altamente dinâmico. No entanto, quando os dados são renovados apenas periodicamente, o tempo de aprendizagem é de menor importância ou até irrelevante (*e.g.*, censo anual).

Facilidade de compreensão dos resultados da classificação (por exemplo, regras) é outra forma de avaliação do processo de descoberta que favorece a credibilidade no sistema por parte do usuário.

Em muitas aplicações a compreensão por parte dos humanos é

extremamente importante, exigindo algum complemento no pós-processamento. As soluções possíveis incluem: representação gráfica, utilização de estrutura de regras, geração de padrões expressos em linguagem natural, e emprego de técnicas para visualização de dados e conhecimento. No caso de se utilizar estrutura de regras, pode-se empregar estratégias de refinamento para eliminar conhecimento descoberto que seja redundante (Gago & Bento, 1998).

Em suma, o processo geral de KDD é formado por três etapas principais, destacando-se a etapa de mineração de dados. Entretanto, um dos primeiros passos em um processo de KDD é a definição da tarefa de *MD* a ser realizada, que determina o tipo de conhecimento a ser descoberto, conforme mostra a Figura 9.

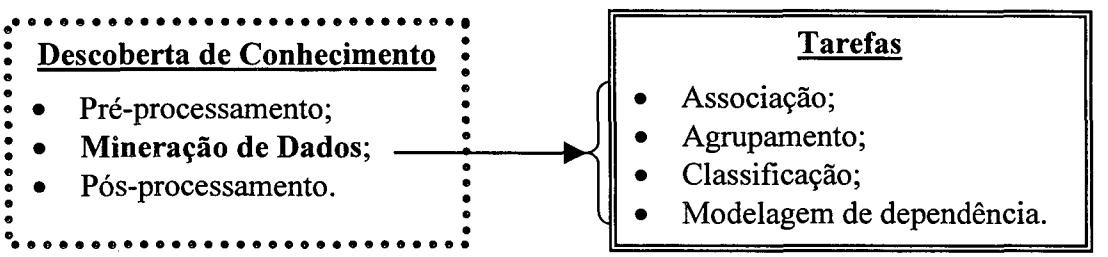


Figura 9: Tarefas de mineração de dados.

Nas seções seguintes apresenta-se uma visão geral de algumas tarefas de mineração de dados e alguns algoritmos empregados nestas tarefas.

3.4 A Tarefa de Associação

Uma das tarefas mais simples e mais conhecidas em MD é conhecida por Extração de Regras de Associação na forma SE X ENTÃO Y, onde X e Y são conjuntos de itens. Exemplo de regra de associação em C&T:

SE <doutor> E <líder_de_grupo> ENTÃO <masculino E idoso>

ou, de forma simplificada:

<doutor>,<líder_de_grupo> ⇒ <masculino, idoso>

A tarefa de associação, na sua forma básica (Agrawal *et al.*,1993), pode ser considerada uma tarefa bem definida e determinística que não envolve

previsão.

> *"Por definição, qualquer algoritmo de associação [na forma básica, padrão dessa tarefa] deve descobrir precisamente o mesmo conjunto de regras, i.e., todas as regras que possuem suporte e confiança maior que um limite especificado pelo usuário, sem exceção. Portanto, todos algoritmos de associação encontram o mesmo conjunto de regras"* (Freitas, 2000a).

Freqüentemente, em grandes bancos de dados que armazenam milhares de itens, como aqueles existentes em redes de supermercados, deseja-se descobrir associações importantes entre os itens comercializados, tal que a presença de alguns deles em uma transação (compra e venda) implique na presença de outros na mesma transação. O objetivo, então, é encontrar todas as regras de associação relevantes entre os itens, do tipo X (antecedente) $\Rightarrow$ Y (conseqüente).

Para tratar desta questão, Agrawal *et al.* (1993) propuseram um modelo matemático, onde as regras de associação devem atender a um *suporte* e *confiança* mínimos especificados pelo usuário. O *suporte* corresponde à freqüência relativa que os padrões ocorrem em toda a base de dados. Exemplo: percentagem de pesquisadores do Diretório que são doutores líderes de grupo (X) e idoso do sexo masculino (Y). A *confiança* é uma medida da força das regras, ou seja, a percentagem de itens de X que possuem Y.

O suporte mínimo (*minsup*) é a fração das transações que satisfaz a união dos itens do conseqüente com os do antecedente, de forma que estejam presentes em pelo menos *s%* das transações no banco de dados. A confiança mínima (*minconf*) garante que ao menos *c%* das transações que satisfaçam o antecedente das regras também satisfaçam o conseqüente das regras.

Um algoritmo de extração de regras de associação, denominado Apriori, foi aplicado à base de dados do Diretório dos Grupos de Pesquisa no Brasil, versão 3.0, a qual foi disponibilizada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para o público em geral.

O trabalho objetivou demonstrar a viabilidade do emprego desta técnica quando aplicada a análise de pesquisadores e grupos de pesquisa,

considerando a possibilidade desta abordagem ser um instrumento valioso no auxílio à gestão e à política de C&T. Os resultados desse trabalho (Romão *et al.*, 1999c) são apresentados no ANEXO A.

3.5 A Tarefa de Agrupamento (*Clustering*)

Quando se trata de aprendizagem de máquina, podemos ter aprendizagem supervisionada ou não-supervisionada. O termo *supervisionado* se refere ao fato da aprendizagem ocorrer em um cenário onde já se conhece a classe. Em caso contrário, quando as classes têm que ser descobertas a partir dos dados, a aprendizagem é dita não-supervisionada ou *agrupamento* (*clustering*).

Agrupamento é uma tarefa onde o algoritmo deve descobrir a classe de cada exemplo baseado na sua similaridade com os demais exemplos do conjunto de dados, dividindo os exemplos em grupos (ou classes).

Através da tarefa de agrupamento pode-se dividir os dados em subconjuntos homogêneos fáceis de descrever e visualizar. Estes por sua vez podem ser mostrados para o usuário em vez de tentar mostrar todos os dados, o que usualmente resultaria na perda de padrões embutidos (Fayyad, 1997).

Um trabalho realizado no contexto de C&T, envolvendo a tarefa de agrupamento, foi desenvolvido por Gonçalves (Gonçalves, 2000), o qual efetuou a extração de conhecimento sobre C&T no Diretório 3.0 (CNPq, 1999).

3.6 Considerações Finais

O crescimento acentuado em número e tamanho dos bancos de dados, principalmente na última década, tem ultrapassado as habilidades humanas de processar as informações acumuladas nestes reservatórios e inviabilizado a utilização de técnicas convencionais para extração de conhecimento (*e.g.*, estatísticas), ocultando conhecimentos valiosos e exigindo o desenvolvimento de novas técnicas capazes de minerar conhecimento relevante em grandes bancos de dados.

Considerando a existência de diversas bases de dados sobre C&T nas diversas agências, é natural procurar explorar estes dados para obter as

informações necessárias. No entanto, é necessário fazer uso de métodos adequados para extração de conhecimentos relevantes e surpreendentes no contexto considerado.

Para selecionar os métodos e algoritmos mais adequados a esta aplicação, realizou-se um estudo sobre KDD (Knowledge Discovery in Databases) (Fayyad, 1996, 1997; Silberschatz & Tuzhilin, 1996) e Mineração de Dados (Graettinger, 1999; Mchugh, 1999; Rendell & Cho, 1990), para descobrir como transformar estas técnicas em subsídios efetivos à gestão de C&T.

No próximo capítulo, discutem-se as tarefas mais importantes de mineração de dados no contexto de descoberta de conhecimento relevante para o planejamento de C&T.

4 DESCOBERTA DE REGRAS DE PREVISÃO

4.1 Introdução

Dentro do processo geral de KDD, descrito no Capítulo 3, os maiores desafios estão na etapa de MD. Freitas (2000b) define MD como "a extração (semi-) automática de conhecimento a partir dos dados". A primeira indagação que surge é: Qual o tipo de conhecimento que se pretende descobrir?

Dentro do contexto considerado, conforme descrito no Capítulo 2, o objetivo é obter conhecimento relevante ao planejamento em C&T. Neste caso, é importante empregar alguma tarefa capaz de extrair regras de previsão.

As tarefas de mineração de dados mais viáveis para obter regras de previsão e responder perguntas de interesse neste segmento são as tarefas de classificação e modelagem de dependência. A tarefa de classificação inclui a essência das duas tarefas e a modelagem de dependência é uma generalização da tarefa de classificação. Portanto, este capítulo concentra-se na apresentação dos principais conceitos e técnicas empregados nas tarefas de classificação (e modelagem de dependência).

4.2 A Tarefa de Classificação

Classificação é uma das tarefas mais referenciadas na literatura de MD e a mais importante para a presente pesquisa. Neste tipo de tarefa, o objetivo é descobrir um relacionamento entre um atributo meta (cujo valor, ou classe, será previsto) e um conjunto de atributos previsores. O sistema deve descobrir este relacionamento a partir de exemplos com classe conhecida. O relacionamento descoberto será usado para prever o valor do atributo meta (ou a classe) para exemplos cujas classes são desconhecidas (Fertig *et al.*, 1999).

Na área de aplicação considerada (Gestão de C&T), pode-se definir classificação como sendo a tarefa de prever corretamente informação sobre a classe de um exemplo da unidade de análise (*i.e., pesquisador, grupo de*

pesquisa, projeto de pesquisa, etc.), a partir de alguns atributos desta unidade de análise, chamados *atributos previsores*, cujos valores são conhecidos. Uma das possibilidades é a descoberta de regras que representem as correlações entre os atributos que definem a unidade de análise (Exemplo considerando "pesquisador" como unidade de análise: *nacionalidade, idade, nível_formação, etc.*).

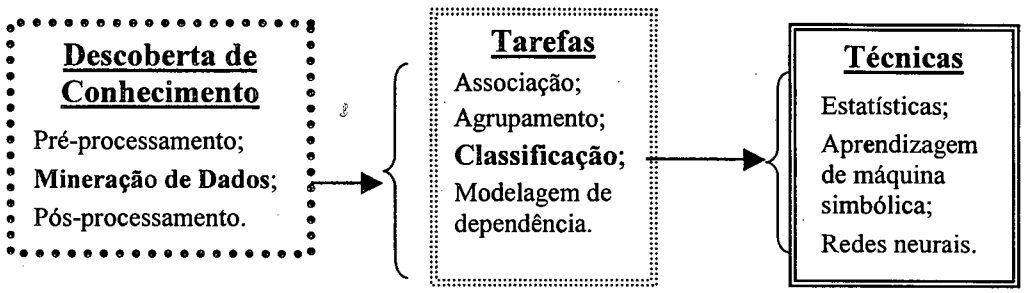


Figura 10: Técnicas de MD para a tarefa de classificação.

A literatura apresenta diversas técnicas de classificação (King *et al.*, 1995; Hand, 1997), conforme mostra a Figura 10. Segundo Michie *et al.* (1994), as principais propostas são originárias de três campos de pesquisa: estatística, aprendizagem de máquina simbólica e redes neurais.

Nesta tese, o interesse é investigar principalmente o campo chamado de *aprendizagem de máquina simbólica*, que abrange procedimentos computacionais automáticos, baseados em operações lógicas ou binárias, capazes de aprender a realizar uma tarefa a partir de uma série de exemplos. Focalizando a tarefa de classificação, muita atenção tem sido dada a técnicas baseadas em árvores de decisão (Quinlan, 1993; etc.). Outras técnicas, tais como AG's (Algoritmos Genéticos) e PLI (Programação Lógica Indutiva) têm sido alvo de mais interesse por parte de pesquisadores recentemente. As técnicas de aprendizagem de máquina simbólica para a tarefa de classificação possuem a vantagem de gerar expressões simples o suficiente para a compreensão humana (Michie *et al.*, 1994).

Cabe ressaltar que nenhum algoritmo é "o melhor" em todas as aplicações. A performance de um algoritmo de classificação depende muito do domínio da aplicação (Freitas, 2000a; Michie et al., 1994; King et al., 1995). Este tópico será discutido em mais detalhes na seção 4.2.2.

4.2.1 Estrutura Geral da Tarefa de Classificação

No escopo desta tese, assume-se que o problema é projetar um algoritmo para ser aplicado em um banco de dados onde as classes são pré-definidas e cada novo dado deve ser associado a uma destas classes. Este processo é conhecido como *reconhecimento de padrões*, *discriminação, aprendizagem supervisionada* ou *classificação.* Na literatura de estatística, a aprendizagem supervisionada usualmente é referenciada como *discriminação.*

Regras de classificação podem ser consideradas uma espécie de regras de previsão onde o antecedente contém uma combinação (tipicamente uma conjunção) de condições envolvendo valores do domínio dos atributos previsores, e o conseqüente contém um valor previsto para o atributo meta (Freitas, 2000b).

Regras do tipo SE... ENTÃO..., também chamadas *regras de produção*, constituem uma forma de representação simbólica e possuem o seguinte formato:

SE <antecedente> ENTÃO <conseqüente>.

O *antecedente* é formado por expressões condicionais envolvendo atributos do domínio da aplicação existentes nos bancos de dados.

O *conseqüente* é formado por uma expressão que indica a previsão de algum valor para um atributo meta, obtido em função dos valores encontrados nos atributos que compõem o antecedente.

Portanto, a tarefa é descobrir regras de classificação capazes de prever o valor de um atributo meta a partir dos valores de atributos previsores. As regras de previsão, portanto, objetivam auxiliar o planejamento de ações futuras.

Como ilustração, considere um banco de dados sobre C&T cujo atributo meta, escolhido por um especialista de alguma agência, indica se uma determinada *região* do país (e.g., sul, sudeste, etc.) é candidata a receber investimentos especificamente para pesquisa, associada às classes "alta" ou "baixa", e considere-se, ainda, que os atributos previsores são: produção científica, idade média e titulação média dos pesquisadores de cada região.

Seguindo esta ilustração, pode-se obter regras, na forma simplificada, como:

(prod_científ. = "alta"), (titulação = "alta") ⇒ Região.Pot_Pesquisa = "alta";

(titulação = "baixa"), (idade = "alta") ⇒ Região.Pot_Pesquisa = "baixa".

Naturalmente a segunda regra indica que a região escolhida não é apta para pesquisa neste momento, mas indicaria uma região apta para receber investimentos em formação ou capacitação.

Comparando com a tarefa de associação, é fácil observar duas diferenças básicas. Uma regra de associação pode conter mais de um item no conseqüente, enquanto que nas regras de classificação isto não é permitido. Segundo, na tarefa de associação qualquer atributo pode aparecer tanto no conseqüente como no antecedente da regra (simetria), enquanto que na classificação o atributo escolhido como meta só pode aparecer no conseqüente e os demais (previsores) apenas no antecedente das regras (assimetria).

Freitas (2000a) considera a tarefa de classificação "uma tarefa de definição difícil e não determinística devido ao fato de envolver previsão", enquanto que a tarefa básica de associação pode ser considerada uma tarefa simples, bem definida e determinística que não envolve previsão.

Freitas utiliza a expressão "definição difícil" para se referir ao fato das regras de classificação serem obtidas utilizando apenas os dados de treinamento, nada garantindo que as regras terão boa exatidão diante dos dados de teste, os quais não foram utilizados durante o treinamento. A expressão "não determinística" se refere ao fato das regras serem obtidas com base em dados passados para prever o futuro, caracterizando uma forma de indução.

4.2.2 Algoritmos de Classificação Convencionais

Existem vários métodos já consolidados na tarefa de classificação. Uma revisão geral sobre este tema pode ser encontrada em Han & Kamber (2000).

Pesquisas comparando diversos algoritmos de classificação convencionais foram publicadas por King *et al.* (1995), descrevendo resultados de um projeto (denominado de Statlog) que compara 17 (dezessete) algoritmos de classificação, incluindo algoritmos de aprendizagem simbólica, estatísticos e RNA, aplicados a problemas do mundo real. King e seus colegas utilizaram 12 (doze) conjuntos de dados: 5 (cinco) de análise de imagens; 3 (três) de

medicina; 2 (dois) de engenharia e 2 (dois) de finanças.

Procurou-se mostrar quais aspectos de um algoritmo o faz ter mais sucesso diante de um conjunto particular de dados ou quais aspectos de um conjunto de dados favorecem a utilização de determinados algoritmos.

King *et al.* (1995) avaliaram 7 (sete) algoritmos de aprendizagem simbólica, sendo cinco chamados de árvores de decisão (C4.5, NewID, AC2, CART/INDCART e Cal5) e dois métodos baseados em regras (CN2 e ITrule); 5 (cinco) da estatística tradicional: k-NN (k-nearest neighbor), Naive Bayes, Discrim, Quadra e LogReg (Logistic Regression); 3 (três) da estatística moderna (SMART, Alloc80 e CASTLE) e 2 (dois) RNA (Redes Neurais Artificiais). Algoritmos genéticos, técnica considerada não convencional naquele trabalho, não foram avaliados.

O desempenho dos algoritmos foi medido considerando aspectos objetivos e subjetivos. O aspecto objetivo incluiu tempo (de treinamento e teste) e a precisão. As medidas subjetivas foram obtidas através de respostas de usuários, leigos e especialistas, quanto à compreensão das regras obtidas, facilidade de uso dos algoritmos e robustez dos parâmetros de entrada requisitados. O principal objetivo do projeto Statlog foi investigar porque certos algoritmos funcionam melhor para determinados tipos de dados.

Quanto à velocidade de aprendizagem, o algoritmo *backprop* (RNA) é mais lento do que algoritmos de árvore de decisão, mas em geral é mais exata do que a aprendizagem simbólica.

Muitos algoritmos necessitam a experiência de um especialista para usá-lo adequadamente, isto é, para ajustar os parâmetros. Idealmente, todos algoritmos deveriam ser automáticos no ajuste destes parâmetros. O usuário deveria apenas alimentar os dados e executar o algoritmo. Entretanto, nem sempre este é o caso. Na visão de King *et al.* (1995), o problema de sintonia de parâmetros é mais grave em RNA e menos agudo para alguns algoritmos simbólicos, mas é um grande problema dos usuários de algoritmos genéticos. Esta hipótese não se confirmou na presente pesquisa utilizando algoritmos genéticos.

Quando a facilidade no entendimento das regras geradas é essencial, os

algoritmos simbólicos (C4.5, NewID, AC2, CN2, ITrule, CART/INDCART e Cal5) apresentam grande vantagem sobre os demais.

ITrule não é um algoritmo apropriado para a tarefa de classificação e não foi projetado para trabalhar com banco de dados de grande porte ou com problemas que apresentam muitas classes. Ele é adequado para extração de regras relevantes isoladas (Michie *et al.*, 1994).

Enfocando a aplicação desta tese, algoritmos de aprendizagem simbólica são métodos adequados em termos de exatidão e facilidade de compreensão. São métodos em geral não paramétricos, ou seja, não assumem qualquer distribuição, e são mais robustos para distribuição extrema de valores de atributos (King *et al.*, 1995).

A maioria dos algoritmos simbólicos trabalha com uma estrutura de dados do tipo árvore. Existe diferença no tamanho das árvores produzidas por diferentes algoritmos, mas isto geralmente não afeta a precisão. Os algoritmos CART (Classification and Regression Tree) e Cal5, em geral, produzem as menores árvores.

Foram avaliados três algoritmos baseados no método ID3: C4.5, NewID e AC2. C4.5 e NewID produzem resultados similares. AC2 é menos preciso e mais lento devido a ser escrito na linguagem LISP, o que provoca *overhead* nas interfaces.

Backpropagation foi o método mais lento analisado. Apresenta dificuldades na sintonia dos parâmetros. Além disso, em geral uma rede neural exige a conversão de atributos lógicos ou categóricos em valores numéricos.

King *et al.* (1995) concluíram que o melhor algoritmo para classificar um conjunto particular de dados depende primordialmente das características dos dados utilizados. Uma das principais conclusões foi que dados com distribuição muito desbalanceada com muitos atributos binários/categóricos favorecem a utilização de algoritmos de aprendizagem simbólica.

Se os dados possuem distribuição acentuada e/ou mais de 38% de atributos categóricos, então os algoritmos simbólicos são uma boa alternativa para maximizar a precisão dos resultados. Se aspectos subjetivos (facilidade de uso e de compreensão humana) forem prioridades, os simbólicos devem ser

escolhidos também.

Em geral regras de decisão são consideradas mais fáceis de entender do que árvores de decisão e ambas são mais fáceis de entender do que coeficientes obtidos por regressão (Michie *et al.*, 1994).

Portanto, considerando os resultados obtidos por King *et al.* (1995), conclui-se que os algoritmos simbólicos são adequados para a tarefa de classificação dos dados sobre C&T, em especial o algoritmo CART. Entretanto, os algoritmos simbólicos tradicionais apresentam algumas limitações, principalmente quando ocorre interação entre atributos, como é o caso nos dados sobre C&T (Romão, 1999c).

Quadro 3: Comparação de algoritmos de classificação

Algoritmo	VP	Custos	Interpr.	Compr.	Param.	User-fr.	Dados
Discrim	N	T	3	4	4	S	N
Quadisc	N	T	2	3	3	S	N
Logdisc	N	T	3	4	4	S	N
SMART	N	AT	1	2	1	N	NC
ALLOC80	N	AT	1	2	2	N	NC
k-NN	N	T	1	5	2	N	N
CASTLE	N	T	3	3	3	S	NC
CART	S	T	5	4	5	S	NC
IndCART	S	T	5	4	5	S	NC
NewID	S	N	5	4	4	S	NC
AC^2	S	N	5	4	4	S	NCH
Baytree	S	T	4	4	5	N	NC
NaiveBay	S	T	3	4	4	S	N
CN2	S	N	5	4	4	S	NC
C4.5	S	N	5	4	4	S	NC
Itrule	N	N	3	4	4	N	NC
Cal5	S	AT	5	4	5	S	NC
Kohonen	N	N	1	1	1	N	N
DIPOL92	N	AT	2	3	2	N	NC
Backprop	N	T	1	3	3	N	N
RBF	N	N	1	1	1	N	N
LVQ	N	N	1	1	1	N	N
Cascade	N	T	1	3	2	N	N

Fonte: Michie *et al.*, 1994, p.215.

Legenda:
VP = se o programa aceita campos com ausência de valores de atributos;
Custos = trabalham com matriz de custos de classificação errônea na (A)prendizagem, (T)este ou (N)enhum;
Interpr = facilidade de interpretação (5 = muito fácil de interpretar);
Compr = facilidade de compreender o princípio do método (5 = muito fácil);
Param = facilidades para seleção de parâmetros importantes (via usuário ou automático) (5 = muito fácil);
User-fr = se é amigável para o usuário;
Dados = dados permitidos (N = numérico, C = Categórico, H = Hierárquico).

O projeto StatLog gerou resultados de pesquisa, na avaliação de algoritmos de classificação, publicados no livro de Michie *et al.* (1994), disponível no endereço da Internet: http://www.amsta.leeds.ac.uk/~charles/statlog/.

Uma visão geral dos resultados das pesquisas de Michie e seus colegas está resumida no Quadro 3 contendo informações sobre diversas características dos algoritmos avaliados por King *et al.* (1995) e também das redes neurais Kohonen e LVQ e do algoritmo híbrido DIPOL92.

4.2.2.1 O Algoritmo J4.8

O algoritmo J4.8 é uma versão modificada do algoritmo C4.5 (Quinlan, 93) que descobre conhecimento na forma de árvore de decisão. O J4.8 é parte de uma ferramenta denominada Weka (Waikato Environment for Knowledlge Analysis), desenvolvida na Universidade de Waikato na Nova Zelância (Witten, 2000), que possui diversos algoritmos de mineração de dados implementados, além de módulos para pré-processamento de dados.

Os principais algoritmos implementados na ferrramenta Weka são algoritmos de classificação (*e.g.*: J4.8), mas possui também algoritmo para extração de regras de associação e algoritmo para resolver a tarefa de agrupamento.

Esta ferramenta, distribuída gratuitamente pela Internet no endereço: http://www.cs.waikato.ac.nz/ml/weka/index.html, foi implementada em Java e está disponível, juntamente com sua documentação, para as plataformas Linux, Macintosh e Windows.

4.3 Avaliação de Regras Descobertas

Independente do tipo de algoritmo utilizado na tarefa de classificação, as regras descobertas devem ser avaliadas. O critério mais empregado é a medição da taxa de acerto destas regras. Existem diversos métodos para isto. Um método simples é:

$$\text{Taxa de acerto} = \frac{\text{N° de instâncias dos dados classificadas corretamente}}{\text{N° de instâncias dos dados aplicáveis}}$$

Este método simples tende a não ser muito eficaz em algumas situações na prática por ignorar situações dos seguintes tipos (Hand, 1997; Michie et al., 1994):

- Em geral, a distribuição de freqüência das classes é desbalanceada, ou seja, as classes mais raras são mais difíceis de serem previstas;
- Custo de classificações errôneas geralmente varia com o tipo do erro de classificação;
- Regras podem ter diferentes custos de aplicação, devido a diferentes custos para medir valores de atributos ocorrendo nas regras. Por exemplo, o custo de erroneamente uma agência de fomento a C&T negar custeio a um projeto viável é bem menor do que o custo de erroneamente conceder custeio a um projeto inviável;

A taxa de acerto descrita acima pode ser calculada tanto sobre os dados de treinamento quanto sobre os dados de teste. Nos dados de treinamento é chamada de taxa de acerto aparente ou de re-substituição, e apresenta sempre um valor superior à taxa de acerto calculada sobre os dados de teste. Este último geralmente envolve dados não considerados na fase de treinamento e fornece uma medida mais precisa da confiabilidade das regras.

4.3.1 Métodos de Validação

4.3.1.1 "*Hold Out*"

Este é um método adequado quando se tem grande quantidade de registros disponíveis. O conjunto dos dados é separado em dois outros subconjuntos: dados para treinamento e dados para teste, conforme mostra a Figura 11.

Os registros do conjunto para treinamento devem ser separados dos registros do conjunto teste de forma aleatória, com uma única possível exceção: manter a mesma proporção de registros de cada classe nos dois conjuntos, o que é denominado de hold-out estratificado. O conjunto para treinamento é utilizado para descobrir as regras de classificação.

O conjunto para teste deve permanecer separado e ser utilizado apenas para validação dos resultados obtidos pelo algoritmo. Eles devem ficar escondidos ("Hold Out") durante a fase de treinamento. Desta forma pode-se obter uma medida da taxa de acerto diante de novos registros.

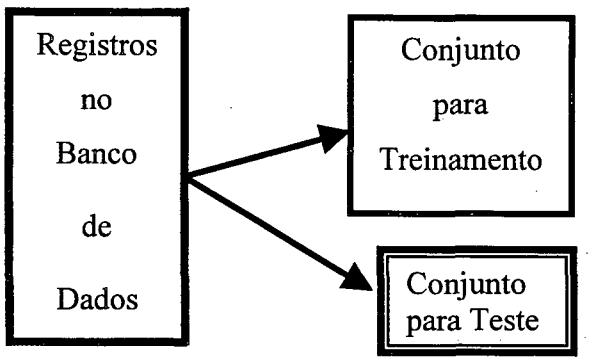


Figura 11: Separação dos Dados.

Este método é simples e apropriado quando se tem abundância de dados. Um conjunto de dados para teste com pelo menos 1000 registros pode ser considerado adequado (Weiss, 1991). Se o conjunto total dos dados disponíveis for suficiente, usualmente divide-se este conjunto em 2/3 para treinamento e 1/3 para teste (Baranauskas, 2001), desde que o conjunto resultante para teste seja de no mínimo aproximadamente 1000 registros. Em caso contrário deve-se utilizar outro método de validação, tal como o método de validação cruzada.

4.3.1.2 Validação Cruzada

Quando há poucos dados disponíveis, pode-se empregar todos os dados na fase de treinamento e realizar validação cruzada para calcular a taxa de acerto na fase de teste.

O método de validação cruzada (Baranauskas, 2001) possui a vantagem de utilizar todos os dados para treinamento e para teste, mantendo a separação entre os dados de treinamento e de teste, mas possui a desvantagem de ser mais caro computacionalmente quando comparado com a forma convencional que utiliza apenas um conjunto de dados de teste.

A validação cruzada consiste em dividir todo o conjunto de dados

disponíveis em k subconjuntos iguais e executar o procedimento do Quadro 4.

Quadro 4: Algoritmo para validação cruzada.

```
Início
Para i = 1 .. k
        Executar o algoritmo de MD usando k-1 subconjuntos de dados;
        Computar a taxa de acerto, das regras obtidas, sobre os dados
           do subconjunto i não utilizado no passo anterior;
Fim Para;
Taxa de acerto = média das taxas de acerto nas k iterações anteriores;
Fim.
```

Para calcular a média e desvio padrão do acerto na validação cruzada pode-se empregar as seguintes equações (Baranauskas, 2001):

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad (4.1)$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} \qquad (4.2)$$

onde:

ā = média;

dp = desvio padrão;

x_i = taxa de acerto da partição i;

k = número de partições dos dados.

As Equações 4.1 e 4.2 podem ser utilizadas também para calcular a média e desvio padrão da cobertura das regras fazendo x_i = cobertura da partição i.

4.3.2 Fator de Confiança

A utilização de um fator de confiança é uma forma simples de se avaliar a precisão das regras obtidas apenas nos dados de treinamento e pode ser calculada pela razão X/Y, onde X é o número de registros que satisfazem o antecedente e o conseqüente da regra e Y é o número total de registros que satisfazem o antecedente da regra.

Entretanto, segundo Freitas (1999), quando uma regra cobre poucos (em torno de cinco) registros, este método apresenta uma limitação. Veja os exemplos abaixo.

Exemplo 1: considere uma regra que cubra apenas um registro corretamente. Neste caso, a confiança nesta regra seria X/Y = 100%, resultado indesejável uma vez que a regra está muito especializada (*overfitting*).

"Overfitting" ocorre quando as regras se ajustam demais a peculiaridades dos dados de treinamento, o que tende a reduzir a taxa de acerto em dados de teste não vistos anteriormente.

Para resolver este problema Freitas utiliza um fator –1/2 no numerador da expressão de cálculo da confiança na regra, baseado na proposta de Quinlan (1987), isto é:

$$\text{Fator de Confiança} = (X - ½)/Y \qquad (4.3)$$

Utilizando esta abordagem, o cálculo da confiança no exemplo 1, aplicando a Equação 4.3, fica:

$$(X - ½)/Y = (1 - ½)/1 = 50\%$$

Exemplo 2: considere outro caso onde uma regra cubra 100 registros corretamente. Neste caso, o cálculo da confiança utilizando a Equação 4.3 seria:

$$(X - ½)/Y = (100 - ½)/100 = 99{,}5\%$$

Comparando a confiança obtida no primeiro exemplo (50%) com a confiança obtida no segundo exemplo (99.5%), fica fácil observar que o fator –1/2 penaliza as regras que cobrem poucos registros mas não penaliza de forma significativa as regras que cobrem muitos exemplos, permitindo a distinção entre regras especializadas e regras generalizadas, desvalorizando as primeiras.

4.3.3 Matriz de Confusão

Há situações onde métodos simples para avaliação dos resultados são

insuficientes. Como ilustração, considere o caso de um sistema automático, para auxiliar no diagnóstico médico, capaz de prever se uma pessoa está doente ou não. Se a previsão for de que uma pessoa não está com uma determinada doença quando na verdade ela está (decisão negativa falsa) é um erro de previsão considerado mais sério do que prever que uma pessoa está com a doença quando na verdade ela está saudável (decisão positiva falsa). Neste último erro o paciente no máximo tomaria um remédio desnecessário enquanto que no primeiro (negativo falso) dependendo da doença poderia proporcionar a morte do paciente por não tomar as medidas necessárias para cura.

Para distinguir entre os dois tipos de erro utiliza-se uma matriz de confusão (Freitas, 2002a). Esta é uma alternativa para atender o critério de qualidade: construir uma matriz de confusão N x N (um tipo de tabela de contingência para análise de previsões corretas e errôneas) com domínio composto por duas classes, denominadas aqui de "sim" e "não", conforme mostra a matriz de confusão do Quadro 5.

Quadro 5: Matriz de confusão.

		Classe real do exemplo	
		Sim	**Não**
Classe prevista pelo método	**Sim**	Soma classificações "sim" corretas (SC)	Soma classificações "sim" erradas (SE)
	Não	Soma classificações "não" erradas (NE)	Soma classificações "não" corretas (NC)

O algoritmo de classificação deve ser executado sobre os dados de treinamento e o resultado da classificação dividido em quatro categorias:

- Classificações "sim" realizadas corretamente (SC);
- Classificações "sim" realizadas erradamente (SE);
- Classificações "não" realizadas erradamente (NE).
- Classificações "não" realizadas corretamente (NC);

As variáveis SC, SE, NE e NC serão utilizadas como contadores internos da matriz de confusão. O contador SC, por exemplo, indicará a quantidade de exemplos "sim" que foram classificados corretamente como exemplos "sim",

enquanto que o contador SE indicará a quantidade de exemplos "não" que foram classificados erroneamente como "sim".

Quanto mais coberturas corretas (SC e NC maiores) e menos erradas (SE e NE menores), maior será a precisão da regra.

A qualidade da regra pode então ser calculada por uma das seguintes equações (4.4 a 4.6):

$$QUALIDADE = \frac{SC}{SC + NE} * \frac{NC}{NC + SE} \quad (4.4)$$

$$QUALIDADE = \frac{SC}{SC + SE} * \frac{SC}{SC + NE} \quad (4.5)$$

$$QUALIDADE = \frac{SC - 1/2}{SC + SE} \quad (4.6)$$

A Equação 4.6 é uma adaptação da equação 4.3, conhecida como Fator de Confiança, visto na seção 4.3.2.

A Equação 4.5 é conhecida como Precision*Recall (Hand, 1997; Freitas, 2002a).

Na Equação 4.4 o primeiro termo [SC/(SC + NE)] é chamado de "sensibilidade" ou "completeza de positivos" ou "taxa de acerto na classe positiva". O segundo termo [NC/(NC + SE)] é chamado de "especificidade" ou "completeza de negativos" ou "taxa de acerto na classe negativa". Os dois termos da equação são multiplicados para forçar a descoberta de regras que tenham alta sensibilidade e alta especificidade (Carvalho & Freitas, 2000).

4.3.4 Avaliação da Compreensibilidade das Regras

Outro aspecto importante na avaliação de regras de classificação é a compreensibilidade das regras descobertas. Em geral considera-se que, quanto menor o número de regras descobertas e menor o número de condições por regra, maior a compreensibilidade do conjunto de regras. No entanto, esse critério é puramente sintático e objetivo, ignorando aspectos semânticos e subjetivos das regras.

Na concepção de Pazzani (2000), "não há nenhum estudo que mostre que pessoas acham modelos menores mais compreensíveis, ou que o tamanho de

um modelo é o único fator que determina sua compreensibilidade".

Muitos algoritmos de mineração de dados possuem parâmetros para controlar o número e tamanho das regras obtidas que indiretamente influenciam a compreensibilidade, mas existem outros fatores tais como:

- ❖ Familiaridade com sistemas de representação de conhecimento;
- ❖ Preferência por modelos consistentes com conhecimento prévio.

O segundo fator aumenta a compreensibilidade mas diminui o grau de surpresa/novidade. Regras surpreendentes também é um aspecto importante na avaliação das regras descobertas e um assunto pouco explorado, tema da seção 4.5.

4.4 A Tarefa de Modelagem de Dependência

Esta é uma tarefa de MD que funciona como uma generalização da tarefa de classificação, ou seja, o objetivo também é descobrir regras de previsão, mas estas regras não estão restritas a apenas um atributo meta. Vários atributos metas podem ser escolhidos, implicando que regras diferentes podem ter atributos diferentes no conseqüente. Na versão desta tarefa adotada neste trabalho, análoga à tarefa de classificação, a quantidade de atributos no conseqüente de uma dada regra será sempre unitária.

Exemplo de regras:

(UF = "RS"), (nível_formação = "alto") ⇒ art_per_nac = "alto";

(idade = "médio"), (art_per_nac = "médio") ⇒ art_per_int = "baixo".

Na tarefa de modelagem de dependência, qualquer atributo pode aparecer tanto no antecedente como no conseqüente das regras, como exemplificado acima com o atributo 'art_per_nac' (artigos publicados em periódicos nacionais), mas não pode aparecer em ambos (antecedente e conseqüente) na mesma regra.

Na tarefa de modelagem de dependência, o espaço de busca é muito maior do que na tarefa de classificação (Noda *et al.*, 1999).

Em algumas situações pode-se restringir o uso de certos atributos apenas ao antecedente ou ao conseqüente das regras. No exemplo acima, pode-se especificar que o atributo 'art_per_nac' poderá ocorrer apenas no conseqüente,

ou que o atributo *idade* poderá ocorrer apenas no antecedente das regras.

Comparando com a tarefa de associação (seção 3.4), aquela permitia qualquer atributo aparecer tanto no antecedente como no conseqüente das regras. Na forma de modelagem de dependência adotada neste trabalho, apenas alguns atributos especificados pelo usuário podem aparecer no conseqüente das regras. De qualquer modo, quando um atributo ocorre no conseqüente, naturalmente ele não pode ocorrer no antecedente da mesma regra como mencionado anteriormente.

Cabe ressaltar que para descobrir regras de previsão é necessário utilizar classificação ou modelagem de dependência – ou outra tarefa relacionada à aprendizagem de máquina - mas não a tarefa de associação padrão (Freitas, 2000a).

O conhecimento extraído deve ser exato e compreensível, mas além disso deve-se utilizar técnicas (em geral pós-processamento) capazes de selecionar regras que sejam relevantes e surpreendentes, assuntos da próxima seção.

4.5 Descoberta de Conhecimento Relevante

Um sistema de KDD pode gerar conhecimento exato, compreensível e de forma eficiente mas que não é do interesse do usuário. É recomendável utilizar algum método que meça o grau de interesse do conhecimento descoberto.

Existem muitos métodos para descoberta de conhecimento que encontram um grande número de regras, mas a maioria destas regras não é relevante para o usuário. Além disso, o grande número de regras dificulta a análise pelo usuário para encontrar as regras relevantes.

É comum primeiro encontrar um grande conjunto de regras e depois extrair um subconjunto bem menor de regras desse conjunto, com o auxílio do usuário, para obter as regras relevantes, caracterizando um pós-processamento das regras descobertas.

Existem métodos objetivos para descoberta de conhecimento relevante, que em geral trabalham de forma autônoma, e métodos subjetivos os quais levam em conta o conhecimento prévio do usuário sobre o domínio da aplicação.

4.5.1 Métodos Objetivos para Descoberta

Existem várias formas objetivas para medir o grau de interesse no conhecimento (Major & Mangano, 1993; Silberschatz & Tuzhilin, 1996), baseadas nos dados.

Uma forma objetiva de avaliação de regras é a confiança e suporte das regras obtidos na tarefa de associação (seção 3.4).

Para a tarefa de classificação, uma alternativa é empregar uma medida baseada em teoria da informação, proposta por Freitas (1998). Nesta proposta, primeiramente, durante o pré-processamento, calcula-se o ganho de informação de cada atributo (Cover & Thomas, 1991). Em seguida calcula-se o grau de interesse na regra, na versão normalizada (Noda *et al.*, 1999), dado por:

$$INTERESSE = 1 - \frac{\sum_{i=1}^{n} Ganho(A_i)}{\log_2(|Dom(G_k)|)} \qquad (4.7)$$

onde:

n = número de atributos do antecedente;

$Dom(G_k)$ = é o domínio do atributo meta G_k;

$|Dom(G_k)|$ = número de valores em $Dom(G_k)$;

$Ganho(A_i) = Info(G_k) - Info(G_k|A_i)$;

$$Info(G_k) = -\sum_{i=1}^{mk}(\Pr(V_{kl})\log_2(\Pr(V_{kl})));$$

$$Info(G_k \mid A_i) = \sum_{i=1}^{n_i}\left\{\Pr(V_{ij})\left[-\sum_{j=1}^{mk}\Pr(V_{kl} \mid V_{ij})\log_2(\Pr(V_{kl} \mid V_{ij}))\right]\right\};$$

mk = número de valores possíveis do atributo meta;

n_i = número de valores possíveis para o atributo A_i;

V_{kl} = valor de cada atributo meta;

V_{ij} = valor de cada atributo A_i;

Pr(X) = probabilidade de X;

Pr(X|Y) = probabilidade condicional de X dado Y.

O interesse na regra, dado pela Equação 4.7, é maior quanto menor a média do ganho de informação dos atributos do antecedente da regra, ou seja, regras constituídas por atributos que apresentam baixo ganho de informação surgem como surpreendentes, uma vez que regras que possuem atributos com alto ganho de informação são intuitivamente previsíveis e evidentes.

Um atributo isolado pode obter baixo ganho de informação e não ser relevante. No entanto, a interação entre atributos pode transformar um atributo irrelevante em relevante, e este fenômeno está intuitivamente associado com o grau de interesse na regra (Noda *et al.,* 1999).

Portanto, dado um "bom" conjunto de regras (exatas, compreensíveis, etc.), aquelas regras que apresentarem um baixo ganho de informação tendem a ser mais relevantes (ou surpreendentes).

Entretanto, para capturar toda a complexidade do processo de descoberta de conhecimento, as medidas objetivas do grau de interesse não são suficientes. São necessárias medidas subjetivas, as quais dependem não só dos dados mas também do usuário.

4.5.2 Métodos Subjetivos para Descoberta

O interesse em uma regra depende do usuário, do conhecimento que o mesmo tem do domínio de aplicação, mas não é fácil identificar regras relevantes a partir de um grande conjunto de regras descobertas.

Uma regra pode ser relevante para um usuário e considerada inútil para outro. Portanto, o interesse em uma regra pode ser considerado essencialmente subjetivo, uma vez que depende dos conceitos atuais que o usuário tem a respeito do domínio, e também de seus interesses. Em MD, grau de interesse subjetivo é considerado um problema importante (Liu *et al.*, 1997).

Silberschatz & Tuzhilin (1996) identificaram duas razões para um padrão ser relevante do ponto de vista subjetivo: deve ser inesperado (causar surpresa ao

usuário) e induzir a alguma ação, este último será chamado de acionável a seguir.

Liu *et al.* (1997) consideram uma regra acionável "se o usuário puder fazer alguma coisa com ela a seu favor" e, inesperada "se ela causar surpresa para o usuário".

Para exemplificar, considere uma avaliação comparativa entre dois grupos de pesquisa onde um grupo possui dez integrantes (Grupo10) e o outro grupo apenas dois (Grupo2), onde os insumos sejam proporcionais ao n.º de integrantes. Suponha ainda um conhecimento descoberto sobre produção: Pgrupo10 < Pgrupo2. Este conhecimento descoberto apresenta dois aspectos significativos: é inesperado e induz a alguma ação por parte do usuário (acionável).

O conhecimento gerado é inesperado porque naturalmente a crença é que um grupo contendo 10 (dez) integrantes deve produzir mais itens do que um grupo com apenas dois integrantes. Além disso, este resultado permitiria que o usuário (possivelmente um analista de C&T de alguma agência) tomasse alguma medida (acionável) para motivar o Grupo10 a aumentar sua produção.

O conceito acionável é muito importante para tomada de decisão, mas difícil de ser medido, enquanto que o conceito inesperado é, em geral, um pouco mais fácil de ser obtido. Silberschatz & Tuzhilin realizaram estudos sobre a relação entre inesperado e acionável e concluíram que a maioria do conhecimento inesperado é acionável também, e vice-versa. Logo, uma forma de medir o grau de acionável é através da medição do grau de surpresa, servindo como uma aproximação. Conclui-se que, se um conhecimento é inesperado, então ele será provavelmente acionável e, portanto, relevante.

Segundo Silberschatz & Tuzhilin (1996), um conhecimento é considerado inesperado quando ele afeta as crenças do usuário de forma significativa. As crenças podem ser firmes ou leves.

As crenças firmes não mudam diante de novas descobertas e servem apenas para confirmar a qualidade dos dados e das descobertas.

Nas crenças leves o usuário está disposto a mudar suas crenças diante de novas evidências, permitindo a aplicação do conceito de acionável.

Assim, é possível definir o interesse no conhecimento obtido em função do grau de influência no sistema de crenças.

Silberschatz & Tuzhilin (1996) apresentaram um método indicado para aplicações onde novos dados são adicionados periodicamente sobre dados coletados anteriormente (*e.g.*, Currículos):

> *"Quando novos dados chegam, todos os graus de crenças são revisados. Se algum dos graus mudou acima de um limite predeterminado, isto significa que há algum padrão relevante nos dados e que o processo de descoberta para extrair padrões relevantes deve ser executado"* (Silberschatz & Tuzhilin, 1996).

Neste método, uma vez obtido um conjunto de regras relevantes, a rotina de MD só será executada novamente quando houver necessidade, ou seja, quando houver alguma mudança nas crenças indicando chegada de conhecimento relevante.

4.5.3 Técnica Subjetiva de Comparação Difusa

Liu & Hsu (1996) propuseram uma técnica de pós-análise de regras de classificação, geradas pelo algoritmo C4.5, baseada na comparação das mesmas com alguns conceitos prévios ou conhecimento a respeito do domínio de aplicação por parte do usuário.

A principal motivação é que, quando o número de regras geradas é grande, fica difícil para o usuário analisar todas estas regras. Além disso, em sistemas dinâmicos, onde as regras podem mudar constantemente, é importante saber o que mudou desde a última aprendizagem. Segundo os autores, estes aspectos da pesquisa foram ignorados no passado, motivando o desenvolvimento de uma técnica de pós-análise.

Os autores propõem uma técnica de comparação difusa para realizar a pós-análise de regras. Nesta técnica, as regras existentes (conhecimento prévio) são consideradas como regras difusas e são representadas utilizando a teoria dos conjuntos difusos. As regras novas são comparadas com as regras difusas existentes utilizando técnicas difusas.

O conhecimento prévio do usuário é representado por E, e o conjunto de regras descobertas é representado por B.

Duas regras, uma de B e outra de E, são consideradas similares se ambas as partes (antecedente e conseqüente) das regras forem similares.

Uma regra B_i e uma regra E_j são consideras diferentes entre si se pelo menos uma parte (antecedente ou conseqüente) de uma regra for diferente da outra. Aqui podem surgir duas situações: conseqüente inesperado e/ou condições (do antecedente) inesperadas.

A primeira situação (conseqüente inesperado) ocorre quando a parte condicional de B_i e E_j são similares, mas os conseqüentes são diferentes.

Na segunda situação (condições inesperadas) os conseqüentes de B_i e E_j são similares mas as partes condicionais das duas regras são diferentes. Nesta situação surgem dois casos: condições contraditórias e condições inesperadas.

Condições contraditórias: há vários atributos semelhantes nas partes condicionais das regras B_i e E_j, mas os valores destes atributos são completamente diferentes.

Condições inesperadas: os atributos nas partes condicionais de B_i e E_j são diferentes.

Em resumo, esta técnica consiste em três passos:

- converter cada regra de E em uma regra difusa;
- comparar cada regra B_i com E, e obter o grau de similaridade para cada regra em B;
- ordenar B em função do grau de similaridade.

4.5.3.1 Computando a Similaridade entre Regras

Para computar a similaridade da regra, Liu & Hsu (1996) definiram W_i como o grau de similaridade entre uma regra descoberta B_i e o conjunto de regras (E) especificado pelo usuário. O cálculo de W_i compreende dois passos: cálculo da similaridade do nome do atributo e cálculo da similaridade de valor do atributo.

A similaridade dos nomes dos atributos do antecedente, L, é calculada comparando os nomes dos atributos em B_i com os atributos em E_j para obter

$A_{(i,j)} = F_i \cap H_j$ (atributos que ocorrem tanto em B_i quanto em E_j), conforme a Equação 4.8.

$$L_{(i,j)} = \frac{|A_{(i,j)}|}{\max(|F_i|,|H_j|)} \tag{4.8}$$

onde: $|F_i|$ = n.º de atributos em B_i;

$|H_j|$ = n.º de atributos em E_j;

$|A_{(i,j)}|$ = tamanho do conjunto $A_{(i,j)}$.

Após o cálculo da similaridade dos nomes dos atributos é feito o cálculo da similaridade entre os valores dos atributos. São considerados dois casos: similaridade entre atributos discretos e similaridade entre atributos contínuos.

Nesta técnica, para calcular a similaridade entre valores de atributos discretos (categóricos), o usuário deve especificar o universo de discurso, cada termo lingüístico e ainda fornecer o grau de pertinência de cada termo, em cada hipótese (Impressão Geral).

Exemplo: (Idade = 'alta') ⇒ ed_public_livro_nac = 'alto'.

Neste exemplo o usuário teria que especificar que o atributo *Idade* pode assumir os valores lingüísticos {baixo, médio, alto} e o significado de cada termo lingüístico.

Para calcular a similaridade entre valores de atributos contínuos, Liu & Hsu (1996) utilizaram conjuntos difusos na forma trapezoidal com distribuição de vértices arbitráveis. Por exemplo, na hipótese do usuário:

(Idade = 'baixa') ⇒ projeto = 'reprovado'.

O termo 'baixa' terá que ser definido pelo usuário. Supondo que Idade = [0.. 80], o usuário terá que fornecer quatro pontos utilizando valores neste intervalo (*e.g.:* a=15, b=20, c=30 e d=35).

A similaridade entre regras pode ser calculada pela Equação 4.9.

$$w_{(i,j)} = \begin{cases} \dfrac{Z_{(i,j)} \cdot L_{(i,j)} \cdot \sum_{k \in A_{(i,j)}} V_{(i,j)k}}{|A_{(i,j)}|}, & |A_{(i,j)}| \neq 0 \\ 0, & |A_{(i,j)}| = 0 \end{cases} \tag{4.9}$$

onde:

$Z_{(i,j)}$ é o grau de similaridade do conseqüente;

$L_{(i,j)}$ é a similaridade dos nomes dos atributos nos antecedentes de B_i e E_j;

$V_{(i,j)k}$ é o grau de similaridade dos valores do k-ésimo atributo de $A_{(i,j)}$;

$|A_{(i,j)}|$ é o tamanho do conjunto de atributos que ocorrem tanto no antecedente de uma regra descoberta (B_i) como no antecedente de uma regra E_j especificada pelo usuário.

As fórmulas usadas para cálculo de $V_{(i,j)k}$ podem ser encontradas em Liu & Hsu (1996). A Figura 12 ilustra as comparações relativas ao cálculo de $W_{(i,j)}$.

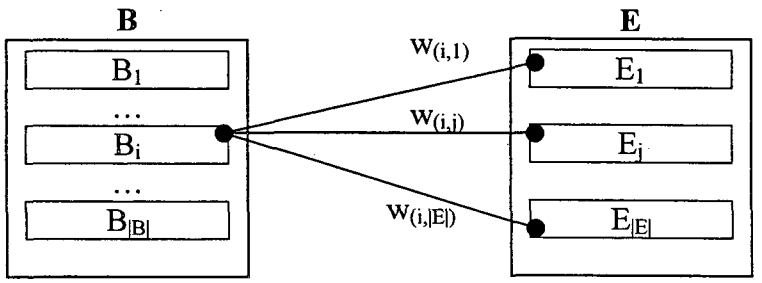


Figura 12: Cálculo da similaridade $w_{(i,j)}$.

Em seguida, pode-se então calcular o grau de similaridade, W_i, entre uma regra nova específica B_i e um conjunto de regras existentes E, dado por:

$$W_i = \max(w_{(i,1)}, w_{(i,2)}, \ldots, w_{(i,j)}, \ldots, w_{(i,|E|)}) \qquad (4.10)$$

4.5.3.2 Computando a Diferença entre Regras

Uma regra pode diferir de uma hipótese (ou regra previamente descoberta) em dois sentidos diferentes: conseqüentes inesperados e/ou condições inesperadas.

O conseqüente inesperado ocorre quando o antecedente da regra nova é similar ao antecedente da hipótese mas o conseqüente é diferente. Neste caso, a similaridade $w_{(i,j)}$ é calculada conforme a Equação 4.11.

$$w_{(i,j)} = \begin{cases} \dfrac{L_{(i,j)} \cdot \sum\limits_{k \in A_{(i,j)}} V_{(i,j)k}}{|A_{(i,j)}|} - (Z_{(i,j)} - 1), & |A_{(i,j)}| \neq 0 \\ -Z_{(i,j)}, & |A_{(i,j)}| = 0 \end{cases} \qquad (4.11)$$

Condições inesperadas ocorre quando o conseqüente é similar mas as condições são diferentes. Podem ocorrer duas situações: condições contraditórias ou condições imprevistas.

Nas condições contraditórias as regras com $|A_{(i,j)}|>0$ são priorizadas e o cálculo de $w_{(i,j)}$ é conforme a Equação 4.12.

$$w_{(i,j)} = \begin{cases} Z_{(i,j)} - L_{(i,j)} \cdot \dfrac{\sum_{k \in A_{(i,j)}} V_{(i,j)k}}{|A_{(i,j)}|} - 1, & |A_{(i,j)}| \neq 0 \\ Z_{(i,j)}, & |A_{(i,j)}| = 0 \end{cases} \quad (4.12)$$

Neste caso, o cálculo de W_i é semelhante Equação 4.10.

Nas condições imprevistas as regras com $|A_{(i,j)}|=0$ são priorizadas e o cálculo de $w_{(i,j)}$ é feito em duas partes, conforme as equações 4.13 e 4.14.

$$wa_{(i,j)} = \begin{cases} Z_{(i,j)} - L_{(i,j)} \cdot \left[\dfrac{\sum_{k \in A_{(i,j)}} V_{(i,j)k}}{|A_{(i,j)}|} + 1 \right], & |A_{(i,j)}| \neq 0 \\ Z_{(i,j)}, & |A_{(i,j)}| = 0 \end{cases} \quad (4.13)$$

$$wb_{(i,j)} = \begin{cases} Z_{(i,j)} \cdot L_{(i,j)} \cdot \left[\dfrac{\sum_{k \in A_{(i,j)}} V_{(i,j)k}}{|A_{(i,j)}|} + 1 \right], & |A_{(i,j)}| \neq 0 \\ 0, & |A_{(i,j)}| = 0 \end{cases} \quad (4.14)$$

Neste caso (regras diferentes), W_i é calculado da seguinte forma:

$W_i = \max(w_{a(i,1)}, w_{a(i,2)}, \ldots, w_{a(i,j)}, \ldots, w_{a(i,|E|)}) - \max(w_{b(i,1)}, w_{b(i,2)}, \ldots, w_{b(i,j)}, \ldots, w_{b(i,|E|)})$ (4.15)

Este método pode ser resumido no seguinte:

- Primeiramente o algoritmo C4.5 é executado para gerar regras;
- O usuário fornece as hipóteses e a distribuição (coordenadas) dos conjuntos difusos;
- O sistema encontra as regras similares;
- O sistema encontra regras considerando três critérios:
  - condições inesperadas;
  - condições contraditórias;

- conseqüente contraditório;

- Ordena as regras de acordo com os diferentes critérios.

4.5.4 Técnica Subjetiva Baseada em Impressões Gerais

Liu *et al.* (1997) propuseram uma técnica para analisar regras descobertas baseado em um tipo específico de conhecimento prévio que o usuário tem a respeito do domínio da aplicação, o qual eles chamam de Impressões Gerais (IG's).

Muitas vezes o usuário não sabe detalhes a respeito do domínio, como alguns sistemas especialistas exigem, mas normalmente ele sempre tem algumas IG's. Por exemplo, em uma agência de C&T um analista pode pensar que um grupo de pesquisa, onde a maioria de seus integrantes tenha titulação alta, deve ter boas chances de obter custeio para projetos.

Os autores propuseram uma forma para especificar as IG's e dois algoritmos de comparação para analisar as regras descobertas. As regras geradas são comparadas uma a uma com as impressões gerais e o resultado da comparação fornece as regras inesperadas, ou seja, regras que contrariam as impressões iniciais do usuário. As regras inesperadas são consideradas relevantes.

Um termo de uma IG é dado por A.Op, onde A é um atributo e Op é um dos operadores {<, >, <<, |, [conjunto]}. Os operadores são definidos conforme descrito no Quadro 6.

Exemplos de IG's:

Titulação > → aprova_projeto;

Idade | → {titulação, produção_científica};

Titulação >, idade << → produção_científica;

Titulação >, estrangeiro [sim] → fala_inglês.

A técnica de Liu *et al.* (1997) pode ser resumida em dois passos:

a) O usuário especifica todas as IG's que ele tem a respeito do domínio usando a linguagem de especificação dada acima (pode-se facilitar o trabalho do usuário com uma interface que utilize termos lingüísticos);

b) O sistema analisa as regras descobertas comparando com as IG's para determinar diversos tipos de regras relevantes. As regras são, então, ordenadas de acordo com os resultados da comparação.

Quadro 6: Operadores definidos em Liu *et al.* (1997).

IG	DIGNIFICADO	EXEMPLO
$A <\rightarrow C_j$	Quanto menor o valor do atributo A, maior a probabilidade de pertencer a uma classe específica C_j.	Quanto menor a titulação do pesquisador, maior a chance dele pertencer à classe de pesquisadores com baixa produção científica.
$A >\rightarrow C_j$	Quanto maior o valor de A, maior a chance de pertencer à classe C_j.	Quanto maior a titulação, maior a chance de obter aprovação.
$A << C_j$	Se A estiver dentro de certo intervalo, então a classe C_j é a mais provável.	Se pesquisador não é nem muito jovem nem muito velho, então conceder bolsa.
$A \mid \rightarrow C_{sub}$	Há algum relacionamento (não exato) entre o atributo A e as classes em C_{sub}.	O tempo de formação do grupo de pesquisa pode determinar a viabilidade de um projeto, ou seja, tempo_formação $\mid \rightarrow$ {viável, inviável}.
$A[S] \rightarrow C_j$	Se A é um elemento do conjunto S, então é mais provável pertencer à classe C_j, onde S é um subconjunto dos valores possíveis do atributo A.	Estrangeiro [sim] $\rightarrow$ alta_produtividade.

Nesta técnica não é considerado se as regras são consistentes, se há interação entre elas, etc. Assume-se que estas análises foram feitas pelo algoritmo de classificação, que no caso foi o C4.5.

O primeiro passo descrito acima (a), é simples e depende do usuário.

O segundo passo (b) inclui a comparação das regras descobertas com as IG's. A forma como esta comparação é feita depende do tipo das regras, que são três:

- regras equivalentes;
- regras com conclusão (conseqüente) inesperada; e
- regras com condições (antecedente) inesperadas.

Regras equivalentes: Uma regra é considerada equivalente a uma IG se as

condições e a conclusão da regra são iguais às condições e à conclusão desta IG. Para isso, é necessário comparar cada regra descoberta com as IG's que levam à mesma conclusão da regra analisada.

Durante a comparação surgem dois casos:

1. Todos os atributos de uma regra aparecem em uma IG, ou seja, não há atributos inesperados;
2. Um subconjunto de atributos da regra não aparece na IG, ou seja, há alguns atributos inesperados.

Todas as regras do caso (1) são ordenadas de acordo com um grau de equivalência (semelhança) com as IG's. Nas regras do caso (2), todas as condições que não aparecem na IG correspondente são removidas. As regras resultantes são ordenadas da mesma forma que as regras do caso (1), mas em um conjunto separado.

Às regras do caso (2) é imposto ainda um esquema de dois níveis: primeiramente cada regra é dividida de acordo com o número de atributos inesperados. Em seguida, dentro de cada partição, as regras são ordenadas como no caso (1).

Regras com conclusão inesperada: as condições da regra são equivalentes às condições de algumas IG's, mas a conclusão é diferente. Exemplo: (Titulação = 'alta', Prod_científica = 'alta' → Não_aprovar_projeto).

Regras com condições inesperadas: não há IG que inclua as condições da regra. Exemplo: (UF = 'SC', Sexo = 'F' ⇒ Aprovar_projeto).

Liu *et al.* (1997) consideraram "improvável obter uma ordenação ótima devido à natureza subjetiva do grau de interesse" e não apresentaram comparações com outras técnicas que realizem este tipo de tarefa.

Muitos sistemas de indução de regras de classificação utilizam conhecimento do domínio no processo de descoberta, mas nestes casos o propósito é produzir regras mais exatas ou melhorar a clareza das regras. No trabalho de Liu, o objetivo é ajudar o usuário a analisar as regras descobertas e encontrar aquelas mais relevantes.

> "*Esta técnica de pós-processamento encoraja a análise interativa e iterativa das regras descobertas, mas a abordagem iterativa não é viável*

porque o processo de descoberta de conhecimento normalmente é de alto custo computacional" (Liu *et al.*, 1997).

Liu e seus colegas concluíram que esta técnica é útil para resolver o problema de grau de interesse em regras. Como tipicamente o número de IG's é pequeno, o custo computacional tende a ser baixo.

Técnicas de pós-processamento de regras descobertas em geral demandam alto custo computacional, tanto na primeira como na segunda etapa. Além disso, muitas das regras obtidas na primeira etapa são descartadas na segunda etapa, com desperdício de processamento.

Neste trabalho propõe-se utilizar uma abordagem capaz de descobrir "pepitas" de conhecimento diretamente durante a descoberta das regras, evitando a necessidade de pós-processamento.

4.6 Considerações Finais

Entre as diversas tarefas de MD há algoritmos bem definidos e determinísticos para encontrar regras de associação, mas para realizar previsão e indução de regras, que é o objetivo nesta pesquisa, geralmente é necessário realizar a tarefa de classificação (ou de modelagem de dependência) utilizando técnicas mais avançadas de MD.

A performance de muitos algoritmos pode ser melhorada através da remoção de atributos irrelevantes, tarefa que pode ser realizada manualmente ou através de algum método automático de seleção dos atributos.

Foram revisados diversos métodos de avaliação da qualidade das regras de previsão descobertas. Destaca-se o método da matriz de confusão, o qual permite a avaliação quantitativa de cada regra de previsão descoberta a partir da execução do algoritmo de MD sobre os dados de treinamento, avaliando o acerto tanto na classe de positivos como de negativos.

Entre as diversas tarefas de MD, uma alternativa menos explorada é a tarefa de modelagem de dependência que se caracteriza como uma generalização da tarefa de classificação, já que em modelagem de dependência pode haver vários atributos metas. A maioria dos conceitos

empregados na tarefa de classificação é igualmente útil à tarefa de modelagem de dependência.

Foram apresentados alguns métodos de pós-processamento para extração de conhecimento relevante a partir do conjunto de regras de previsão obtidas com algoritmos de MD, abordando tanto métodos objetivos como subjetivos, destacando-se o último.

Os métodos subjetivos avaliados são importantes na presente pesquisa, mas inadequados na aplicação considerada, exigindo-se a adaptação dos mesmos para aplicação no contexto de C&T.

No próximo capítulo é apresentada uma revisão bibliográfica sobre algumas abordagens híbridas genético-difusas visando obter subsídios para se propor um novo algoritmo genético para descoberta de regras difusas.

5 SISTEMAS HÍBRIDOS GENÉTICO-DIFUSOS

5.1 Introdução

Para resolver o problema de descoberta de regras de previsão uma boa alternativa está no uso de técnicas de aprendizagem de máquina simbólica, conforme descrito no Capítulo 4. Neste contexto existem os algoritmos evolucionários e os conjuntos difusos que podem ser empregados nesta missão, conforme mostra a Figura 13.

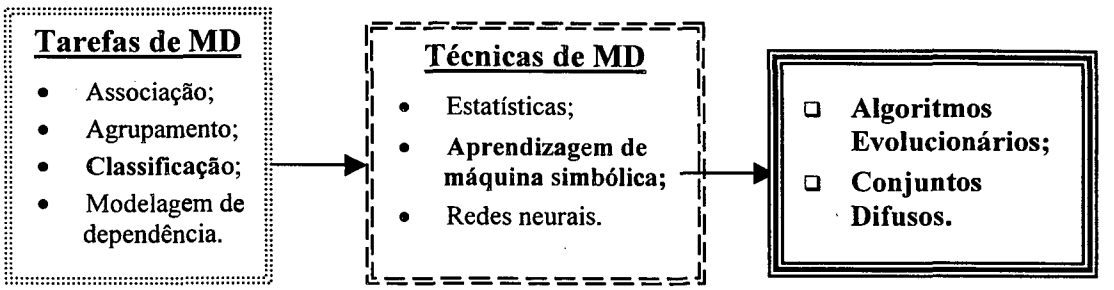


Figura 13: Algumas técnicas para aprendizagem de máquina simbólica.

A maioria dos métodos de MD é baseada no paradigma de indução de regras, onde se realiza uma espécie de busca local (*e.g.*, *Hill Climbing*), que não leva em consideração a interação entre atributos, considerando apenas um atributo por vez no processo de seleção dos mesmos. Neste campo é bem conhecido o problema da separabilidade de classes não-linearmente separáveis. Por exemplo, no problema tradicional do OU exclusivo (XOR), mostrado no Quadro 7, saber o valor de apenas um atributo não é suficiente para prever o valor da função.

Quadro 7: Função OU exclusivo

A1	A2	XOR
0	0	0
0	1	1
1	0	1
1	1	0

Os algoritmos evolucionários surgem como uma alternativa motivada pelo fato dos mesmos realizarem busca global e explicitamente considerarem a interação entre atributos. Em particular, a função de Fitness avalia um indivíduo solução candidato como um todo, em vez de um atributo por vez.

Existem vários tipos de algoritmos evolucionários (AE) disponíveis na literatura aplicados em diversas áreas. Na tarefa de descoberta de regras de classificação os mais utilizados são: os algoritmos genéticos e a programação genética (Freitas, 2002b).

Os algoritmos genéticos (AG's) podem ser usados para otimizar parâmetros de vários tipos de algoritmos de mineração de dados, tais como: encontrar um conjunto de pesos de atributos para algoritmos de aprendizado baseado em instância; encontrar uma topologia de interconexões para uma rede neural, etc. Além disso, AG's podem ser usados como um paradigma para descoberta de conhecimento por si só.

A programação genética (PG) possui como principal característica, que a difere dos AG's, a sua estrutura de dados. Os indivíduos podem ser considerados como "programas" de comprimento variável, uma vez que podem conter tanto dados como operações. Cada indivíduo geralmente é constituído de uma árvore onde cada nó interno da árvore corresponde a uma função (*e.g.*, AND, >, +) e cada folha da árvore corresponde a um atributo previsor (*e.g.*, Teses_Orient, Dissert_Orient) ou a um valor de atributo (*e.g.*, "baixo", "alto"), conforme mostra a Figura 14.

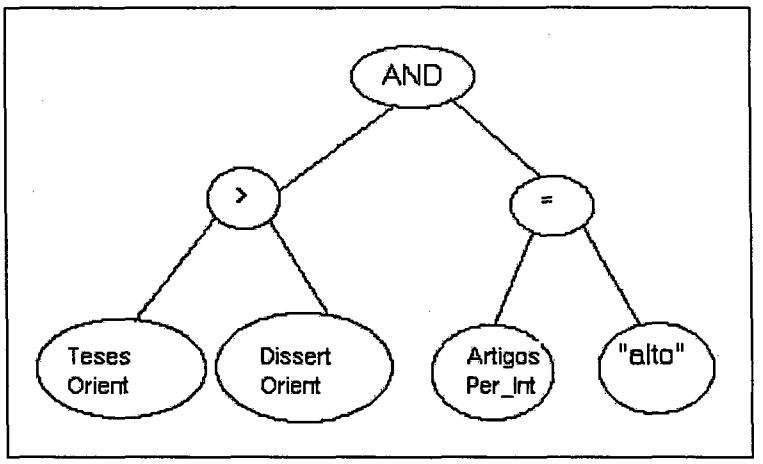


Figura 14: Exemplo de indivíduo em PG.

O indivíduo da Figura 14 representa o antecedente da seguinte regra:

(Teses_Orient>Dissert_Orient), (Art_Per_Int="alto") ⇒ Cap_livros_int="alto".

Apesar do grande potencial da PG, ela apresenta um problema quando aplicada a MD: requer que todos os nós da árvore retornem valores do mesmo tipo. Como na aplicação considerada nesta tese há diversos atributos de diferentes tipos, PG não foi empregada. Algumas possíveis soluções podem ser encontradas em (Bojarczuk *et al.*, 2000).

Para obtenção de conhecimento relevante ao planejamento de C&T, foram avaliadas e adaptadas duas técnicas da área de inteligência artificial que podem ser aplicadas à tarefa de classificação: algoritmos genéticos e conjuntos difusos, ambos descritos neste capítulo.

O interesse é descobrir conhecimento que tenha um certo poder de predição. A idéia básica é prever os valores que alguns atributos irão ter baseado nos dados observados previamente.

Segundo Fayyad *et al.* (1996b), os bancos de dados atuais estão povoados por atributos estruturados hierarquicamente, os atributos muitas vezes são fortemente relacionados e cada vez mais são requeridos meios mais sofisticados para a representação do conhecimento. Isto requer algoritmos que sejam efetivamente capazes de explorar estas informações. No entanto, os algoritmos de MD convencionais foram desenvolvidos para registros de atributos-valor simples, exigindo a criação e adaptação de novas técnicas que considerem a relação entre atributos.

Os algoritmos genéticos, os quais têm ganhado crédito no meio científico, exigem pouco conhecimento do problema tratado para encontrar uma boa solução em espaços de busca complexos multivariáveis onde há interação entre variáveis.

Neste capítulo, o objetivo é demonstrar a aplicabilidade de algumas abordagens genético-difusas (*i.e.*, técnicas fundamentadas na combinação de Algoritmos Genéticos (AG's) e Conjuntos Difusos), para extração de regras de previsão.

5.2 Algoritmos Genéticos

Os algoritmos genéticos (AG's) são algoritmos de busca e otimização baseados na analogia com os processos de seleção natural e genética evolucionária (Goldberg, 1989). A essência do método consiste em manter uma população de indivíduos (cromossomos), onde cada indivíduo representa uma possível solução para um problema específico. A melhor solução é atingida através de um processo de seleção competitiva, envolvendo cruzamentos e mutações (Herrera *et al.*, 1996).

A genética define cromossomos como corpúsculos que aparecem no núcleo de uma célula, considerados a sede dos genes, responsáveis em transmitir características hereditárias. Cada uma das partículas cromossômicas que encerram os caracteres hereditários é chamada gene.

A constituição hereditária de um indivíduo é dada pelo seu *genótipo*, enquanto que *fenótipo* é o nome dado àqueles que têm o mesmo aspecto geral de outros da mesma espécie, diferindo dela apenas por certos caracteres exteriores resultantes de condições mesológicas (meio ambiente).

O sentido exato do termo fenótipo em AG é um pouco controverso. Como exemplo, considere os dois indivíduos semelhantes abaixo:

Ind. A: (Idade='baixo'),(Nível_formação='médio') --> Artigos_per_nac='alto'

Ind. B: (Idade='baixo'),(Nível_formação ='baixo') --> Artigos_per_nac='alto'

Para avaliar a semelhança destas regras, há duas formas de se definir fenótipo:

1) Fenótipo é a regra "decodificada" a partir do genótipo. Suponha que o genótipo tenha 10 genes, cada um deles uma condição de regra em potencial, mas cada condição pode ser ativa ou não (que é o caso desta pesquisa). Se apenas dois genes estão ativos, a decodificação daqueles dois genes nas duas condições do antecedente da regra caracteriza o fenótipo da regra. Logo, as duas regras acima têm diferentes fenótipos, já que uma das condições é diferente com relação a um valor de atributo (mesmo atributo não é suficiente, é necessário que o par atributo-valor inteiro seja igual).
2) Fenótipo é determinado pelos exemplos cobertos pela regra. Assim, duas

regras diferentes podem ter o mesmo fenótipo se elas cobrirem exatamente os mesmos exemplos. Isso é improvável, mas é possível. Nesse caso o fenótipo seria definido pela interação entre a regra e o "meio ambiente" (os exemplos do banco de dados).

Em ambas definições acima, pode-se afirmar que as duas regras do exemplo têm fenótipos diferentes.

AG simula uma população genética organizada, composta por indivíduos que competem entre si para sobreviver e cooperam para alcançar uma adaptação melhor. Estes agentes também são chamados de cromossomos ou indivíduos. Similares aos cromossomos originais, os indivíduos de um AG são compostos por genes (genótipo). O significado de um indivíduo é definido externamente pelo usuário e fornece a solução aproximada de um problema específico.

Um AG é um procedimento iterativo para evoluir uma população de indivíduos. Em cada iteração, os indivíduos da população atual são avaliados por uma função de "Fitness" (aptidão), que mede a qualidade da solução associada ao indivíduo. Em seguida são aplicados operadores genéticos para criar novos indivíduos a partir dos indivíduos atuais.

Os AG's utilizam dois mecanismos adaptativos: seleção e herança. Os fatores determinantes do processo adaptativo são: as variações hereditárias e a seleção natural. Esta última age sobre as variações hereditárias eliminando os portadores de variações que dificultam ou impedem a sobrevivência e mantendo aqueles cujas variações permitem melhor explorar o ambiente. As variações hereditárias ocorrem devido principalmente a dois fatores: crossover e mutação.

A seleção, ou competição, é um processo estocástico de sobrevivência de um indivíduo proporcional à sua capacidade de adaptação. A adaptação é medida através da avaliação do *fenótipo* no ambiente em que está inserido, ou seja, sua capacidade de resolver o problema. Esta seleção promove a sobrevivência dos melhores indivíduos (melhores soluções) que irão gerar descendentes (novas soluções) baseados nas soluções "pais".

A herança é obtida através de cruzamento (*crossover*) do material genético

entre dois indivíduos selecionados, com a expectativa de produzir indivíduos melhor adaptados ou melhores soluções.

Além de cruzamento, pode-se empregar mutação, em menor proporção, para introduzir variação adicional. A mutação possibilita a geração de genes não presentes na população corrente, aumentando a robustez do algoritmo.

Um dos aspectos que torna um AG atrativo é sua habilidade para acumular informações sobre um espaço de busca inicial desconhecido e explorar este conhecimento para levá-lo a buscas subseqüentes em subespaços úteis (Perneel *et al.*, 1995).

Os AG's têm sido utilizados com sucesso na otimização de parâmetros em diversos tipo de aplicações, como pode ser visto em Punch *et al.* (1993), Turney (1995), Janikow (1995) e Romão (1999b).

5.2.1 AG's na Descoberta de Regras de Classificação

A utilização de AG na tarefa de descoberta de regras de classificação (previsão) é motivada pelo fato dos mesmos realizarem uma busca global que implicitamente considera a interação entre os atributos, enquanto que a maioria dos métodos tradicionais realiza busca local que não considera a importante interação entre os atributos. Por exemplo, o fato de um atributo isolado indicar que um pesquisador produziu mais artigos do que outro não garante que o primeiro é mais produtivo sem que se considere a interação deste com outros atributos, tais como "idade" e "titulação".

Quando AG é aplicado à tarefa de classificação, cada regra é construída e avaliada como um todo (conjunto de condições), geralmente lidando melhor com a interação entre os atributos do que a maioria dos algoritmos de árvore de decisão que constróem regras uma condição por vez.

Os AG's podem descobrir regras de previsão, do tipo SE... ENTÃO..., no formato semelhante às regras descobertas por indução de regras, garantindo a fácil inteligibilidade do conhecimento descoberto.

Quando AG é utilizado para descoberta de conhecimento por si só, os indivíduos representam regras de previsão ou outra forma de conhecimento. A função de Fitness mede a qualidade das regras ou do conhecimento associado

com os indivíduos. Por exemplo, o indivíduo abaixo pode representar uma regra para prever se determinado pesquisador poderá ou não assumir a liderança de determinado projeto:

(NIVEL_FORMAÇÃO = 'doutorado'), (ARTIGOS_PUBLICADOS = '20').

O indivíduo do exemplo acima representa apenas o antecedente de uma regra. O conseqüente pode ser escolhido de forma mais determinística. Uma opção é escolher a classe mais freqüente no conjunto de exemplos satisfazendo as condições da regra. Por exemplo, se o antecedente da regra (parte SE) é satisfeita por 100 exemplos, 80 da classe "+" e 20 da classe "-", escolhe-se classe "+" para o conseqüente da regra (ENTÃO classe = '+).

5.2.2 Michigan x Pittsburgh

Para utilização de AG's na tarefa de classificação, existem duas abordagens básicas que receberam os nomes das universidades de Pittsburgh e Michigan onde foram desenvolvidas.

A primeira abordagem, *Pittsburgh*, conhecida como LS (*Learning System*), foi proposta por Smith (1980). Nesta abordagem a codificação é complexa, implicando em operadores mais complexos, uma vez que cada indivíduo representa uma solução do problema, ou seja, cada indivíduo representa um conjunto de regras, onde cada regra é uma conjunção de condições e cada condição pode ter uma disjunção interna, conforme mostra a Figura 15.

Titulo = "doutor"	40 < Idade < 55	Sexo = "M"	outras condições	outras regras

Figura 15: Indivíduo representando um conjunto de regras.

Uma vantagem na abordagem de *Pittsburgh* é que a função de Fitness do indivíduo permitirá a avaliação do conjunto de regras como um todo, incluindo implicitamente a interação entre estas regras (Freitas, 2001b). Logo, a otimização desta função implicará automaticamente a busca de regras que, combinadas, fornecerão os melhores resultados com um mínimo de redundância.

A segunda abordagem, *Michigan*, também conhecida como CS (*Classifier*

Systems), foi desenvolvida originalmente por Holland e seus colegas (Holland, 1986). Aqui cada regra é representada por um indivíduo, sendo que um conjunto de regras (uma solução do problema) é representada por um conjunto de indivíduos, possivelmente a população inteira de indivíduos.

Artigos publicados > 10	Titulo = "mestre"	outras condições

Figura 16: Indivíduo representando uma regra (Michigan).

A abordagem de *Michigan* facilita a codificação dos indivíduos (Figura 16) permitindo a construção de indivíduos simples e pequenos, mas apresenta dificuldades em lidar com o problema de interações entre regras (a qualidade da solução representada por um indivíduo depende de outros indivíduos). Ocorre que a avaliação de cada indivíduo (Fitness) corresponde a cada regra, não considerando o conjunto de regras solução como um todo.

O problema da interação entre regras pode ser visualizado através da Figura 17, onde A1 e A2 seriam dois atributos previsores e R1 e R2 seriam regras (indivíduos).

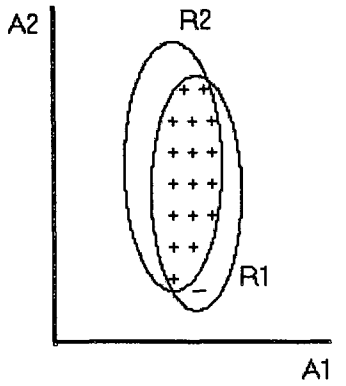


Figura 17: Interação entre duas regras.

Ambas as regras R1 e R2 têm uma boa qualidade cobrindo vários exemplos positivos e nenhum ou apenas um exemplo negativo. Porém, naturalmente se a abordagem levasse em conta a interação entre regras, a regra R1 seria descartada uma vez que a regra R2 já cobre todos os exemplos positivos de R1 sem misturar qualquer exemplo negativo.

Na abordagem de Michigan, para descobrir um conjunto de regras é necessário um mecanismo especial para manter a diversidade da população e assim evitar a convergência para uma única regra.

5.2.3 Representação do Conhecimento

Em um AG, os indivíduos são representados por códigos, compondo um alfabeto ou *string*, geralmente formado por dígitos binários (0 e 1). A codificação binária inicialmente dominou a pesquisa em AG (Houck *et al.*, 1996), pois é de fácil implementação e apresenta resultados promissores (Herrera *et al.*, 1996). Entretanto, é possível encontrar outros alfabetos, como os de ponto flutuante (codificação real) e a codificação em alto nível representando as informações no seu formato original, com diversos tipos de dados diferentes codificados no mesmo indivíduo.

Quando o assunto é MD, há necessidade de algumas considerações no projeto dos AG's. A representação do indivíduo, os operadores genéticos e as funções de aptidão devem ser adaptadas para extração de conhecimento de alto nível compreensível para o usuário.

Se o AG é aplicado à tarefa de classificação, os indivíduos podem representar apenas os antecedentes das regras de classificação, desde que os conseqüentes das regras sejam determinados por algum critério pré-definido. Sendo todos os atributos categóricos (os valores contínuos seriam discretizados), poder-se-ia utilizar codificação binária, onde o número de bits de determinado atributo é definido igual à quantidade de valores possíveis para este atributo. Nesse caso, em cada condição mais de um bit pode estar "ligado" (tendo valor 1), indicando uma disjunção implícita dos valores do respectivo atributo. Por exemplo, o valor "0 0 1 1 0" pode representar a condição: SE (titulação = "mestre" OU "doutor"), onde o atributo *titulação* assume os valores "graduado", "especialista", "mestre", "doutor" ou "pós-doutor".

Este esquema poderia ainda ser estendido para representar regras com várias condições, ligadas implicitamente pelo operador de conjunção, através da simples inclusão de mais bits no genoma.

Outra alternativa é representar o genoma codificando as condições das

regras diretamente em uma linguagem de alto nível. Exemplo:

(idade = "alta"), (titulação = "especialista"), (prod_científ = "baixa")

Seja qual for a forma de codificação, em geral é necessário utilizar uma estrutura de dados que permita variar o tamanho do vetor que representa o indivíduo, uma vez que não se sabe previamente quantas condições serão produzidas em cada regra.

O operador de cruzamento também deve ser adaptado para trabalhar com indivíduos de tamanho variável. Além disso, para evitar a criação de filhos "deformados", o indivíduo deve ser programado para manter, na sua representação interna, a mesma ordem dos atributos que se encontra no conjunto de dados de treinamento.

Os dados podem conter atributos categóricos (ordenados ou não) e atributos contínuos. Os atributos categóricos ordinais podem ser associados com os operadores "≥" e "≤", além do operador "=" nas condições das regras (*e.g.,* TÍTULO ≥ 'mestre'). Os atributos puramente categóricos só podem ser associados com o operador "=" (*e.g.,* NACIONALIDADE = 'estrangeiro').

Em MD, as regras podem ser representadas de duas formas: lógica proposicional (ordem 0) ou lógica de 1ª ordem.

Na lógica proposicional cada condição nas regras pode comparar apenas atributos com valores de seus domínios (*ex.:* Idade>20), enquanto que na lógica de 1ª ordem é possível representar a comparação entre atributos diferentes.

As condições proposicionais devem ter a forma $[A_i \; op \; V_{i,j}]$ onde $V_{i,j}$ é o j-ésimo valor do domínio do i-ésimo atributo previsor (*e.g.,* TESES_ORIENT ≤ 5).

O operador *op* pode assumir um dos valores =, ≥ ou ≤.

As condições de primeira ordem podem ter a forma $[A_i \; op \; A_j]$ onde A_i e A_j são dois atributos compatíveis (*e.g.,* DISSERT_ORIENT ≥ TESES_ORIENT).

Em AG pode-se utilizar tanto a lógica proposicional (ordem zero) como a lógica de 1ª ordem. Os AG's convencionais utilizam apenas lógica proposicional, o que os tornam eficientes em termos computacionais, mas limitados em expressividade. Considere um exemplo com dois parâmetros monetários: FOMENTO e GASTOS, conforme a Figura 18.

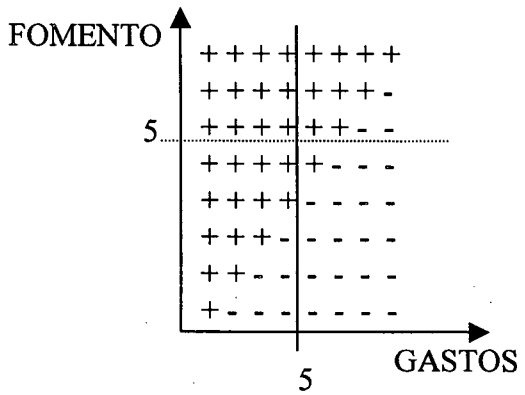


Figura 18: Exemplo com uma regra.

Pelo exemplo, uma condição do tipo [(FOMENTO > 5), (GASTOS < 5)] permite selecionar um subgrupo dos exemplos positivos mas não é possível cobrir todos os exemplos positivos com uma única regra. Seria necessário diversas regras para classificar todos os exemplos, conforme mostra a Figura 19, o que é contrário ao critério de facilidade de compreensão do conhecimento obtido.

É fácil observar que, pela lógica proposicional, não é possível criar uma regra genérica que possa expressar a relação entre os atributos FOMENTO e GASTOS. Não se pode descobrir uma regra com a condição, por exemplo, FOMENTO > GASTOS, uma vez que esta condição envolve uma relação (">") entre dois atributos não prevista pela lógica proposicional.

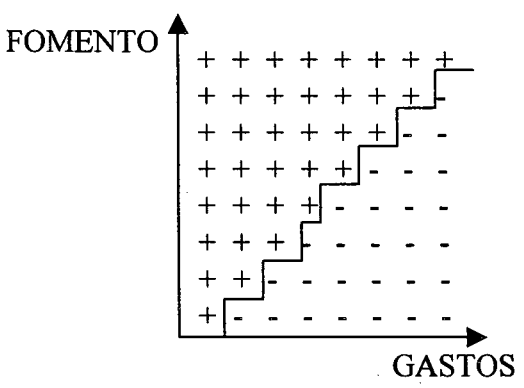


Figura 19: Exemplo com muitas regras.

Uma solução é empregar representação com lógica de primeira ordem, o qual tem o poder de expressar condições comparando dois atributos distintos.

No exemplo acima, uma única regra contendo a condição FOMENTO > GASTOS representaria o conhecimento necessário. Entretanto, esta solução tem como contrapartida o aumento no espaço de busca que, por conseqüência, provocará aumento no tempo de processamento do algoritmo de MD.

Além disso há outro problema: a incompatibilidade entre atributos de tipos diferentes. A comparação entre atributos só pode acontecer quando os mesmos forem do mesmo tipo, ou seja, possuírem o mesmo domínio. Por exemplo, considere os atributos contínuos SALÁRIO e IDADE. Eles não podem ser considerados pela lógica de primeira ordem com condições do tipo SALÁRIO > IDADE. Eles possuem tipos e domínios diferentes, e são incompatíveis.

Para resolver este problema, Freitas (1999a) propôs a criação de uma tabela baseada no dicionário de dados, contendo os tipos de todos os atributos utilizados pelo algoritmo de MD. Nesta abordagem, cada vez que o algoritmo cria uma população ou aplica um operador, terá que realizar uma consulta a esta tabela de tipos para verificar a possibilidade de utilizar lógica de primeira ordem na criação de regras/condições com os atributos envolvidos. Veja um exemplo na Tabela 2, onde as células marcadas com X indicam compatibilidade entre os atributos correspondentes à sua linha e coluna.

Tabela 2: Exemplo de atributos de tipos compatíveis

	Salário	Fomento	Idade	Art. Nac.	Art. Inter.	Teses orient.	Dissert. orient.
Salário	x	x					
Fomento	x	x					
Idade			x				
Art. Nac.				x	x		
Art. Inter.				x	x		
Teses orient.						x	x
Dissert. orient.						x	x

Outro problema é a existência de condições nulas dentro de uma regra. Quando um atributo não fizer parte de um indivíduo, ele deverá ser substituído por uma condição nula na sua representação. Como ilustração, suponha que o

conjunto de dados possua apenas os atributos produção, idade, salário, título e região. Considere os dois indivíduos (c1 e c2) abaixo:

c1: [(produção = "baixa") ("cond_nula") **(salário > 2000) ("cond_nula") (região = "norte")**]

c2: [(produção="alta") ("cond_nula") **("cond_nula") (título = "pós-doutor") ("cond_nula")**]

Aplicando, por exemplo, o operador de cruzamento (veja seção a)) após o segundo atributo, pode-se obter os filhos (c3 e c4) abaixo:

c3: [(produção="baixa") ("cond_nula") ***("cond_nula") (título="pós-doutor")("cond_nula")***]

c4: [(produção = "alta") ("cond_nula") ***(salário > 2000) ("cond_nula") (região = "norte")***]

Para codificar "cond_nula" dentro do indivíduo pode-se utilizar um bit adicional que indica a ocorrência ou não da condição em cada atributo, e este bit também fica sujeito aos operadores genéticos (Freitas, 2000b).

Nos AG's, cada indivíduo usualmente é representado apenas por uma lista (ou "string") podendo conter condições de regras, onde cada gene é constituído por um par atributo-valor, no caso das condições serem expressas em lógica proposicional (Freitas, 2002b).

Sumarizando, em AG cada indivíduo pode representar uma regra candidata (ou um conjunto candidato de regras) à solução de um dado problema (por exemplo, classificação). O indivíduo é composto por diversos genes, conforme a Figura 20, onde cada gene pode representar uma condição da regra.

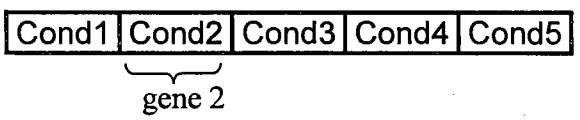


Figura 20: Exemplo de indivíduo com cinco genes.

5.2.4 Função de Aptidão (Fitness)

Todo AG exige a definição de uma função de aptidão, denominada função de Fitness. Ela fornece um valor que irá indicar a qualidade do indivíduo (solução candidata) avaliado. Esta função é específica para cada aplicação do AG e, portanto, deve ser definida para cada tipo de problema tratado. No contexto de mineração em banco de dados de grande porte, o maior tempo de processamento dos AG's é gasto na computação da função de Fitness (Freitas, 2002b).

Como ilustração, considere o seguinte problema simples: descobrir o valor de x que resulte em um máximo valor da função x^4 no intervalo [0, 63].

Considerando uma representação binária com 6 (seis) bits, para avaliação da função de Fitness de um indivíduo pode-se decodificar os seis bits produzindo um número x entre 0 a 63 e medir o resultado de x^4 que é a função de Fitness. Quanto maior o valor da função, melhor a qualidade do indivíduo.

Supondo que a população inicial seja composta pelos indivíduos da Tabela 3, gerados aleatoriamente, e que o método de seleção seja o da roleta (ver seção 5.2.5.1), então os indivíduos são copiados para a próxima geração de acordo com probabilidades proporcionais ao seu valor de Fitness. Neste exemplo, o terceiro indivíduo possui a maior probabilidade de se reproduzir (62%) por ter obtido o melhor resultado na função de Fitness.

Tabela 3: Exemplo de população com quatro indivíduos.

No.	Indivíduo	x	Fitness (x^4)	% do total
1	000110	6	16	1
2	001111	15	169	12
3	100000	32	900	62
4	010101	21	361	25

Existem diversos fatores que podem ser medidos na função de Fitness. Assumindo que cada indivíduo representa uma regra na tarefa de classificação, pode-se utilizar a taxa de acerto, generalidade e simplicidade sintática da regra.

A taxa de acerto é a proporção de exemplos corretamente classificados pela regra. A generalidade é o número de exemplos cobertos pela regra. A simplicidade sintática da regra é considerada, em geral, o inverso do número de condições por regra, ou seja, quanto menos condições na regra mais simples e compreensível ela se torna.

Pode-se fazer uma combinação ponderada desses fatores:

Fitness = a1*Acerto + a2*Generalidade + a3*Simplicidade (5.1)

Os parâmetros a1, a2 e a3 podem ser pesos pré-definidos pelo usuário, caracterizando uma forma subjetiva de avaliação.

Vale ressaltar que, se cada indivíduo representa uma regra (Michigan), a

função de Fitness é utilizada para avaliar o indivíduo (regra) como um todo, em vez de cada parte do indivíduo (condição de regra). Se cada indivíduo representa um conjunto de regras (Pittsburgh), a função de Fitness avalia o conjunto de regras solução como um todo, e não cada regra ou condição individualizada.

5.2.5 Métodos de Seleção

5.2.5.1 Roleta

A partir dos resultados da função de Fitness é possível comparar diversos indivíduos e escolher os melhores através de um método de seleção. O método é aplicado de forma que indivíduos mais aptos são usados para gerar os indivíduos da próxima geração. Ele é responsável por selecionar dois indivíduos, dentre aqueles que obtiveram os melhores valores da função de Fitness, para serem usados pelos operadores de cruzamento e mutação. O método de seleção mais conhecido é baseado no princípio do *jogo da roleta*.

Seguindo o exemplo do problema simples anterior (Tabela 3), a distribuição de probabilidade para os quatro indivíduos da população inicial seria conforme mostra a Figura 21.

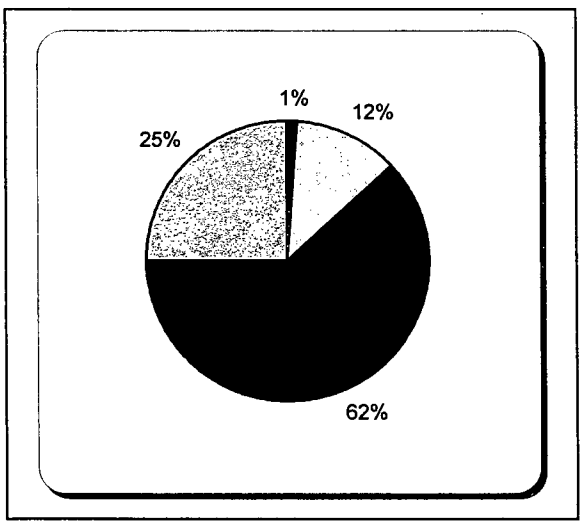


Figura 21: Método da roleta para seleção de indivíduos.

Para escolher um indivíduo para reproduzir, usa-se um gerador de números aleatórios que simula o giro de uma roleta. Assim, quanto maior a área da fatia (*slot*) da roleta associada com um indivíduo, maior a probabilidade daquele indivíduo ser escolhido. O procedimento da roleta deve ser repetido uma quantidade de vezes suficiente para formar uma nova população.

5.2.5.2 Torneio

Neste método de seleção k indivíduos são escolhidos aleatoriamente da população corrente. Os indivíduos escolhidos são comparados entre si, caracterizando um torneio, e aquele com maior valor da função de Fitness é escolhido para reprodução (Back, 2000). O indivíduo vencedor do torneio pode ser ou não removido da população corrente. Se não for removido ele poderá competir no próximo torneio.

O torneio deve ser repetido uma quantidade de vezes suficiente para formar uma nova população.

5.2.6 Operadores Genéticos

A escolha dos operadores genéticos, juntamente com a determinação da função de *Fitness* (aptidão) e da representação apropriada dos indivíduos, é determinante para o sucesso de um AG. Os operadores genéticos de seleção, cruzamento e mutação fornecem o mecanismo básico de busca e são usados para criar novas soluções baseadas nas melhores soluções existentes na população atual do AG.

a) Operador de Cruzamento

Uma vez que os indivíduos já foram selecionados, é possível aplicar o operador de cruzamento, o qual realiza a troca de material genético entre pares de indivíduos selecionados para reprodução. Para isto, aleatoriamente é escolhido um ponto entre dois genes onde se efetua um corte, conforme mostra a Figura 22.

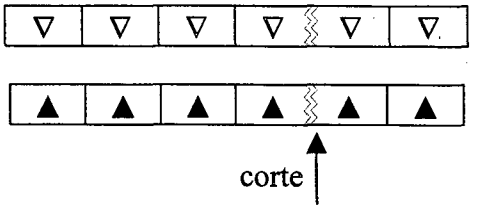


Figura 22: Indivíduos antes do cruzamento.

Supondo que o ponto sorteado para o corte seja entre o quarto e quinto genes, o primeiro indivíduo receberia os genes 5 e 6 do segundo indivíduo e vice-versa. Após o cruzamento, os indivíduos ficariam conforme mostra a Figura 23.

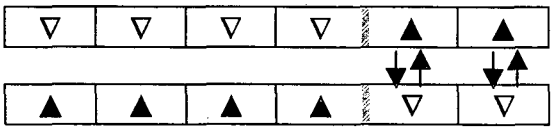


Figura 23: Indivíduos após o cruzamento.

O objetivo do operador de cruzamento é combinar "blocos de construção" de indivíduos diferentes, tentando combinar partes de cada indivíduo.

b) Operador de Mutação

O objetivo no uso do operador de mutação é aumentar a diversidade genética da população. Ele pode ser aplicado individualmente sobre cada indivíduo selecionado, realizando a alteração de algum gene escolhido aleatoriamente. Este operador permite que novas combinações de genes, que ainda não existam dentro dos indivíduos da população, sejam viabilizadas.

A mutação geralmente é aplicada com uma freqüência bem baixa, pois freqüências muito altas implicariam busca aleatória.

5.2.7 Características dos AG's

Tanto o operador de cruzamento como o de mutação são aplicados conforme probabilidade definida pelo usuário, sendo a probabilidade de cruzamento normalmente definida bem maior do que a de mutação.

A cada nova geração, indivíduos melhores são criados fazendo com que os indivíduos evoluem para soluções melhores. Em suma, os AG's apresentam as seguintes características principais:

- ❖ Em cada iteração, os AG's avaliam e modificam uma grande quantidade de soluções candidatas;
- ❖ Soluções de várias partes do espaço de busca são consideradas e combinadas em cada iteração, caracterizando uma busca global;
- ❖ A função de Fitness avalia a qualidade de um indivíduo como um todo, e não de suas partes;
- ❖ AG's são não determinísticos, ou seja, operadores são aplicados baseados em certas probabilidades, indivíduos da população inicial geralmente são criados aleatoriamente, etc.;
- ❖ Usam uma heurística extremamente genérica baseada no princípio da seleção natural, ou sobrevivência do mais apto;
- ❖ Um AG convencional usa operadores genéticos "cegos" sobre a semântica de um indivíduo, mas alguns AG's mais elaborados usam operadores específicos para o problema alvo.

5.2.8 Tendências (*biases*) nos AG's

Uma tendência (do termo em inglês "bias") é qualquer critério, explícito ou implícito, diferente de rigorosa consistência com os dados, usado para favorecer uma hipótese sobre outra (Mitchell, 1997). A eficácia de uma tendência depende do domínio de aplicação (Domingos, 1998).

Todo algoritmo de aprendizado indutivo deve ter pelo menos uma tendência. Os tipos de tendências mais comuns são de representação e de preferência. Nos estudos de Mitchell (*apud* Domingos, 1998, p.38) está implícito que "é improvável ocorrer aprendizagem livre de alguma tendência". Qualquer algoritmo de classificação deve ter alguma tendência indutiva para favorecer determinadas hipóteses.

Por exemplo, os AG's, na sua forma convencional (lógica proposicional), não podem descobrir regras comparando dois atributos diferentes (*ex.:* ANO_PRODUÇÃO > ANO_FORMAÇÃO). Neste caso eles possuem uma

tendência de representação. Além disso, a implementação da função de avaliação das regras, bem como o método de busca, representam tendências de preferência.

Os sistemas de aprendizagem de conceitos requerem a incorporação de várias tendências (biases) fortes para induzir regras de classificação de forma eficiente. No entanto, nem todas as tendências pré-definidas são apropriadas para todas as tarefas de aprendizagem (De Jong *et al.*, 1993).

5.3 Conjuntos Difusos

A lógica binária comum trabalha com dois valores: verdadeiro ou falso. No entanto, no mundo real nem sempre esta representação corresponde à realidade, pois em geral proposições são verdadeiras com um certo grau de veracidade. Conjuntos Difusos, juntamente com a Lógica Difusa e o raciocínio aproximado, são uma alternativa para a abordagem clássica booleana (Zadeh, 1965). Eles podem representar informação de acordo com seu grau de verdade e permitem codificar expressões do tipo "mais ou menos", "aproximadamente", "quase", "pouco", "muito", etc..

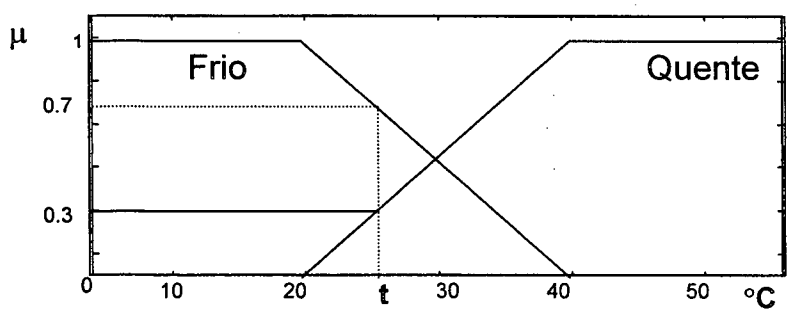


Figura 24: Conjuntos difusos de temperatura.

Na teoria de conjuntos difusos uma mesma informação pode ser representada por mais de um conjunto difuso. Por exemplo, um dado indicando que a temperatura foi medida como sendo de 25°C pode simultaneamente ser expressa por termos lingüísticos tipo "pouco fria" e "pouco quente". Esta dualidade da informação reflete-se nas funções de pertinência (FP's) de cada temperatura em cada conjunto (frio e quente). Os graus de pertinência (μ)

indicam o quanto (pouco, muito) uma temperatura corresponde a cada conjunto. As formas mais utilizadas de funções de pertinência são a trapezoidal e a triangular. A Figura 24 ilustra a interação entre dois conjuntos (frio e quente).

Neste exemplo diz-se que a temperatura está mais fria do que quente, ou seja, 0.3 quente e 0.7 fria. Conjuntos Difusos normalizados têm máximo valor de pertinência igual a 1, ou seja:

$$\mu_A(x) + \mu_{\neg A}(x) = 1$$

Quando os conjuntos difusos representam palavras da linguagem natural, a variável mapeada no conjunto é chamada *variável lingüística* (no exemplo acima, temperatura é uma variável lingüística).

Para criação de expressões lógicas, de forma análoga aos conjuntos clássicos, a Teoria dos Conjuntos Difusos forma a base científica para a *Lógica Difusa*. Nesta, uma implicação de fatos difusos é representada por regras difusas. Os sistemas de lógica difusa, em geral, são mapeamentos não lineares de um vetor de entrada (dados) em um escalar de saída, formando uma coleção de sistemas independentes de múltiplas entradas/única saída (Mendel, 1995).

Neste campo, são bem conhecidos os controladores difusos, com várias aplicações práticas nas engenharias.

A cada variável lingüística corresponde uma série de conjuntos difusos, de formas variadas, denominados "termos lingüísticos" que descrevem os diversos estados das variáveis lingüísticas.

A qualidade do sistema difuso depende principalmente da escolha dos seguintes fatores: operadores difusos; tipo das funções de pertinência (FP's); método de *defuzzificação* (*i.e.*, procedimento de extração do valor mais característico de um conjunto difuso); variáveis relevantes; número de funções de pertinência e quantidade de regras.

Uma vez construída a base de conhecimento de um sistema difuso, pode-se descrever a essência do método como sendo composta por quatro estágios:

- Determinação do grau de pertinência (μ) na regra antecedente (*fuzzificação*). Nesta etapa o valor informado é confrontado com cada

conjunto difuso do sistema (Por exemplo, Temperatura = 23° pode ser $\mu_{QUENTE} = 0.31$ e $\mu_{FRIO} = 0.69$);

- Cálculo dos conseqüentes das regras. Nesta etapa, empregam-se métodos do raciocínio aproximado (implicação) para derivar conclusões a partir da intercessão dos fatos apurados na primeira etapa com os antecedentes das regras difusas;
- Agregação dos conseqüentes das regras no conjunto difuso. Nesta etapa, os fatos apurados são conjugados para a construção do conjunto difuso resultante; e
- Defuzzificação. Esta etapa ocorre no estágio final quando se necessita retornar o valor típico do conjunto difuso derivado do sistema difuso (*e.g.*, problemas de controle).

O método mais utilizado para agregação é a implicação mínima de Mandani (Mendel, 1995), dada por:

$$\mu_{A \to B}(x,y) = \min[\mu_A(x), \mu_B(y)] \qquad (5.3)$$

onde a função de pertinência $\mu_{A \to B}(x,y)$ mede o grau de verdade da relação de implicação entre a entrada x e a saída y.

Fertig *et al.* (1999) vêem dois ingredientes que motivam a utilização da abordagem difusa em mineração de dados:

- a lógica difusa é um método flexível e poderoso para tratar com incertezas;
- regras difusas são um meio natural para representar regras com atributos contínuos.

5.4 Sistemas Híbridos Genético-Difusos

Um sistema é identificado como híbrido quando incorpora duas ou mais diferentes técnicas em sua estrutura, tais como: AG, redes neurais, lógica difusa, aproveitando a potencialidade de cada uma das técnicas de forma a realizarem um trabalho sinérgico.

Nesta tese, o objetivo é construir um sistema difuso de classificação, ou seja, encontrar um conjunto de regras com termos lingüísticos onde um objeto

desconhecido possa ser classificado utilizando raciocínio difuso.

A forma direta de se obter este sistema é dividir o espaço de exemplos (ou amostras) em uma grade difusa. Entretanto, quando há muitos atributos a combinação aumenta exponencialmente, tornando o problema complexo.

Os algoritmos genéticos aparecem como uma alternativa eficiente para tratar problemas com muitos atributos. Foi provado teórica e empiricamente que AG's são eficientes e robustos para encontrar soluções ótimas ou desejáveis em espaços complexos (Xiong & Litz, 1999).

Portanto, uma combinação interessante é a otimização de sistemas difusos através de algoritmos genéticos, conforme a Figura 25.

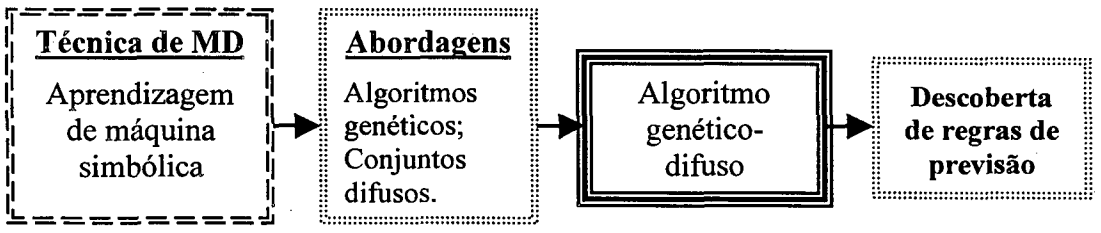


Figura 25: Abordagem genético-difusa.

Aplicações de AG na modelagem de sistemas difusos surgiram por volta de 1990, principalmente em controle de processos, seguidos por aplicações em química, medicina, telecomunicações, biologia e geofísica (Peña-Reyes & Sipper, 1999). Na atualidade, sistemas híbridos têm sido utilizados em muitas aplicações nas mais diferentes áreas.

Focalizando a tarefa de classificação, a combinação de AG com sistemas difusos apresenta qualidades marcantes, como a busca global e a facilidade de compreensão dos resultados, motivando diversos pesquisadores da área de MD a usufruírem destas técnicas.

5.4.1 Métodos Genético-Difusos

As seções anteriores forneceram subsídios para o entendimento da presente seção onde se discute vários sistemas híbridos genético-difusos para descoberta de regras de previsão. A discussão é apresentada com foco na aplicabilidade desses sistemas à descoberta de regras de previsão.

Na literatura existem diversas abordagens desta natureza. Os operadores genéticos de reprodução, cruzamento e mutação (descritos na seção 5.2.6) são apresentados como poderosos mecanismos para descoberta de regras relevantes em aplicações como otimização (Janikow, 1995), aproximação de funções (Delgado *et al.*, 1999), obtenção de resumos (Kacprzyk & Strykowski, 1999), diagnóstico médico (Peña-Reyes & Sipper, 1999), robótica (Deb *et al.*, 1998) e, mais especificamente, classificação (Lee, 1998; Mota *et al.*, 1999 e Xiong & Litz, 1999). Veja no Quadro 8 (seção 5.5) um resumo comparativo dos principais métodos genético-difusos analisados.

Na seqüência, são apresentados alguns comentários sobre algumas abordagens genético-difusas analisadas, para extração de regras, onde a maioria apresenta as FP's codificadas no indivíduo juntamente com as regras.

Na maioria das publicações analisadas as funções de pertinência foram codificadas no indivíduo juntamente com as regras. Uma exceção foi o trabalho de Kacprzyk & Strykowski (1999), os quais utilizaram cinco termos lingüísticos com funções de pertinência fixas para representar a imprecisão da linguagem natural. Eles aplicaram AG para realizar a busca na base de dados, na implementação de um sistema gerador de resumos de vendas de artigos de computador. A função de Fitness foi calculada baseada em cinco indicadores de qualidade dos resumos. Outra exceção foi o trabalho de Delgado *et al.* (1999), discutido abaixo.

Janikow (1995) apresentou um método, baseado em AG, que otimiza componentes de árvores de decisão difusas, o qual é incorporado na rotina de construção da própria árvore. Foram identificadas diversas restrições usadas para reduzir o espaço de busca. Ele utilizou FP's na forma de trapézios codificados no indivíduo através de quatro parâmetros correspondentes aos quatro vértices do trapézio, conforme a Figura 26.

A definição da representação de um indivíduo do AG incluiu restrições nos valores dos parâmetros $\beta_1 ... \beta_4$ mostrados na Figura 26. Além disso, restrições foram utilizadas para provocar a produção apenas de descendentes viáveis a partir de pais viáveis. O autor demonstrou que esta abordagem garante uma solução possível e melhora a eficiência da busca, podendo ser estendida para

otimização de sistemas baseados em regras.

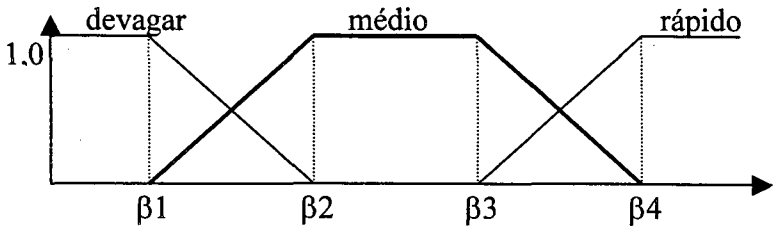


Figura 26: Conjuntos difusos utilizados em *Janikow* (1995).

Delgado *et al.* (1999) utilizaram AG para evoluir uma população de modelos difusos baseados em regras, através de uma abordagem intermediária entre Pittsburgh e Michigan, empregada na aproximação de funções.

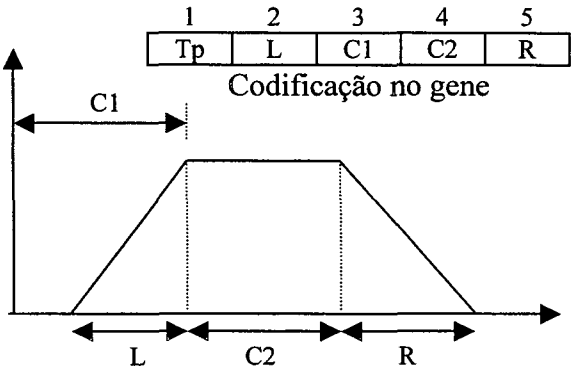


Figura 27: Representação das FP's em *Delgado et al.* (1999).

Foram utilizados sete termos lingüísticos para representar cada variável numérica e a forma de cada FP foi codificada no gene através de 5 (cinco) parâmetros: Tp, L, C1, C2 e R, onde Tp é um parâmetro adaptativo que indica o tipo da FP, podendo ser trapezoidal, triangular ou gaussiana. Os demais parâmetros definem a distribuição da FP, conforme a Figura 27.

Peña-Reyes & Sipper (1999) mostraram resultados da aplicação de técnicas genético-difusas no diagnóstico do câncer de mama a partir de características de tumores.

Os autores seguiram a abordagem de Pittsburgh, utilizando AG para realizar a busca de três parâmetros (conforme a Figura 28):

P indica o ponto de início da segunda FP;

d define a distribuição das FP's;

A_j^i representa as regras.

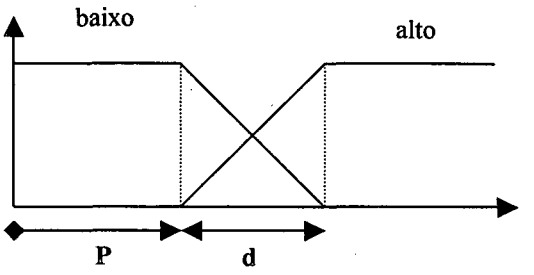


Figura 28: Representação das FP's em Peña-Reyes (1999).

Cada indivíduo foi codificado segundo a ilustração da Figura 29.

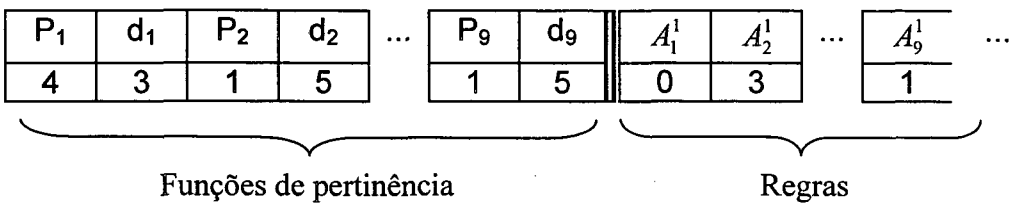


Figura 29: Indivíduo contendo várias regras *fuzzy* (Peña_Reyes, 1999).

Para a avaliação dos resultados foram combinados três critérios:

a) Fc = percentagem de casos classificados corretamente;

b) Fe = (valor estimado – valor correto)^2;

c) Fv = número médio de variáveis / regras ativas.

A função de Fitness foi definida como:

$$F = Fact - \alpha Fv - \beta Fe \qquad (5.4)$$

onde: $\alpha = 0.05$ e $\beta = 0.01$ (estimados empiricamente).

Deb *et al.* (1998) utilizaram a abordagem Genético-Difusa para otimizar off-line o tempo de viagem de um robô programado para atuar em um meio dinâmico onde outros objetos também estariam em movimento. Como nos demais trabalhos citados, AG foi aplicado para encontrar simultaneamente um conjunto ótimo de regras difusas e a distribuição das funções de pertinência.

Os autores fizeram uma analogia entre um robô e uma criança que aprende algumas regras, através de tentativas (treinamento), para depois conseguir se

locomover o mais rápido possível entre dois pontos específicos (teste), evitando se chocar em outras crianças ou objetos em movimento. Para isto é necessário saber a distância do objeto e o ângulo entre as direções de movimento de ambos.

Foram empregados quatro termos lingüísticos para distância e cinco para ângulo utilizando-se funções de pertinência triangulares. Para cada variável (distância e ângulo) as funções são mantidas com distribuição uniforme. A otimização é feita apenas sobre a largura da base do triângulo, bastando apenas dois parâmetros (b1 e b2) para codificar a forma das funções de pertinência, conforme Figura 30.

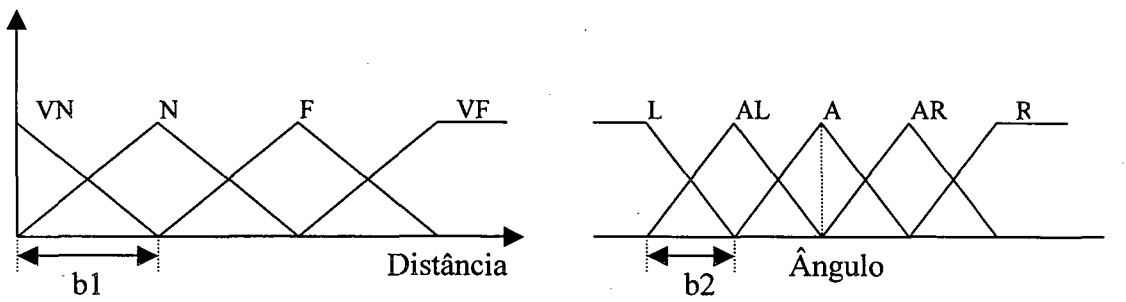


Figura 30: Conjuntos difusos utilizados por Deb *et al.* (1998).

Uma vantagem deste método é o fato da otimização ocorrer off-line obtendo-se primeiro uma base de regras para depois o robô se movimentar evitando, assim, o problema do alto tempo computacional para obter as regras quando comparado com o tempo de decisão para definição da rota do robô.

Lee (1998) propôs um método de classificação que difere dos demais por otimizar não só as regras e a forma das FP's, mas também a quantidade de FP's que nos outros métodos é fixa. Além disso, o conseqüente das regras é obtido através do uso do método do gradiente descendente, que minimiza o número de regras resultante.

As FP's, utilizadas na forma triangular, foram representadas por dois parâmetros, a saber: **a** (centro do triângulo) e **b** (largura da base), conforme a Figura 31.

A saída do sistema (y) é obtida por:

$$y = \frac{\sum_{i=1}^{n} (\mu_i . \omega_i)}{\sum_{i=1}^{n} \mu_i} \tag{5.5}$$

onde: μ_i = valor da FP;

ω_i = valor contínuo do conseqüente.

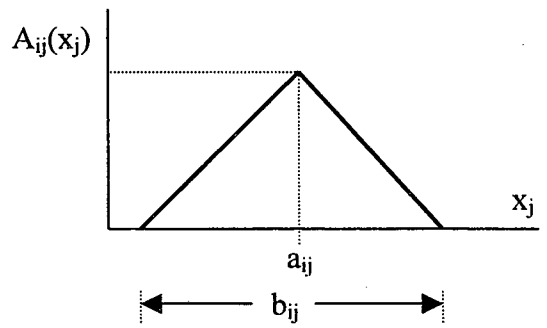


Figura 31: Representação da FP em Lee (1998).

A função de Fitness (E) é então obtida por:

$$E = \frac{1}{2}\left(y - y^p\right)^2 \tag{5.6}$$

onde: y^p = saída desejada.

Para testar o sistema foram selecionados cinco problemas (funções de duas variáveis) de classificação [f(x)] divididos em duas classes cada um, denominadas G1 e G2, onde f(x) ≥ 0 implica que x pertence à classe G1, caso contrário G2.

Mota *et al.* (1999) desenvolveram um sistema difuso de classificação para automatizar um estágio moroso, conhecido no contexto clínico como *Sleep*, de análise de 2000 páginas de registros poligráficos. O sistema é composto por quatro classificadores independentes e um módulo para integrar suas saídas.

A entrada do sistema, constituída por 12 variáveis, e a saída são representados por três FP's, sendo duas trapezoidais e uma triangular, conforme mostra a Figura 32.

Foi utilizado o método da soma e do produto nas fases de composição e inferência. A defuzzificação foi realizada pela média dos máximos e o número

máximo de regras foi limitado em 10.

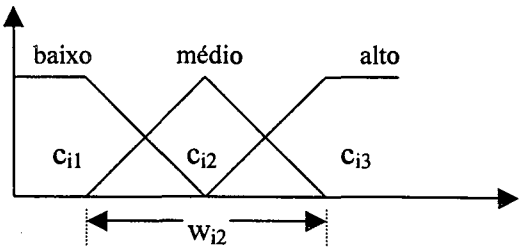


Figura 32: Conjuntos difusos em Mota *et al.* (1999).

Cada indivíduo foi formado por dois genomas: um para os parâmetros da FP e outro para as regras. O genoma das FP's foi codificado pelos parâmetros c_{ij} e w_{ij} (veja Figura 32), onde: c_{ij} = centro da FP j da variável i; e w_{ij} = largura da FP j da variável i.

Neste método, a função de Fitness (*f*) foi calculada por:

$$f = \frac{1}{\overline{D} + k} \tag{5.7}$$

onde:

$$\overline{D} = \frac{\sum_{i=1}^{P} D(d(x_i) - o(x_i))}{P}$$

$$d(x_i) = \begin{cases} 1, C(x_i) = s \\ 0, C(x_i) \neq s \end{cases}$$

$o(x_i)$ = saída do classificador o qual reflete a taxa de acerto;

$C(x_i)$ = classificação correta do padrão x_i;

$\overline{D}$ = distância média até a meta;

k = constante limite para o valor da função de Fitness;

P = número total de exemplos no conjunto de treinamento;

s = estágio.

A escolha dos indivíduos foi executada pelo método da roleta (*roulette wheel*), mantendo-se o elitismo. O cruzamento foi uniforme e a mutação foi realizada de acordo com uma distribuição normal.

Identificou-se a necessidade de eliminar as características de entrada irrelevantes para aumentar o desempenho do AG.

Morales (1997) apresentou um modelo para identificação de sistemas, baseado em regras, aplicando a teoria dos conjuntos difusos para incorporar incertezas utilizando FP's triangulares. O processo de busca é realizado por um AG onde, na representação de cada indivíduo, todos os valores lingüísticos são codificados nos lócus dos cromossomos como valores inteiros. Por exemplo, se o sistema que se deseja identificar tem duas variáveis, cada uma com cinco valores lingüísticos, sua representação é conforme a Figura 33.

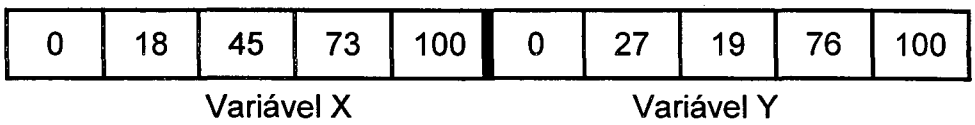


Figura 33: Representação do Indivíduo em Morales (1997).

Xiong & Litz (1999) apresentaram um método baseado em AG para classificação a partir de dados numéricos. Embora tenham aplicado apenas nos dados sobre a Íris, o método serve para classificação em geral. Eles seguiram a abordagem de Pittsburgh, codificando regras e FP's no mesmo indivíduo, conforme a Figura 34 (m = n° de atributos; k = n° de FP's).

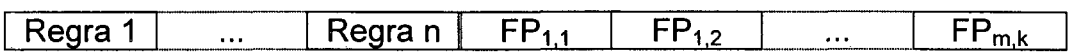


Figura 34: Representação das regras e FP's em Xiong & Litz (1999).

As regras foram codificadas por representação binária. Por exemplo, a regra: SE [x_1 = (pequeno OU grande)] E [x_3 = médio] E [x_4 = (médio OU grande)] foi codificada por: (101; 111; 010; 011). O padrão "111" na segunda condição indica que ela é efetivamente removida da regra (já que o respectivo atributo pode assumir qualquer um de seus valores).

Foram utilizadas FP's triangulares com distribuição determinada pelo parâmetro **p**, conforme mostra a Figura 35.

O parâmetro **p** foi codificado utilizando um número inteiro (e não binário como normalmente é codificado), o que permite reduzir o tamanho do indivíduo.

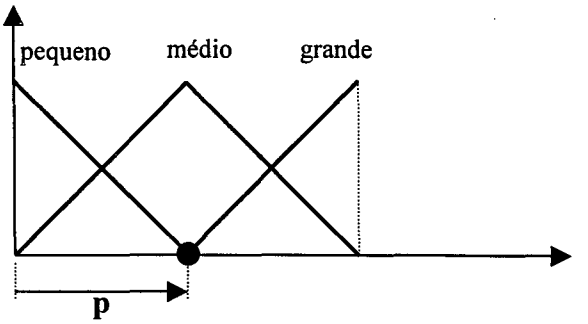


Figura 35: Conjuntos difusos definidos em Xiong & Litz (1999).

A operação de cruzamento foi realizada considerando a distinta natureza de cada uma das duas partes do indivíduo (regras e FP's). Para isto, foram aplicados cortes em três pontos: um corte fixo, para separar as duas *substrings* (uma contendo as regras, e outra contendo as FP's), e dois cortes correspondentes a um corte em cada *substring*, caracterizando os pontos reais de corte para cruzamento.

Devido ao mesmo motivo (a diferença entre as duas partes do indivíduo), a operação de mutação também foi realizada de forma diferente em cada parte. Como os valores do parâmetro **p** das FP's são contínuos, foi utilizado uma pequena mutação com alta probabilidade e uma perturbação em cada bit gerada por uma função Gaussiana. Na parte das regras, a mutação foi inserida através da simples inversão de bits com pequena probabilidade.

A função de Fitness foi calculada com base na taxa de acerto na classificação dos padrões de treinamento, seguindo os seguintes passos:

a) Decodificar a substring binária para a estrutura inicial de regras e a substring de inteiros nas funções de pertinência de entrada;
b) Determinar as classes dos conseqüentes a partir das condições das regras através da maximização dos valores verdade das regras;
c) Classificar os exemplos de treinamento com as regras geradas nos passos anteriores;
d) Computar a taxa de acerto, na classificação dos exemplos de treinamento, como o valor da função de Fitness.

Um resumo das abordagens genético-difusas analisadas, com as principais características de cada uma, encontra-se no Quadro 8 da próxima seção.

5.5 Discussão

Nesta seção é apresentada uma visão crítica dos principais pontos positivos e negativos dos métodos genético-difusos vistos na seção anterior. Essa visão crítica é apresentada no contexto de *mineração de dados*, culminando em sugestões sobre como adaptar os métodos acima para que eles se tornem mais adequados à descoberta de regras de previsão em C&T.

Quase todos os métodos analisados utilizam a abordagem de Pittsburgh, mas os dados utilizados para efetuar a extração de regras eram de pequeno volume. Para adaptar alguma abordagem genético-difusa sobre banco de dados de maior porte, como os dados sobre C&T, é recomendável utilizar a abordagem de Michigan para reduzir o espaço de busca e, consequentemente, o tempo de processamento.

A maioria das abordagens genético-difusas foram aplicadas a dados numéricos (veja coluna 3 do Quadro 8). Logo é aconselhável que se aplique abordagem deste tipo apenas a domínios que contenham número significativo de variáveis contínuas.

Restrições dentro do AG, semelhantes às utilizadas por Janikow (1995), também reduzem o espaço de busca e poderão ser empregadas em casos de extração de regras de previsão, onde se tem conhecimento do domínio, utilizando a experiência de um especialista para determinar restrições para cada atributo. Por exemplo, em alguns dados de C&T, o atributo NÍVEL_DE_FORMAÇÃO de um pesquisador possui um domínio contendo 15 (quinze) valores possíveis, mas apenas cinco ou seis destes valores são relevantes para a busca.

Apesar de quase todos os métodos analisados apresentarem as FP's codificadas dentro do indivíduo junto com as regras (veja 7ª coluna do Quadro 8), a otimização destas FP's é pouco significativa perante a importância do processo de evolução das regras. Em casos de banco de dados de maior porte é recomendável otimizar estas FP's em separado, utilizando algum método rápido de busca local, ou utilizar conhecimento do usuário para especificar a distribuição destas funções.

Quadro 8: Características de algumas abordagens genético-difusas.

AUTORES	*Objetivo*	*tipo dos dados*	*Forma da FP*	*N° de FP/var.*	*Parâmetro por FP*	*Conteúdo do Indivíduo*
Deb, K. *et al.*, 1998	Aprendizagem de regras p/ guiar robôs	Num	Triang	Dist=4; Âng=5	2	Regras + FP FP = (bases dos triângulos)
Delgado, M. *et al.*, 1999	Aproximação de funções	Num	Trap, Triang, Gauss	7	5	Regra + FP FP = (Tp,L, C1,C2,R)
Janikow, C. Z., 1995	Otimização de árvore de decisão difusa	Num	Trap	3	4	Regras + FP FP = 4 vértices nos trapézios
Lee, M-R., 1998	Classificação	Num	Triang	variável	2	Regras + FP = (3 parâmetros)
Mota, C. *et al.*, 1999	Classificação	Num	Triang ou Trap	3	4	Regras + FP FP = (centro; largura)
Morales, A. B. T., 1997	Identificação de Sistemas	Num	Triang	5	1	Picos das PF's
Peña-Reyes, C.A. & Sipper, M., 1999	Diagnóstico médico	Cat_ Ord	Trap	2	2	Regras + FP FP = (dois parâmetros)
Xiong, N. & Litz, L., 1999	Classificação	Num	Triang	3	1	Regras + FP FP = único ponto.

Legenda: Cat_Ord=Categórico ordenado; Num=Numérico; Dist=Distância; Âng=Ângulo; Mich=Michigan; Pitt=Pittsburgh;Trap=Trapezoidal; Triang=Triangular; Gauss=Gaussiana.

A utilização de métodos automáticos para otimizar as FP's possui um aspecto negativo: geração de FP's "contra-intuitivas", ou seja, gerar domínios que não fazem sentido. Por exemplo, um sistema automático poderia gerar uma FP para o atributo Idade com "baixa" até 40 anos.

A retirada das FP's de dentro do indivíduo simplifica a construção dos operadores genéticos e melhora o desempenho do AG, uma vez que o espaço de busca se torna menor. Além disso, no caso de extração de regras de previsão, o formato das FP's pode ser fixo, sendo os formatos trapezoidal e triangular os mais comuns (veja 4ª coluna do Quadro 8). No caso trapezoidal, a distribuição dos trapézios pode ser semelhante à utilizada por Janikow, onde quatro parâmetros são suficientes para codificar a distribuição de três FP's (veja coluna 6).

A quantidade de FP's também não precisa necessariamente ser otimizada, como proposta por Lee (1998). Em vários casos três FP's são suficientes (veja

5ª coluna do Quadro 8) e viabiliza a especificação manual das FP's diretamente pelo usuário.

A representação do indivíduo deve ser de mais alto nível possível. Isto aumenta a compreensão das regras e facilita a especificação de restrições para redução do espaço de busca. Uma alternativa seria a utilização de variáveis lingüísticas (*ex.:* Idade) com termos simples, tais como, "baixa", "média" e "alta".

5.6 Considerações Finais

Com o objetivo de obter conhecimento relevante a partir de banco de dados sobre C&T, neste capítulo procurou-se investigar as técnicas de MD utilizadas na tarefa de classificação, em especial as técnicas híbridas genético-difusas.

Os AG's possuem comprovada eficiência no trato com interação entre atributos, o que parece ser uma característica dos dados em C&T, apesar de serem cerca de duas a três ordens de magnitude mais lentos do que os algoritmos de indução de regras convencionais (Dhar *et al.*, 2000). No entanto, utilizando as restrições e simplificações discutidas acima, AG's tornam-se uma alternativa promissora.

Os conjuntos difusos incorporam os termos lingüísticos intrínsecos da linguagem natural e o raciocínio aproximado que o torna um método flexível e poderoso para tratar com incertezas.

Apesar da existência de várias técnicas convencionais de MD para realizar a tarefa de classificação, as discussões apresentadas indicam que a aplicação de AG's, em combinação híbrida com conjuntos difusos, é uma alternativa viável que pode ser estendida à tarefa de modelagem de dependência.

A abordagem genético-difusa, para realizar a tarefa de extração de regras de previsão, deverá fornecer conhecimento novo, compreensível e exato, representado por regras de tamanho e quantidade variáveis, a partir de banco de dados de C&T.

6 O ALGORITMO GENÉTICO PROPOSTO

6.1 Introdução

No Capítulo 5 efetuou-se levantamento bibliográfico sobre o uso de AG e Lógica Difusa no contexto da tarefa de classificação e estabeleceram-se estudos sobre a forma de representar regras através dos indivíduos (cromossomos) de um algoritmo genético.

Neste capítulo são descritos os principais detalhes da composição do modelo proposto e sua implementação como protótipo, na linguagem de programação *Delphi*, de um sistema híbrido denominado AGD (Algoritmo Genético para Descoberta de Regras Difusas).

No algoritmo AGD utiliza-se a abordagem de Michigan (descrita na seção 5.2.2) na codificação dos indivíduos no AG, onde cada indivíduo representa uma regra, assunto da seção 6.3. Os indivíduos poderão conter genes representando atributos categóricos ou contínuos. Os valores contínuos são "*fuzzificados*", isto é, representados por valores lingüísticos (conjuntos difusos) adequados à linguagem natural utilizada pelo usuário do sistema. Neste caso cada gene poderá assumir um valor lingüístico, e não o valor contínuo. Por outro lado, os atributos categóricos não são "fuzzificados" e, dependendo do atributo, cada gene correspondente assume um valor categórico do seu domínio original ou de um domínio derivado ou de um domínio construído a partir da agregação de atributos.

Nas seções 6.4 a 6.6 são apresentados, respectivamente, os métodos utilizados para seleção de indivíduos, os operadores genéticos empregados e elitismo.

Na seção 6.7 apresenta-se uma forma de integrar conjuntos difusos para representar as informações difusas da linguagem natural existentes em algumas variáveis contínuas. Exemplos de tais variáveis, no caso de C&T, são: idade, quantidade de artigos publicados (baixo, médio ou alto), etc..

Em MD, é desejável que o conhecimento extraído satisfaça pelo menos três propriedades: deve ser correto, compreensível e relevante. A primeira propriedade é atendida pela aplicação do conceito de matriz de confusão

adaptado para o objetivo do algoritmo proposto (seção 6.8). A segunda será facilitada pelo uso de regras do tipo SE... ENTÃO... e incrementada pelo uso de termos lingüísticos difusos. A terceira propriedade (relevante) é mais desafiadora e uma das principais contribuições desta tese (seção 6.9).

Na seção 6.10 discute-se a função de Fitness empregada e na seção 6.11 descreve-se um resumo do algoritmo proposto para descoberta de regras de previsão relevantes. Finalmente, na seção 6.12, são apresentadas algumas considerações finais deste capítulo.

A motivação para se propor um novo algoritmo genético para descoberta de regras de previsão difusas é discutida na próxima seção (6.2), onde são apresentadas diversas considerações, definições e particularidades que justificam a proposta.

6.2 Justificativa

As agências de fomento à C&T têm formado vários bancos de dados sobre a pesquisa brasileira, como exposto no Capítulo 2. As informações para tomada de decisão em gestão de C&T podem ser geradas por um módulo de MD desenvolvido com o objetivo de extrair conhecimento relevante para planejamento em C&T a partir de bancos de dados do CNPq.

O que pode ser feito? Para atender ao caráter previsor do planejamento, pode-se realizar experimentos utilizando técnicas de MD para resolver a tarefa de classificação ou modelagem de dependências, cuja característica principal é a descoberta de conhecimento que permita a previsão de futuros valores de uma ou mais variáveis chave, o que é essencial ao planejamento.

A intenção não é extrair um grande número de regras, como obtido com o algoritmo Apriori na extração de regras de associação (veja seção 3.4), e nem tão pouco descobrir um conjunto "completo" de regras de classificação, cobrindo todo o espaço de dados, mas sim extrair um pequeno número de regras surpreendentes, ou seja, regras de previsão novas e potencialmente relevantes para o usuário.

Descoberta de conhecimento relevante é um problema em muitos algoritmos de MD e uma área pouco explorada. Nesta tese apresenta-se uma

contribuição para resolução deste problema, tanto no contexto da aplicação considerada como no contexto da pesquisa básica da comunidade científica que pesquisa novas técnicas para descoberta de conhecimento relevante.

Apresenta-se um algoritmo genético como elemento principal na composição do sistema, o que é justificado considerando-se um dos pontos cruciais ao sucesso da aplicação de técnicas de extração de conhecimento: levar em consideração a interação entre atributos no processo de busca. Esta é a principal motivação para se utilizar algoritmos genéticos (Romão *et al.*, 1999b) na extração de regras de previsão. De fato, no AG a função de Fitness avalia o indivíduo como um todo, ou seja, todas as interações entre os atributos que ocorrem nas regras são consideradas.

O mecanismo de busca de conhecimento, empregado no algoritmo, é regido pelo processo evolutivo do AG que é baseado na analogia com o processo reprodutivo da natureza. Nesta abordagem, os cromossomos (indivíduos) representam regras e cada gene pode representar uma condição, que é uma informação direta do problema tratado. Foram considerados e adaptados os operadores genéticos mais adequados a esta representação.

Existem inúmeros algoritmos de indução de regras de previsão, mas a maioria trabalha apenas com a lógica clássica. Para tratar com os parâmetros contínuos existentes nos bancos de dados propõe-se a inclusão da lógica difusa ao algoritmo de mineração de dados. A lógica difusa será empregada de forma relativamente simples, principalmente na validação das novas regras diante do banco de dados. A essência do modelo proposto está no AG. Os conjuntos difusos são utilizados também como linguagem natural para obtenção de informações do usuário.

Conforme mencionado anteriormente, esta decisão está baseada em duas motivações - Fertig *et al.* (1999):

- a lógica difusa é um método flexível e poderoso para tratar com incertezas;
- regras difusas são um meio natural para representar regras envolvendo atributos contínuos.

A utilização da lógica difusa no contexto de mineração de dados se justifica

também pelo fato de intuitivamente as regras geradas serem mais compreensíveis para o usuário. Por exemplo, uma condição do tipo: [18 <= idade <= 29 ⇒ ...] pode ser substituída por uma condição mais simples, compreensível e natural utilizando uma variável difusa, do tipo: [idade = 'baixa' ⇒ ...], onde 'baixa' é uma informação difusa representando o grau ao qual uma pessoa é jovem, dado por uma função de pertinência.

A ênfase é encontrar conhecimento novo, compreensível e relevante para ser utilizado por humanos (*e.g.,* membro de comitê assessor) para tomada de decisão em alto nível.

Na publicação de Deb *et al.* (1998), foi apresentada a comparação de resultados da aplicação de abordagens genético-difusas e verificou-se que, para uma aplicação específica à área de robótica, encontrar um bom conjunto de regras é mais importante do que encontrar um bom conjunto de funções de pertinência, podendo considerar o primeiro como uma otimização grosseira e o segundo como uma otimização fina. Partindo desta hipótese, decidiu-se priorizar o processo de busca para encontrar um bom conjunto de regras, deixando as funções de pertinência para serem obtidas de uma outra forma mais simples: a partir do próprio usuário. Essa abordagem tem também a vantagem de que as funções de pertinência são facilmente compreensíveis para o usuário, já que ele as definiu; evitando o risco do sistema gerar automaticamente funções de pertinência contra-intuitivas para o usuário.

Veja na seção 6.3, a seguir, a forma utilizada para se representar os indivíduos no AG.

6.3 Representação do Indivíduo

A complexidade do conhecimento descoberto pode ser simplificada pela representação do mesmo. De Jong *et al.* (1993, p. 163) expressam bem esta idéia:

> *"Uma forma natural de expressar conceitos complexos é como um conjunto de disjunções de possíveis regras de classificação sobrepostas, isto é, na forma disjuntiva normal. O lado esquerdo de cada regra (i.é., disjunção ou condições) consiste de uma conjunção de um ou mais*

testes envolvendo valores de características (atributos). O lado direito da regra indica o conceito (valor de um atributo meta) a ser associado com os exemplos que se casam (são cobertos por) com o lado esquerdo da regra".

Todos os indivíduos possuem o mesmo tamanho fixo equivalente ao número de atributos considerados na tarefa de MD. Para simplificar, mesmo que um atributo não faça parte de determinada regra ele é representado no indivíduo, ou seja, o genótipo dos indivíduos é fixo mas o fenótipo pode variar. Como estamos interessados na tarefa de modelagem de dependência, são codificados tanto os antecedentes como o conseqüente das regras.

Cada atributo do antecedente é representado por um par de genes <Valor, Flag>. O Flag é utilizado para estabelecer se o atributo faz parte ou não da regra. Quando o atributo faz parte do conseqüente, o Flag correspondente indica que este atributo está inativo no antecedente.

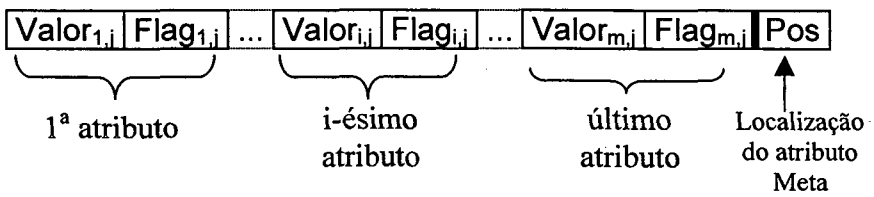


Figura 36: Codificação do indivíduo.

Cada indivíduo é representado conforme mostra a Figura 36. O conseqüente é codificado por um gene especial no final do indivíduo que recebe o valor da posição (correspondente aos atributos do antecedente) do atributo no próprio indivíduo, ou seja, o gene especial tem apenas um valor. O mesmo par <Valor, Flag> utilizado para representar os atributos previsores é utilizado para representar o atributo meta, mas neste caso o Flag é 0 para indicar que este atributo está desativado no antecedente.

Como exemplo, considere a seguinte regra de previsão:

(Idade = 'alta'), (UF = 'SC') ⇒ Art_per_nac = 'alto.

Considere ainda que os atributos que aparecem nesta regra são os atributos das posições 1, 5 e 7 dentro do indivíduo. Neste caso, a

codificação desta regra no indivíduo ocorre conforme mostra a Figura 37.

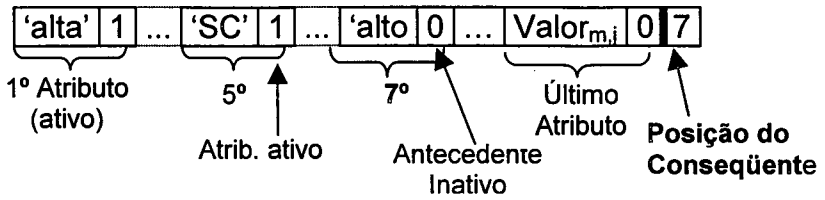


Figura 37: Exemplo de codificação de uma regra.

Cada indivíduo representa uma regra, conforme Figura 38.

$Valor_1^1$	$Flag_1^1$	...	$Valor_i^1$	$Flag_i^1$	...	$Valor_m^1$	$Flag_m^1$	Pos
$Valor_1^j$	$Flag_1^j$	...	$Valor_i^j$	$Flag_i^j$	...	$Valor_m^j$	$Flag_m^j$	Pos
$Valor_1^R$	$Flag_1^R$	...	$Valor_i^R$	$Flag_i^R$	...	$Valor_m^R$	$Flag_m^R$	Pos

Figura 38: Representação de uma população.

Da Figura 36 a Figura 38, a seguinte notação é utilizada:

i = é o i-ésimo atributo da regra j;

j = é a j-ésima regra (ou indivíduo);

$Valor_i^j$ = é um valor do domínio do atributo i na j-ésima regra;

$Flag_i^j$ = assume um dos valores: 0, 1 ou 2. O valor 0 indica que o atributo i não faz parte do antecedente da regra j (inativo), mas pode ser um conseqüente. O atributo que for conseqüente deve ficar inativo no antecedente. O valor 1 indica que o atributo i faz parte do antecedente da regra j (ativo), e o valor 2 indica que o atributo correspondente está desabilitado, ou seja, o usuário optou por não considerar este atributo;

m = quantidade de atributos relevantes (selecionados manualmente para os experimentos) no banco de dados;

R = quantidade de regras na população.

Cada condição de regra é da forma "Atributo$_i$ = Valor$_{i,j}$". Como o nome do atributo é implicitamente determinado por seu índice i, e como o operador usado na condição é sempre "=", essas informações não precisam ser

codificadas no genoma. Cabe ressaltar que o operador é "=" tanto para atributos categóricos quanto para atributos contínuos fuzzificados, já que nesse último a condição conterá um valor lingüístico – por exemplo, "idade = baixo".

O número máximo de condições ativas em cada regra foi fixado em cinco.

Portanto, em uma população com "R" indivíduos, cada indivíduo representa uma regra, onde cada regra será composta por "m" condições e cada condição será formada por dois genes:

$< Valor_i^j , Flag_i^j >$

onde: i = 1 .. m; j = 1 .. R.

Todos os indivíduos possuem o mesmo número fixo de condições (m), mas cada regra pode ter condições ativas, inativas e até desabilitadas dependendo do valor do gene *Flag* em cada indivíduo. Logo, o número de genes (ativos, inativos ou desabilitados) em cada população é R*(2*m **+ 1)**, conforme ilustra a Figura 38, onde o fator 2 corresponde ao par (valor, Flag) e +1 se refere ao atributo meta (Pos).

6.4 Seleção de Indivíduos

Após alguns experimentos, optou-se pelo método do torneio para seleção de indivíduos, descrito na seção 5.2.5.2, o qual proporcionou populações com bastante diversidade, onde k (o tamanho do torneio) foi definido como sendo igual a 2. Neste método os k indivíduos (k = 2) são escolhidos aleatoriamente da população, e o indivíduo de maior Fitness é escolhido como o vencedor (escolhido para reprodução).

Este método é executado pelo menos duas vezes para cada cruzamento a ser realizado, a fim de selecionar dois pais para o cruzamento: na primeira execução seleciona-se diretamente o vencedor do torneio para ser o primeiro pai e em seguida o operador de seleção é executado repetidas vezes até selecionar um segundo indivíduo pai diferente do primeiro, evitando o cruzamento de indivíduos iguais.

6.5 Definição dos Operadores Genéticos

6.5.1 Operador de Cruzamento

Para explorar o espaço de busca e realizar a mudança do fenótipo das regras, inicialmente utilizou-se cruzamento uniforme baseado na proposta de Goldberg (1989), consistindo na troca de condições ativas dentro das regras. No entanto, o cruzamento implementado na sua forma original mascara a sua probabilidade de ocorrer devido à fragmentação inerente do indivíduo implementado (específico para mineração de dados), ou seja, como há mais atributos inativos do que ativos em cada indivíduo podem ocorrem cortes a direita ou à esquerda dos ativos provocando a troca de todos os valores ativos de um indivíduo para o outro, o que caracteriza um cruzamento mascarado (nenhum gene é trocado). Diante deste fato decidiu-se modificar a rotina do cruzamento adotando uma variante da forma original conforme implementada por Noda (1999). Nesta implementação, além da probabilidade externa de ocorrer cruzamento aplica-se uma probabilidade interna diretamente em cada atributo ativo.

Após alguns experimentos iniciais e ajustes, a implementação do operador de cruzamento ficou da seguinte forma: dada a probabilidade externa (85%, determinado empiricamente), quando ocorrer o cruzamento entre um indivíduo 1 e outro indivíduo 2 a probabilidade interna para cada atributo ativo de cada indivíduo sendo cruzado é de 50%, garantindo uma distribuição uniforme de atuação do operador, ou seja, todos os atributos ativos (com exceção do meta) dos dois indivíduos têm a mesma probabilidade de sofrer a ação do cruzamento. Isto oferece a vantagem de independer da posição em que o atributo foi codificado no indivíduo, uma vez que todas as condições na regra possuem a mesma probabilidade de sofrer a ação deste operador.

Quando um atributo escolhido para cruzamento está ativo em apenas um dos indivíduos selecionados, a operação de cruzamento funciona como especialização de uma regra e generalização da outra regra selecionada. Por exemplo, se um indivíduo selecionado para cruzamento possui dois atributos

ativos (Atr_1 e Atr_3) e o outro indivíduo selecionado possui apenas um atributo ativo (Atr_3) e apenas o Atributo 1 foi escolhido para receber o operador de cruzamento, então ocorre uma generalização do Indivíduo 1, e uma especialização do Indivíduo 2, conforme mostra a Figura 39. No entanto, neste exemplo apenas um atributo ativo sofreu cruzamento. Na prática, dois ou mais atributos ativos podem sofrer cruzamento, e se esses atributos ocorrerem em indivíduos diferentes o efeito de especialização ou generalização associado com a troca de um único atributo pode ser cancelado pela troca de outro(s) atributos(s) ativos.

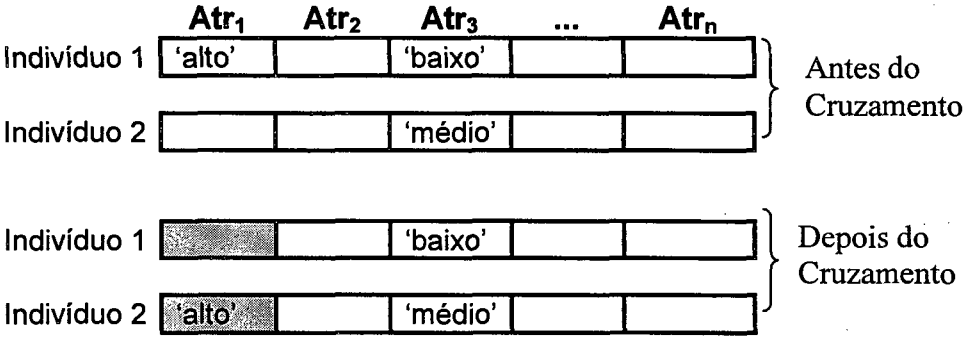


Figura 39: Exemplo de atuação do operador de cruzamento.

Os resultados preliminares com este operador de cruzamento foram satisfatórios, reduzindo-se a um mínimo o número de indivíduos repetidos na população, evitando-se a convergência prematura e obtendo-se mais indivíduos relevantes na população final.

6.5.2 Mutação

O operador de mutação altera o valor de atributos ativos dentro da regra. Sua ação ocorre de forma aleatória apenas no valor de atributos ativos (ignorando os atributos inativos ou desabilitados), mantendo a quantidade de condições ativas inalterada (fenótipo). Ele pode ser aplicado sobre todos os indivíduos da população, menos dois indivíduos copiados por elitismo (que possuem maior valor da função de Fitness), independente da execução ou não da operação de cruzamento, com base em duas probabilidades: genérica e específica.

A probabilidade genérica é de aproximadamente 2% (determinada empiricamente) válida para todos os indivíduos.

A probabilidade específica de ocorrer mutação em um atributo ativo de uma regra é calculada dinamicamente para cada condição ativa dentro da regra em função do tamanho do domínio de cada atributo. Quanto maior o domínio de um dado atributo maior será a probabilidade específica (PrEspec) de ocorrer mutação neste atributo, conforme a seguinte relação:

$PrEspec = (TamDom_i - 1) * Pmutação$

onde: $TamDom_i$ = Tamanho do Domínio de um atributo específico;

Pmutação = Probabilidade genérica de ocorrer mutação.

Portando, se um atributo possui dois valores possíveis (menor domínio existente nos atributos utilizados) a probabilidade específica será igual à probabilidade genérica. Se um atributo possui três valores possíveis a probabilidade específica será duas vezes a probabilidade genérica e assim sucessivamente.

Para escolher um novo valor de atributo diferente do valor atual foi criada uma função que consulta uma tabela contendo o domínio de cada atributo possibilitando a escolha apenas de valores válidos pertencentes ao conjunto de valores possíveis de dado atributo.

6.5.3 Operadores de Inserção e Remoção de Condições

Além dos operadores de mutação e cruzamento utilizaram-se também operadores de inserção e remoção de condições para melhorar a compreensão das regras e aumentar a diversidade da população, possibilitando a exploração de espaços talvez não alcançados pelas operações de cruzamento e mutação. A inserção caracteriza uma operação de especialização da regra, enquanto que a remoção caracteriza uma operação de generalização. Estes operadores atuam depois dos demais, ou seja, após o cruzamento e mutação o antecedente da regra resultante poderá sofrer aleatoriamente uma especialização ou generalização.

O operador de inserção de condições realiza a inclusão de condições nas regras através da alteração do valor do Flag de 0 para 1. A probabilidade de se

realizar a inserção (*Pinsere*) deve ser fornecida pelo usuário. O operador de inserção apenas inclui condições em uma regra nas posições (atributos) que estiverem habilitadas e inativas e que não fizerem parte do conseqüente desta regra.

O operador de remoção realiza o contrário, retira condições das regras através da alteração do valor do Flag de 1 para 0.

O usuário deve definir o número máximo de condições nas regras desejado (*MaxCondAtivas*).

Os operadores de inserção e remoção atuam logo após a execução do operador de cruzamento, conforme segue:

- Se o número de condições ativas for menor que o máximo permitido então o operador de inserção é ativado com a probabilidade (*Pinsere*) fornecida pelo usuário. Nesta aplicação utilizou-se *Pinsere* = 2%, determinado empiricamente;
- Se a quantidade de condições ativas for maior do que um, o operador de remoção é acionado com a probabilidade (*ProbRemover*) fornecida pelo usuário. Nesta aplicação utilizou-se *ProbRemover* = 1%, determinado empiricamente;
- Se não houver pelo menos uma condição ativa em uma regra então o operador de inserção é ativado incondicionalmente (com uma probabilidade de 100%) de forma a garantir que a regra fique com pelo menos uma condição ativa evitando a formação de regras vazias;
- Se a quantidade de condições ativas for maior do que o máximo, o operador de remoção efetuará a retirada incondicional de condições ativas, ou seja, aleatoriamente serão escolhidas condições ativas para serem desativadas até que se atinja o número máximo de condições ativas desejado;
- Se após o cruzamento surgirem dois filhos iguais, há duas situações possíveis. Se o número de condições ativas na regra for menor ou igual a *MaxCondAtivas*/2 então o operador de inserção é acionado para um

filho com 100% de probabilidade de ocorrer. Em caso contrário, se o número de condições ativas for maior do que *MaxCondAtivas*/2, então o operador de remoção é acionado para um filho com 100% de probabilidade. Este mecanismo foi empregado para evitar a aglomeração de indivíduos iguais, incentivando uma maior diversidade da população.

6.6 Atributo Meta e Elitismo

Há pelo menos duas formas de se gerar o conseqüente de uma regra associado com um indivíduo: meta fixado ou meta escolhido dinamicamente.

No primeiro caso (meta fixo), o AG é executado uma vez para cada par de atributos metas e seu valor (denotado abreviadamente meta/valor), não havendo qualquer complicador.

No segundo caso, na população inicial é garantida a criação de pelo menos um indivíduo para cada par meta/valor e após aplicar os operadores genéticos sobre o antecedente da regra, descritos na seção 6.5, o desafio é determinar o atributo meta adequado para o novo indivíduo gerado. Para resolver este problema efetua-se a escolha do par <atributo meta, valor> de forma determinística de forma a maximizar a qualidade da regra, ou seja, é calculado o valor da qualidade da regra para cada meta e escolhido o meta que permite o maior valor da função de qualidade. Esta é basicamente a abordagem seguida em Noda *et al.* (1999).

Esta abordagem provoca o surgimento de outro problema: perda da diversidade genética do conseqüente com a dominação da população por alguns metas específicos. Para resolver este problema pode-se utilizar elitismo em nível de conseqüente.

O elitismo é implementado garantindo, em cada geração, a evolução de pelo menos uma cópia de cada conseqüente, ou seja, a cada geração é transferido para a próxima geração, por elitismo, uma cópia inalterada da melhor regra para cada tipo de conseqüente (para cada meta/valor). Os demais indivíduos da população são gerados através dos operadores descritos na seção 6.5. Assim, quando o processo evolutivo chega ao seu final há pelo menos uma regra para cada meta/valor na população final.

Entretanto, experimentos preliminares revelaram problemas com esta abordagem (meta determinístico) obtendo-se resultados inferiores comparados com a abordagem do meta fixo, razão pela qual se decidiu adotar meta fixo em todos os experimentos relatados.

6.7 Definição dos Conjuntos Difusos

No modelo proposto a lógica difusa foi empregada de forma relativamente simples apenas para validação das regras dentro do AG, de forma híbrida.

Apesar de se adotar uma abordagem simples da lógica difusa, a eficiência da máquina de inferência difusa depende da forma das funções de pertinência. As formas mais empregadas são: triangular, trapezoidal e Gaussiana. Em muitos projetos esta forma é escolhida e mantida fixa durante toda a vida útil do sistema.

Nesta tese adotou-se para cada atributo numérico dois ou três valores lingüísticos representados por duas ou três funções de pertinência, respectivamente, no formato trapezoidal. Neste caso, assumindo três FP's, cinco atributos contínuos e um categórico (sexo), o conjunto de valores possíveis seria: {(A_1 = baixo), (A_1 = médio), (A_1 = alto), ... , (A_5 = baixo), (A_5 = médio), (A_5 = alto), (sexo = M), (sexo = F)}.

O formato trapezoidal pode ser representado de diversas maneiras. Exemplos: representar trapézios simétricos por seu centro, pelo ponto de cruzamento das funções, pelos vértices, ou suas bases inferior e superior.

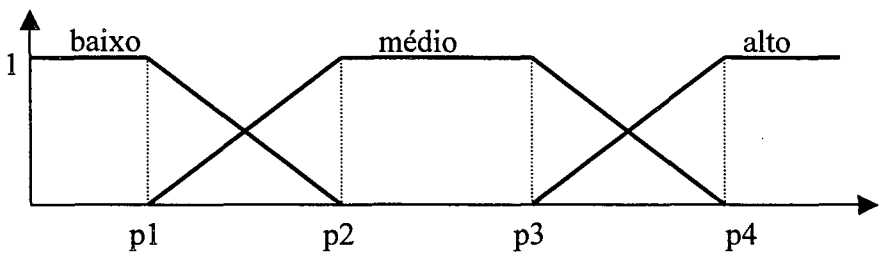


Figura 40: Conjuntos difusos para três termos.

Tomando a base dos trapézios como referência, e impondo-se algumas restrições, é possível representar os três valores lingüísticos de cada atributo

com apenas quatro parâmetros: p1, p2, p3 e p4. Considerando os valores lingüísticos: 'baixo', 'médio' e 'alto', pode-se definir a interação dos conjuntos difusos conforme mostra a Figura 40.

A simetria entre os trapézios é dada pelas seguintes coordenadas:

Ponto de interseção das curvas baixo e médio = $\frac{p1 + p2}{2}$

Ponto de interseção das curvas médio e alto = $\frac{p3 + p4}{2}$

Para obter os valores de p1 .. p4, poder-se-ia:

- empregar alguma técnica de busca local (*Simulate Annealing*; Hill Climbing, etc.) para encontrar a melhor combinação dos vértices; ou
- realizar uma co-evolução de outro AG para realizar a busca da melhor combinação destes parâmetros; ou
- incluir os mesmos dentro dos indivíduos junto com as regras.

Qualquer uma destas alternativas demandaria mais tempo de processamento sem garantia de ganhos significativos nos resultados.

Nesta tese empregou-se o formato trapezoidal para as FP's, mas com uma abordagem que permite a adaptação, pelo usuário, do tamanho da função de pertinência, ou seja, o formato será sempre trapezoidal mas o usuário poderá escolher a extensão de cada conjunto difuso.

Esta abordagem (FP's pré-definidas pelo usuário) possui a desvantagem de perda de generalidade e autonomia, mas no contexto deste trabalho isso não chega a ser uma grande desvantagem, pois o sistema visa uma única aplicação (gestão de C&T). Em contrapartida, esta abordagem apresenta as seguintes vantagens (Ishibuchi & Nakashima, 1999), bastante importantes no contexto deste trabalho:

- incorpora conhecimento do usuário sobre o domínio da aplicação (*background knowledge*);
- reduz consideravelmente o tempo computacional; e
- evita geração de FP's "contra-intuitivas", conforme discutido na seção 5.5 do capítulo anterior.

6.8 Avaliação da Qualidade da Regra

A avaliação de cada regra, que compõe cada indivíduo, é realizada em cada geração. A avaliação das regras é feita segundo dois critérios: Qualidade das regras e Grau de interesse nas regras.

Quadro 9: Matriz de confusão difusa

Exemplo coberto pelo Antec	Meta Igual	
	SIM	NÃO
Sim	Soma grau de correto (SC) de coberturas "sim"	Soma grau de errado (SE) de coberturas "sim"
Não	Soma grau de errado (NE) de coberturas "não"	Soma grau de correto (NC) de coberturas"não"

Para atender o critério da qualidade, construiu-se uma matriz de confusão (ver seção 4.3.3) para cada regra descoberta/construída. Quando o atributo é difuso a soma considera o grau de correto (ou errado) de cobertura, conforme mostra a matriz de confusão difusa do Quadro 9, adaptada a partir da matriz de confusão 'crisp' descrita na seção 4.3.3.

A primeira coluna (Exemplo coberto pelo antecedente) contém a soma de registros cobertos (*Sim*) pelo antecedente da regra e a soma de registros não cobertos (*Não*) pelo antecedente da regra. As colunas "SIM" e "NÃO" fazem separação entre a soma dos registros com valor do atributo meta equivalente ao previsto pela regra e a soma dos registros com valor do atributo meta diferente do previsto pela regra, respectivamente.

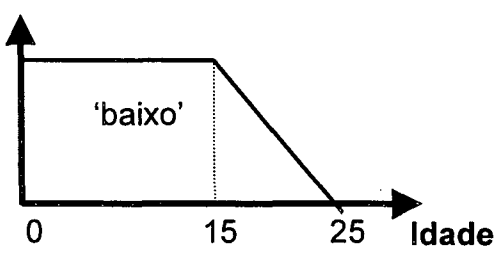


Figura 41: Função de pertinência para Idade = 'baixa'.

Como ilustração, considere-se um exemplo de como calcular uma matriz de confusão difusa para a regra:

(Idade = "baixo"), (Artigos = "alto") ⇒ Classe = 'Sim'

Supondo que o usuário forneceu, como significado do termo difuso "baixo" do atributo Idade, o conjunto difuso dado pela Figura 41.

Supondo ainda que o usuário tenha definido o termo "alto" do atributo Artigos conforme o conjunto difuso da Figura 42.

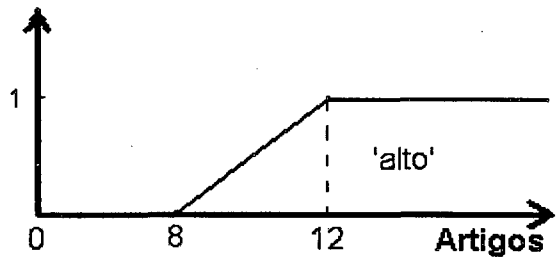


Figura 42: Função de pertinência para Artigos = 'alto'.

Considerando também que a regra exemplo dada acima possui implicitamente uma conjunção, é necessário adaptar para uma conjunção difusa representada pela interseção mínima (veja seção 5.3), ou seja: AND difuso = *min*(cond1, cond2, ...).

Tabela 4: Conjunto de dados fictício.

Reg.	Idade	μ_{Id}	Artigos	μ_{Art}	*min*	Classe
01	15	1.0	12	1.0	1.0	Sim
02	20	0.5	10	0.5	0.5	Sim
03	40	0.0	2	0.0	0.0	Não
04	18	0.7	15	1.0	0.7	Não

Considere-se também o conjunto de dados da Tabela 4, incluindo-se o cálculo das FP's, onde: μ_{ID} = FP do atributo Idade; μ_{Art} = FP do atributo Artigos; *min* = cálculo do AND difuso, dado pela implicação mínima (Equação 5.3).

Para calcular a matriz de confusão toma-se o primeiro registro do conjunto de dados: Idade = 15, Artigos = 12, Classe = "Sim". A conclusão da regra é igual ao valor do atributo meta do registro (Classe = 'Sim'). Logo os valores correspondentes deverão ser incrementados na primeira coluna da matriz de confusão, conforme mostra o Quadro 10. Como o antecedente no registro é

100% igual ao antecedente na regra então o valor de *min* = 1 é lançado na primeira linha.

O segundo registro forneceu *min* = 0.5 e a classe da regra também é igual à classe do registro. Logo se deve lançar 0.5 na posição (1,1) e 0.5 na posição (2.1) da matriz de confusão, ou seja, o complemento de *min* deve ser lançado na segunda linha da mesma coluna dado que o antecedente no registro difere em 50% (em termos difusos) do antecedente na regra.

No terceiro registro, como a classe do registro é "NÃO" o valor de *min* deve ser lançado na segunda coluna, mas na segunda linha pois, o antecedente no registro é completamente diferente (*min* = 0) do antecedente na regra.

Um processamento análogo é aplicado ao quarto registro.

A soma de todas as colunas corresponde ao número total de registros do conjunto de dados de treinamento, que no exemplo é igual a 4.

Quadro 10: Exemplo de matriz de confusão difusa.

Exemplo coberto pelo Antec	Meta Igual	
	SIM	NÃO
Sim	Reg1 + Reg2 = 1.0 + 0.5 = Soma = 1.5	Reg3 + Reg4 = 0.0 + 0.7 = Soma = 0.7
Não	Reg1 + Reg2 = 0.0 + 0.5 = Soma = 0.5	Reg3 + Reg4 = 1.0 + 0.3 = Soma = 1.3
	$\sum = 2.0$	$\sum = 2.0$

Para avaliação de uma regra (indivíduo), ela é aplicada aos dados de treinamento, e o resultado dessa avaliação é dividido em quatro categorias:

SC (Sim Correto) = antecedente cobre o exemplo, meta igual;

SE (Sim Errado) = antecedente cobre o exemplo, meta diferente;

NE (Não Errado) = antecedente não cobre o exemplo, meta igual;

NC (Não Correto) = antecedente não cobre o exemplo, meta diferente.

Em outras palavras:

Se (Valor do atributo Meta na regra) = (Valor do atributo Meta no registro)

então:

- Soma Grau de Correto dos atributos previsores → SC;
- Soma Grau de Errado dos atributos previsores → NE.

Se (Valor do atributo Meta na regra) ≠ (Valor do atributo Meta no registro) então:

- Soma Grau de Errado dos atributos previsores → SE;
- Soma Grau de Correto dos atributos previsores → NC.

As variáveis SC, SE, NE e NC são utilizadas como contadores internos da matriz de confusão difusa. O contador SC, por exemplo, indica a quantidade de exemplos que são corretamente cobertos por uma regra com o mesmo valor do atributo meta, enquanto que o contador SE indica a quantidade de exemplos que são erroneamente cobertos por uma regra que prevê um valor de atributo meta diferente daquele do registro.

Em seguida pode-se calcular a qualidade da regra. Para isto foram realizados experimentos utilizando três tipos diferentes de funções de avaliação da qualidade das regras (descritas na seção 4.3.3). Entre as três funções, os melhores resultados (taxa de acerto, cobertura) foram obtidos com a função Fator de Confiança, dada pela Equação 4.4 e repetida aqui para facilitar a exposição, conforme segue:

$$QUALIDADE = (SC - 1/2)/(SC + SE) \quad (6.1)$$

onde: SC = nº difuso de registros classificados corretamente;

SE = nº difuso de registros classificados erroneamente.

Todos os experimentos com o AGD relatados nesta tese utilizam a Equação 6.1 para o cálculo da qualidade das regras.

Esta função de qualidade (Equação 6.1) penaliza fortemente regras com coberturas muito pequenas através da subtração do fator 1/2 no numerador uma vez que, à medida que a cobertura aumenta, a subtração de 1/2 no numerador não faz praticamente nenhuma diferença.

Em seguida apresenta-se o método utilizado para calcular o grau de interesse na regra.

6.9 Avaliação do Grau de Interesse na Regra

A maioria das pesquisas sobre aprendizagem indutiva tem enfoque em técnicas que visam apenas gerar regras corretas a partir de banco de dados. Nessa abordagem, assume-se que todas as regras geradas serão utilizadas diretamente por um ser humano para inferir soluções para um problema específico de um dado domínio, independente do interesse nas mesmas. Pouca pesquisa tem sido feita sobre o interesse nas regras descobertas.

Não basta apenas obter regras corretas e compreensíveis. Para resolver problemas da vida real é fundamental que o conhecimento descoberto seja, também, de interesse estratégico para o usuário.

> "É *fácil gerar um grande número de padrões a partir de um banco de dados, mas a maioria destes padrões (ou regras) é inútil ou não interessam ao usuário. Além disso, devido ao grande número de padrões, é difícil para o usuário compreender estes padrões e identificar aqueles que são relevantes*" (Liu & Hsu, 1996).

As regras descobertas podem ser numerosas, na ordem de centenas ou até milhares de regras, das quais, muitas inúteis, dificultando para o usuário a escolha das regras relevantes. Além disso, regras diferentes possuem diferentes graus de interesse e este interesse pode variar uma vez que o usuário pode estar interessado em coisas diferentes em momentos diferentes.

Em geral não é adequado retornar todas as regras descobertas. É melhor retornar para o usuário apenas um pequeno conjunto de regras de alta qualidade e de alto interesse (verdadeiras "pepitas" de conhecimento), em vez de um grande conglomerado de regras.

Nesta tese, o modelo proposto retorna apenas a regra mais relevante encontrada para cada um dos valores de atributo meta a serem previstos, mesma abordagem adotada em Noda *et al.*, (1999).

Para descobrir regras relevantes, o sistema deve enfrentar um grande desafio: Como um sistema pode saber o que é novo em um domínio de aplicação e o que é considerado relevante em um momento particular para o usuário?

Para responder a essa questão, foi adaptada uma abordagem subjetiva

(Liu *et al.*, 1997) que considera as impressões gerais do usuário. Uma abordagem deste tipo assume que o usuário inclui, nas impressões gerais, algum conhecimento sobre o domínio e seus interesses em dado momento.

6.9.1 Avaliação Subjetiva do Grau de Interesse

Para atender o critério de interesse nas regras, deve-se medir o quanto elas são surpreendentes. Este é um assunto menos explorado e pode envolver avaliações objetivas e subjetivas. A avaliação objetiva envolve fatores tais como cobertura, confiança e uma medida objetiva de simplicidade. A maioria destes fatores já foi descrita na seção 4.3.

Piatesky-Shapiro & Matheus (1994) observaram que medidas objetivas são úteis mas insuficientes para determinar o interesse nas regras descobertas. É necessário considerar medidas subjetivas. Poucos métodos de avaliação de regras envolvem avaliação subjetiva. Algumas abordagens, disponíveis na literatura, foram exploradas na seção 4.5.2.

A motivação para propor uma técnica subjetiva para induzir regras relevantes provém do fato de que a utilização de uma técnica de aprendizagem não implica que o usuário não sabe nada sobre o domínio e os dados, principalmente neste estudo de caso onde o usuário pode ser um analista de C&T, um pesquisador ou um funcionário de uma agência de fomento a C&T. Normalmente ele tem conceitos pré-concebidos ou conhecimento sobre o domínio de aplicação.

Após a descoberta das regras, naturalmente o usuário poderia tentar responder as seguintes perguntas:

- As regras representam o que eu sabia?
- Se não, que parte do conhecimento prévio é correta e qual é errada?
- Em que sentido as regras descobertas são diferentes do conhecimento anterior?

Além de permitir ao usuário determinar se as novas regras confirmam/recusam conceitos existentes, a presente técnica permite medir o grau de interesse nas regras.

Os métodos existentes para descoberta de conhecimento relevante,

relatados no Capítulo 4 (seção 4.5), fazem isto na forma de um pós-processamento, ou seja, algum algoritmo é aplicado para descoberta de um conjunto de (muitas) regras de previsão e depois outro algoritmo é aplicado sobre este conjunto (em geral de grande tamanho) para extrair as regras relevantes.

Nesta tese apresenta-se uma nova abordagem onde um mesmo algoritmo realiza a busca de regras de previsão, através de um processo evolutivo, e ao mesmo tempo avalia cada indivíduo gerado quanto ao interesse na regra por ele representada (além da avaliação da qualidade da regra), permitindo a evolução apenas de indivíduos que atendam ao interesse do usuário.

Com esta abordagem, a população final fornece apenas regras de interesse do usuário e não um conjunto com muitas regras inúteis mascarando ou até inviabilizando a localização das regras de real interesse.

Para viabilizar a aplicação da técnica proposta, assume-se que o usuário possui algum conhecimento prévio ou hipóteses sobre a área de C&T. Para testar o algoritmo utilizou-se algumas Impressões Gerais obtidas de três usuários potenciais: um pesquisador, o Pró-Reitor de Pesquisa da PUC-PR, o coordenador de Pesquisa também da PUC-PR e o Pró-Reitor de Pesquisa e Pós-Graduação da UEM (Universidade Estadual de Maringá). Estas Impressões Gerais estão descritas no Capítulo 7.

6.9.2 Representação das Impressões Gerais do Usuário

Adaptando a técnica proposta por Liu & Hsu (1996), o conhecimento prévio do usuário é representado por um conjunto de impressões gerais denominado IG, e cada população de novas regras representada pelo conjunto R.

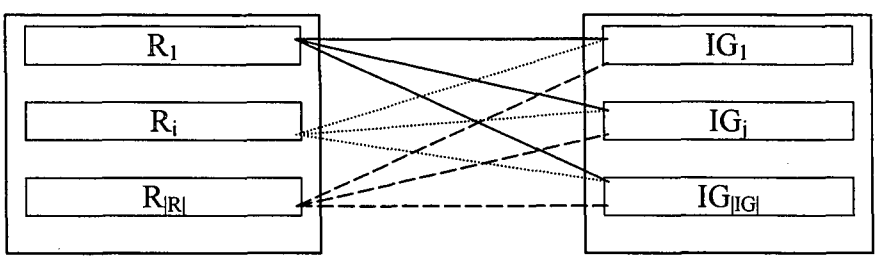


Figura 43: Comparação de cada indivíduo com IG.

Conforme ilustra a Figura 43, para avaliação dos indivíduos interessa saber o grau de similaridade e diferença entre cada regra em R comparada com cada impressão geral existente no conjunto IG. As regras em IG e R devem ter a mesma sintaxe e semântica. Cada regra gerada pelo algoritmo é chamada R_i, enquanto que cada impressão geral (conhecimento prévio) é identificada por IG_j.

Como se utiliza a lógica difusa, importa formalizar alguns conceitos. Segundo Liu & Hsu (1996), uma variável lingüística difusa pode ser representada por cinco termos:

(x, T(x), U, G, $\tilde{M}(x)$)

onde:

x = nome da variável difusa;

T(x) = conjunto de termos de x (valores lingüísticos);

U = universo de discurso de x;

G = regra sintática para gerar o nome dos valores de x;

$\tilde{M}(x)$ = regra semântica para associar significado para cada termo em x, o qual é um subconjunto difuso de U.

Por exemplo, se o atributo "artigos_publicados" for interpretado como uma variável difusa x com U = [0,50], então o conjunto de termos T(x) pode ser:

T(artigos_publicados) = {baixo, médio, alto}.

Neste exemplo, para o termo difuso 'baixo', $\tilde{M}$ (baixo) pode ser:

$\tilde{M}$ (baixo) = $\{(u, \mu_{baixo}(u)) \mid u \in [0,50]\}$

onde:
$$\mu_{baixo}(u) = \begin{cases} 1, & u \in [0, 2); \\ (-u + 6)/4, & u \in [2, 6]; \\ 0, & u \in (6, 50]. \end{cases}$$

$\mu_{baixo}(u)$ expressa o grau de pertinência de u no termo difuso '*baixo*', conforme a Figura 44.

A distribuição dos conjuntos difusos é fornecida pelo usuário, conforme definido na seção 6.7.

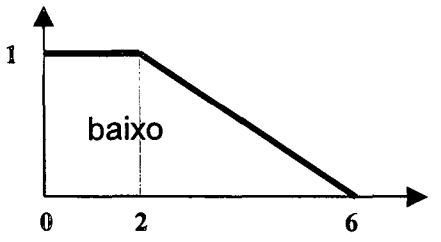


Figura 44: Exemplo de um termo difuso.

Considere, como exemplo, as seguintes regras de previsão descobertas:

R1: (UF = 'RS'), (Idade = 'alta') ⇒ ART_PER_INT = 'medio';

R2: (Nacionalidade='E'), (UF='SC'), (Idade='alta') ⇒ ART_PER_INT ='alto';

R3: (UF = 'PR'), (Idade = 'alta') ⇒ ART_PER_INT = 'baixo'.

Assumindo que o usuário acredita que os pesquisadores idosos que são estrangeiros ('E') em geral publicam bastante artigos em periódicos internacionais (ART_PER_INT), esta impressão geral (hipótese) poderia ser representada pela IG difusa:

IG_j: (Nacionalidade = 'E'), (Idade = 'alta') ⇒ ART_PER_INT = 'alto'.

O significado do termo difuso 'alto' deve ser fornecido pelo usuário.

Dependendo do propósito do usuário, pode-se implementar algoritmos diferentes. Por exemplo, para confirmação de hipóteses, deve-se considerar as regras mais similares às IG's como sendo as melhores. Retomando o exemplo anterior, e considerando a similaridade com IG_j, as regras descobertas poderiam aparecer na seguinte ordem decrescente de preferência:

R2: (Nacionalidade='E'), (UF='SC'), (Idade='alta') ⇒ ART_PER_INT ='alto';

R1: (UF = 'RS'), (Idade = 'alta') ⇒ ART_PER_INT = 'medio';

R3: (UF = 'PR'), (Idade = 'alta') ⇒ ART_PER_INT = 'baixo'.

Neste caso, a regra R2 é bastante equivalente à IG, facilitando a sua escolha baseada no grau de similaridade. A regra R1 é preferida à R3 devido à quantidade de artigos publicados ser maior na primeira (médio > baixo).

Por outro lado, se o propósito é encontrar regras que contradizem as hipóteses, a ordenação resultante deverá ser:

R3: (UF = 'PR'), (Idade = 'alta') ⇒ ART_PER_INT = 'baixo';

R1: (UF = 'RS'), (Idade = 'alta') ⇒ ART_PER_INT = 'medio';

R2: (Nacionalidade='E'), (UF='SC'), (Idade='alta') ⇒ ART_PER_INT ='alto'.

Neste caso, a regra R3 contradiz bastante a hipótese de publicação alta, facilitando sua escolha. Em caso de empate, a escolha é viabilizada pelo valor do grau de interesse. O cálculo do grau de interesse será visto na seção 6.9.4.

Portanto, há duas visões possíveis:

- Usuário quer conhecer as regras similares às IG's;
- Usuário quer conhecer as regras que contradizem as IG's.

Ambas as abordagens exigem a quantificação do grau de similaridade do antecedente da regra, assunto da próxima seção.

Nesta tese optou-se pela segunda abordagem – buscando por regras que contradizem as IG's – a fim de descobrir conhecimento novo e surpreendente (e, portanto, com maior chance de ser relevante) para o usuário.

6.9.3 Cálculo da Similaridade do Antecedente da Regra

Independente do tipo de interesse do usuário (se está interessado em regras similares às IG's ou que contradizem às IG's), importa calcular o grau de similaridade do antecedente, que é utilizado nos dois casos.

Para calcular a similaridade do antecedente (SA) da regra (comparado com cada IG) adaptou-se a equação 4.6, obtendo-se:

$$SA_{(i,j)} = \frac{|A_{(i,j)}|}{\max(|R_i|, |IG_j|)} \qquad (6.2)$$

onde: $|R_i|$ = n.º de atributos ativos no antecedente da regra R_i descoberta;

$|IG_j|$ = n.º de atributos ativos no antecedente da impressão geral IG_j;

$|A_{(i,j)}|$ = n.º de atributos ativos de R_i que são iguais (nome e valor) aos atributos ativos da IG_j.

A Equação 6.2 denota o grau de similaridade do antecedente completo de uma regra (considerando nome e valor dos atributos) e não apenas a similaridade dos nomes dos atributos, como na sua forma original.

Cabe ressaltar que, mesmo o conseqüente podendo assumir qualquer um dos atributos metas (escolhido pelo usuário), sempre o conseqüente terá apenas um atributo meta, já que seria muito difícil descobrir uma regra com

dois ou mais atributos metas no conseqüente (teria que existir um antecedente muito relevante capaz de prever mais de um atributo meta ao mesmo tempo, o que parece ser bastante improvável em alguns casos). Quando isso ocorrer, é mais prático gerar duas regras com mesmos antecedentes e diferentes conseqüentes.

Como ilustração, considere o par R_i (regra) e IG_j (impressão geral):

R_i: (cidade = '5'), (anos_formado = 'baixo') ⇒ artigos_per_nac = 'baixo';

IG_j: (Idade = 'médio'), (anos_formado='baixo') ⇒ artigos_per_nac='baixo'.

O cálculo da similaridade do antecedente, no exemplo acima (aplicando a Equação 6.2), fornece:

$$SA_{(i,j)} = \frac{1}{\max(2,2)} = \frac{1}{2} = 0,5$$

O numerador de $SA_{(i,j)}$ representa a cardinalidade da interseção de atributos no antecedente de R_i e IG_j com o mesmo nome/valor ('anos_formado = 'baixo'). O denominador representa o número de atributos no antecedente de R_i e IG_j respectivamente, extraído o máximo entre eles. O resultado obtido por SA indica que há 50% de similaridade entre os antecedentes de R_i e IG_j.

6.9.4 Cálculo do Grau de Interesse

Quando o usuário está interessado em regras diferentes das IG's, há três situações possíveis (veja seção 4.5.3.2) abordadas por Liu & Hsu (1996):

- antecedente inesperado.
- antecedente contraditório.
- conseqüente contraditório;

Como nesta tese se trabalha com modelagem de dependência, onde o conseqüente pode assumir diversos atributos e diversos valores, adaptou-se a situação de conseqüente contraditório, proposta por Liu & Hsu para a tarefa de classificação. Neste caso, o atributo meta do conseqüente deve ser o mesmo na IG e na regra, mas o valor do atributo meta deve ser diferente.

Nesta abordagem, são consideradas as regras onde o antecedente da IG contém pelo menos uma condição igual no antecedente da regra (tanto o nome

quanto o valor do atributo), mas o valor do conseqüente difere de forma contraditória. Considerou-se que esse caso tem maior potencial para a descoberta de regras altamente surpreendentes para o usuário, mas os demais casos deverão ser considerados em pesquisas futuras.

O cálculo do Grau de Interesse de uma regra R_i é obtido comparando o conseqüente da regra com o conseqüente das IG's. Se o conseqüente da regra (atributo e valor) for igual ao conseqüente de uma IG, a regra é considerada irrelevante com relação àquela IG (Grau de Interesse = 0). Se o nome do atributo do conseqüente da regra for igual ao nome do atributo do conseqüente de uma IG mas o valor for diferente, o interesse nesta regra será maior que zero, e será determinado da seguinte forma: se a diferença entre os valores for de 'baixo' para 'alto', então o interesse será máximo, equivalente à similaridade do antecedente. Em caso contrário, se um dos valores for igual a 'médio' (diferença entre 'baixo' e 'médio' ou entre 'alto' e 'médio'), então o interesse nesta regra será reduzido em 50%.

Em seguida, pode-se então calcular o Interesse em uma regra nova específica R_i dado um conjunto de impressões gerais IG, conforme segue:

$$\text{Interesse} = \text{Max}(SA_{(i,1)}, SA_{(i,2)}, \ldots, SA_{(i,j)}, \ldots, SA_{(i,|iG|)}) \quad (6.4)$$

onde |IG| é o número de impressões gerais especificadas pelo usuário.

Cabe ressaltar que uma regra pode ser considerada relevante mesmo se apenas algumas condições forem equivalentes a uma IG e a conclusão for diferente.

6.10 Cálculo da Função de Fitness

A partir dos resultados dos cálculos da qualidade da regra e do grau de interesse na regra é possível calcular o valor da função de Fitness que será utilizado para avaliação de cada regra gerada.

Inicialmente o cálculo foi feito de forma ponderada, valorizando mais a qualidade da regra, conforme a Equação 6.3.

$$\text{Fitness} = c_1.\text{QUALIDADE} + c_2.\text{INTERESSE} \quad (6.3)$$

onde c_1 e c_2 são pesos definidos pelo usuário.

Os valores de c_1 e c_2 utilizados foram: $c_1 = 2/3$ e $c_2 = 1/3$.

Após alguns experimentos preliminares constatou-se que a Equação 6.3 não é adequada em todos os casos. O problema é sutil e ocorre apenas quando o torneio se dá entre indivíduos com qualidade em torno de 1/20 (0.05), possibilitando um desvio na tendência da evolução para regras piores.

Para demonstrar o problema considere o exemplo da Tabela 5 contendo indivíduos de uma geração intermediária do processo evolutivo:

Tabela 5: Efeito da função de Fitness.

Indiv.	Qualidade	Interesse	Fitness Antiga	Fitness Nova
1	0.01	1.00	0.34	0.010
2	0.01	0.30	0.11	0.003
3	0.05	0.20	0.10	0.010
4	0.09	0.20	0.13	0.018

Quando o torneio se dá entre os indivíduos 1 e 4 a escolha pela Fitness Antiga resultaria no indivíduo 1 (0.34 > 0.13) o que é indesejado: apesar do interesse no indivíduo 1 ser máximo, o indivíduo 4 possui qualidade muito superior apesar do interesse menor.

Semelhantemente, comparando os indivíduos **2** e **3,** a fórmula antiga para o cálculo da Fitness resultaria na escolha do indivíduo 2 enquanto que intuitivamente o indivíduo 3 é superior (qualidade 5 vezes maior apesar do interesse levemente menor).

Esses resultados podem ser considerados indesejáveis pois de nada adianta uma regra ser muito relevante (surpreendente) para o usuário se a regra tem uma precisão preditiva muito baixa.

Para solucionar este problema criou-se uma nova função (produto em substituição à soma) descontínua para o cálculo da Fitness, conforme segue:

```
SE Interesse > 0
  ENTÃO  Fitness = Qualidade * Interesse
  SENÃO  Fitness = Qualidade/20;
```

A descontinuidade é necessária para evitar Fitness = 0 quando o Interesse = 0 (produto da multiplicação), garantindo a consideração da qualidade e a conseqüente evolução das regras sem interesse, mas em um patamar menos significativo. Quando a ordem de grandeza da qualidade é de 1/20 (ou menor) a nova equação para o cálculo da fitness fornece um resultado melhor sem

afetar os demais resultados. Veja na última coluna da Tabela 5 os valores da Fitness com a nova fórmula.

Resultados empíricos demonstraram que, quando a qualidade é da mesma magnitude do Interesse (0.1 a 1), a troca da função de Fitness não provoca alteração nos resultados sendo indiferente o tipo da função de fitness (multiplicação ou soma). No entanto, quando a ordem de grandeza da qualidade é menor ou igual a 1/20 (0.05), a nova função de fitness fornece valores que valorizam mais regras de maior qualidade, apesar da redução no grau de interesse (apenas nos casos onde a nova fórmula deve provocar mudanças).

Na nova equação, onde Fitness = Qualidade*Interesse, o Interesse funciona como um depreciador do valor da qualidade quando o interesse é menor que a unidade.

No caso da descontinuidade, dividir a qualidade por 20 é o mesmo que multiplicar por 0.05. Este número não é arbitrário. Observe que multiplicar por 0.05 não muda os resultados uma vez que não temos Interesse = 0.05, ou seja, não há possibilidade de conflito entre uma regra com interesse e outra regra sem interesse com a mesma Fitness (confirmado empiricamente).

A utilização de 0.05 se justifica por deixar espaço a futuras modificações na rotina de interesse permitindo valores intermediários, talvez 0.1 ou valores próximos.

6.11 Resumo do Modelo Proposto

O Sistema pode ser resumido da seguinte forma:

- O usuário especifica um conjunto de Impressões Gerais;
- O usuário fornece o significado semântico para os valores difusos;
- Obtém-se uma matriz de confusão difusa para cada regra (indivíduo), conforme descrito na seção 6.8, contendo o número de classificações corretas e errôneas feitas pelas regras candidatas (veja Quadro 9);
- Efetua-se o cálculo da qualidade de cada regra;
- Comparam-se cada regra nova com as IG's, considerando regras com

atributo do conseqüente igual ao atributo do conseqüente das IG's, mas com valores diferentes;

- calcula-se o grau de interesse para cada regra;
- calcula-se a Fitness da regra considerando a qualidade e o interesse;
- ao final da evolução mostra-se para o usuário apenas a melhor regra para cada par de atributos metas e seu valor. A Figura 45 ilustra a organização do método.

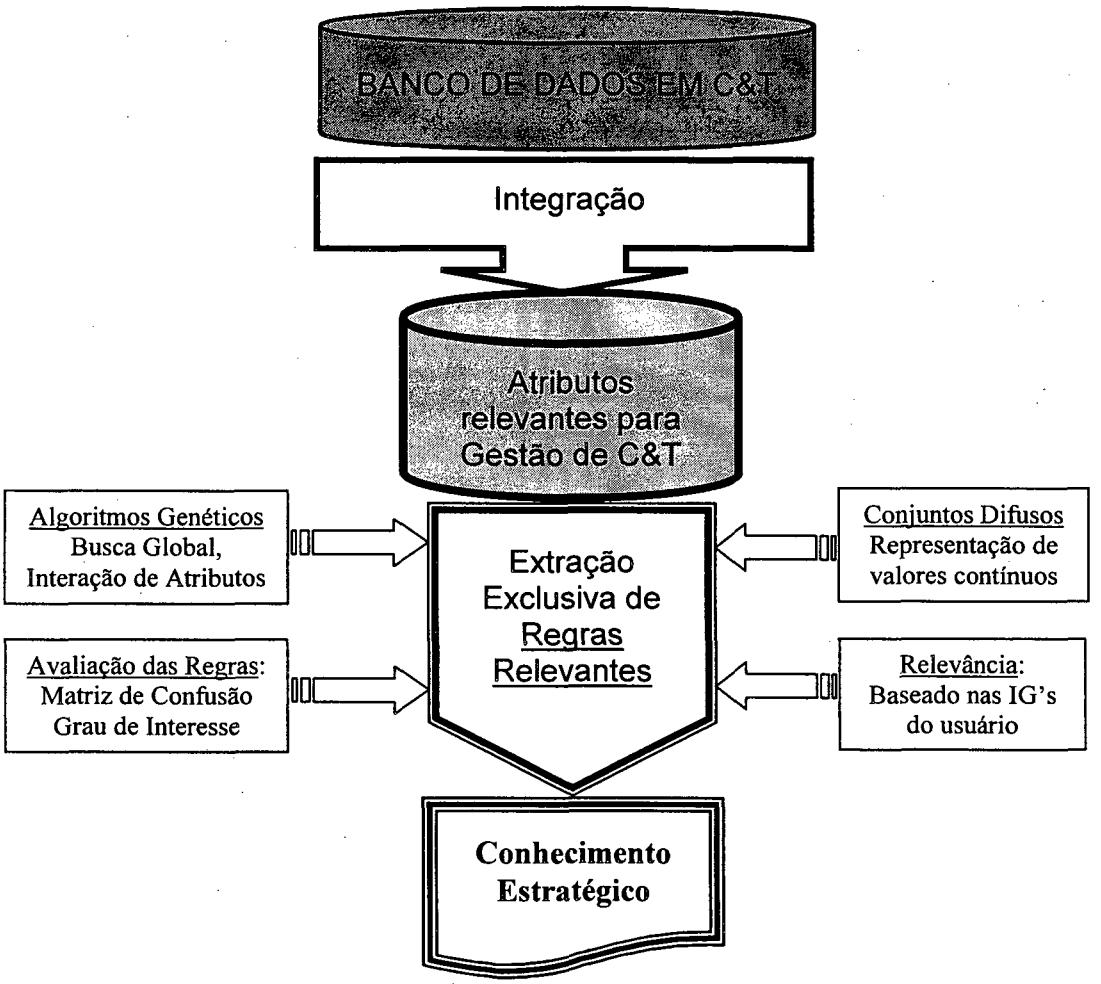


Figura 45: Organização do modelo proposto.

O algoritmo final deverá ter uma estrutura geral conforme o Quadro 11.

No penúltimo passo do algoritmo do Quadro 11, onde se seleciona a melhor regra da população final para ser mostrada para o usuário, há duas exigências: Interesse > 0 e AcertoTreinam > Max(0.5, FreqRelativa). A primeira exigência

(Interesse > 0) é para selecionar apenas regras que tenham grau de interesse maior que zero e, portanto, sejam de algum interesse do usuário.

Quadro 11: Resumo do Algoritmo Proposto

- Obter as IG's do usuário;
- Obter os significados semânticos dos atributos difusos do usuário;
- Repetir para cada par <atributo meta, valor>:
  - Calcular a freqüência relativa do par meta/valor;
  - Formar a população inicial;
  - Computar a qualidade das regras da população inicial;
  - Calcular o grau de interesse das regras da população inicial;
  - Repetir para cada geração:
    - Seleção;
    - Cruzamento;
    - Mutação;
    - Inserção e/ou Remoção de condições nas regras;
    - Computar a qualidade das regras;
    - Calcular o grau de interesse nas regras;
  - Ordenar a população final em ordem descendente de Fitness;
  - Selecionar a melhor regra (maior Fitness) que tenha (Interesse>0) e AcertoTreinam > Max(0.5, FreqRelativa);
  - Mostrar a melhor regra da população final para cada meta/valor;
- Fim do algoritmo.

Na segunda exigência, AcertoTreinam deve ser maior do que o máximo entre 0.5 e FreqRelativa, onde:

$$\text{AcertoTreinam} = \frac{\text{n.º de registros deste meta/valor classificados corretamente}}{\text{n.º de registros com este meta/valor}}$$

$$\text{FreqRelativa} = \frac{\text{n.º de registros com este meta/valor}}{\text{nº de registros total}}$$

Esta segunda exigência [AcertoTreinam > Max(0.5, FreqRelativa)] é para procurar regras que, além de atender a primeira exigência, possua acerto nos dados de treinamento maior do que a freqüência do atributo meta/valor considerado. A motivação para esse critério é que, quanto maior a freqüência relativa de uma classe, mais fácil é prever aquela classe. Assim, para compensar, exige-se uma taxa de acerto maior para classes mais freqüentes.

Caso esta exigência não possa ser atendida, exige-se que o acerto seja pelo menos maior do que 0.5 para garantir que pelo menos 50% dos registros com este meta/valor tenham sido classificados corretamente pela regra corrente.

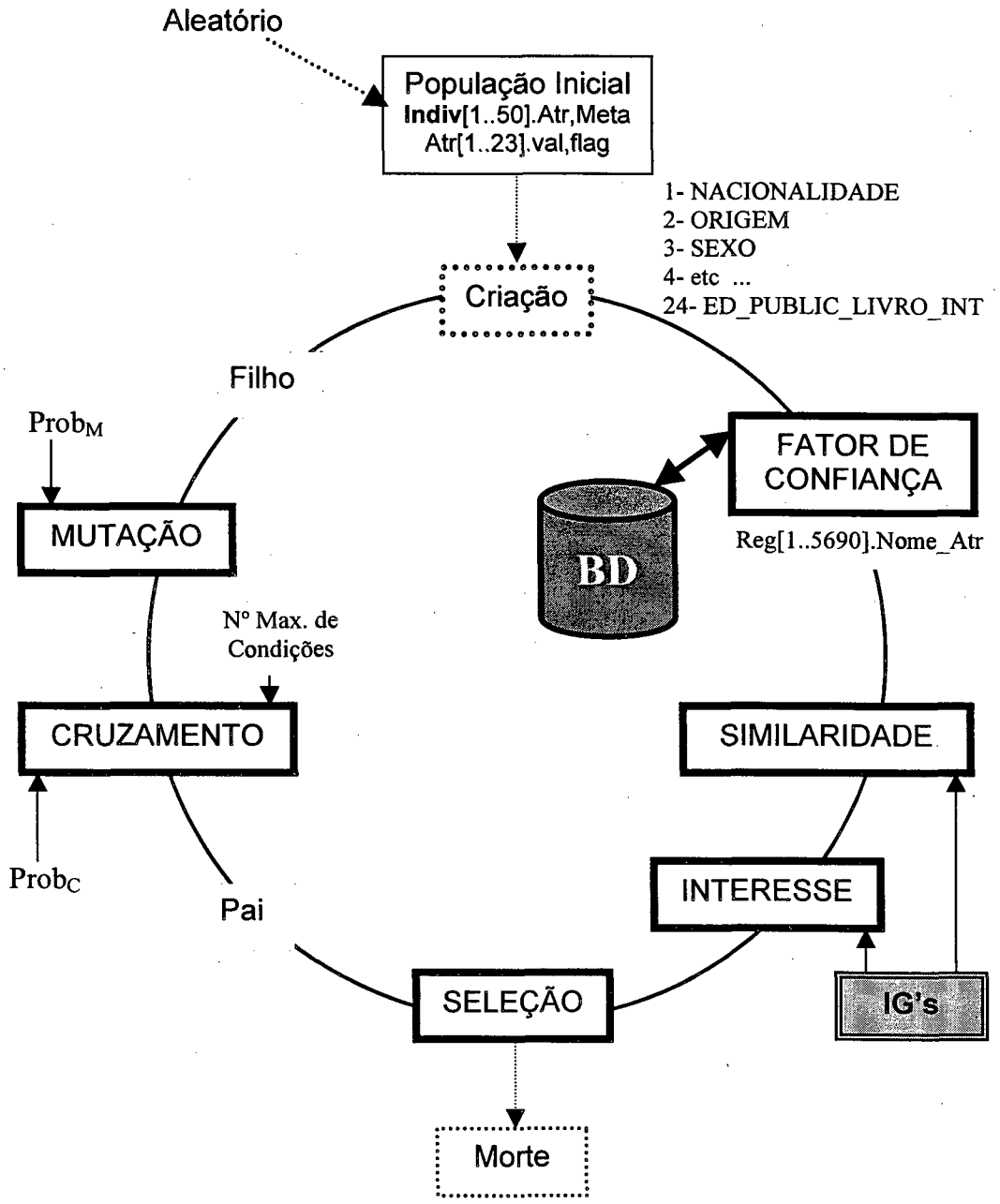


Figura 46: Ciclo de vida de um indivíduo do AGD.

Na Figura 46 apresenta-se o ciclo de vida de um indivíduo.

Um indivíduo da população final pode ser originário da população inicial ou gerado no ciclo de vida de outros indivíduos. Durante o seu ciclo de vida, cada indivíduo é avaliado, através da função de Fitness, e submetido aos

operadores genéticos implementados, podendo passar para a próxima geração ou desaparecer (morte) caso não seja selecionado. Em cada geração é garantida a sobrevivência (elitismo) de pelo menos dois indivíduos de cada meta/valor da geração anterior correspondentes às melhores regras.

6.12 Considerações Finais

Neste capítulo propõe-se um algoritmo genético, denominado AGD, para descoberta de regras de previsão difusas. Foram descritos todos os operadores genéticos implementados e a definição dos conjuntos difusos. Alguns operadores (mutação e cruzamento) foram adaptados e outros (remoção e inserção) foram criados para auxiliar na descoberta de regras de previsão.

Os operadores genéticos convencionais não proporcionaram bons resultados exigindo-se a adaptação dos mesmos no contexto de mineração de dados. Para evitar o problema da convergência para uma única regra (descrito no final da seção 5.2.2) foram implementados dois operadores: inserção e remoção de condições nas regras. Resultados preliminares de execução do protótipo demonstraram que os operadores implementados estão adequados ao objetivo, colaborando para uma maior diversidade da população.

A matriz de confusão "crisp", descrita na seção 4.3.3, foi adaptada para trabalhar com uma abordagem difusa e o método para cálculo da qualidade da regra foi apresentado na seção 6.8.

Destacou-se um dos problemas mais sérios em MD, a interação entre atributos, que muitas vezes não é levada em consideração pelas técnicas convencionais de extração de conhecimento. Isto também justifica a aplicação de AG que possui a habilidade de considerar interações complexas entre atributos.

Na seção 6.9 foi descrito o método para cálculo do grau de interesse na regra, adaptação de um dos métodos descritos na seção 4.5.3. O método leva em consideração as impressões gerais do usuário para guiar o AGD na descoberta apenas de regras relevantes.

Portanto, o algoritmo implementado tem como função descobrir um pequeno número de regras de previsão compreensíveis e surpreendentes capazes de prever o valor de um atributo meta a partir dos valores de atributos previsores, sejam eles difusos ou não.

No processo de KDD há problemas de exatidão, compreensão de resultados e interesse no conhecimento obtido. Estes problemas foram considerados na avaliação das regras descobertas, com destaque ao problema de descoberta de conhecimento relevante. Além disso, a compreensão foi facilitada pelo uso de regras SE... ENTÃO....

Uma ferramenta deste tipo pode ter como público-alvo profissionais ligados à gestão de C&T, os quais podem ser agrupados em dois grandes blocos:

- Usuários diretos, que se constituem de profissionais vinculados às FAP's, tais como: diretores, técnicos, membros de comitês assessores e consultores *ad hoc*;
- Usuários indiretos, que são os profissionais vinculados às instituições executoras de P&D, tais como: universidades (Pró-Reitores de pesquisa e pós-graduação, coordenadores de cursos de pós-graduação, chefes de departamento, pesquisadores, etc.), empresas estatais e privadas que desenvolvam pesquisa.

7 EXPERIMENTOS E RESULTADOS

7.1 Introdução

A implementação de algoritmos de mineração dos dados exige, além dos próprios dados, a interação do analista/programador com o usuário para conhecer o conteúdo, significado e características dos dados disponíveis. Utilizou-se dados da Plataforma Lattes: Diretório de Grupos de Pesquisa 4.0 e Currículos de Pesquisadores e estabeleceu-se a interação com um funcionário do CNPq que trabalha com avaliação e com Pró-Reitores de Pesquisa e Pós-Graduação.

Com relação aos dados, a maior dificuldade foi a seleção dos atributos relevantes, pois só o banco de dados do Diretório de Grupos possui 37 tabelas com dezenas de atributos, além das tabelas do currículo. Para contornar este problema, houve o apoio de analistas do Grupo Stela[2].

Escolheu-se um conjunto de atributos mais comuns de descrição do pesquisador e de sua produção para realizar os experimentos, envolvendo 5 tabelas do Diretório e 4 tabelas do Currículo Lattes.

Os dados da Plataforma Lattes, utilizados nesta pesquisa, foram obtidos junto ao Grupo Stela que possui acordo com o CNPq para utilização restrita à pesquisas e com controle de privacidade das informações.

Cabe esclarecer a terminologia utilizada para diferenciar entre 4 tipos de atributos: *originais, selecionados, candidatos* e *ativos*.

Os atributos *originais* são em grande número e estão nos dados brutos do banco de dados, sem qualquer modificação.

Os *selecionados* são os potenciais candidatos para fins de MD escolhidos manualmente. É um subconjunto do conjunto dos atributos originais.

Os *candidatos* são os atributos fornecidos para o sistema. Eles podem ser o conjunto completo de potenciais candidatos (selecionados) ou podem ser um subconjunto daqueles. Por exemplo, pode-se fazer experimentos utilizando

[2] O Grupo Stela consiste no Laboratório de Desenvolvimento de Sistemas de Informações e de Inteligência Artificial do Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina. Desde 1997 o Grupo Stela é o responsável pelo desenvolvimento da Plataforma Lattes do CNPq.

todos os potenciais candidatos ou fazer experimentos com diferentes números de atributos candidatos. Os candidatos podem também ser atributos selecionados modificados ou derivados daqueles.

Os atributos *ativos* são aqueles selecionados internamente pelo sistema (AGD) como atributos que aparecem nas regras. Se o AG recebeu, por exemplo, 20 atributos candidatos, uma regra pode ter 2 atributos ativos, outra pode ter 5 atributos ativos, etc.

A tarefa mais trabalhosa na preparação dos dados foi a construção de alguns atributos candidatos a partir dos atributos originais. Muitos atributos originais são eliminados diretamente pelo administrador do banco de dados (ex.: CPF e nome do pesquisador, que não são fornecidos por razões de sigilo), outros são eliminados pelo analista por conter poucos registros válidos (ex.: estado civil), outros exigem a simples conversão para valores do tipo inteiro, alguns são derivados de outro atributo (ex.: atributo idade derivado de data de nascimento) e alguns são construídos a partir de vários atributos originais (ex.: artigos publicados em periódicos internacionais).

A construção de atributos candidatos exige tratamento com consultas SQL ao banco de dados e a construção de rotinas que ficam armazenadas no próprio banco de dados. Por exemplo, o atributo selecionado "cod_municipio" da tabela EN_ENDERECO_GR exigiu a implementação de uma função PL/SQL para agrupar em regiões maiores os 66 municípios referenciados no banco de dados, reduzindo para 20 o número de códigos de municípios válidos.

Na seção 7.3 são descritos todos os atributos selecionados para os experimentos realizados, seguido de uma análise de resultados preliminares na seção 7.4.

O AGD foi implementado usando a linguagem de programação Delphi (Reisdorph, 1998; Longo, 1999). Esta linguagem foi escolhida pela facilidade de se programar em um ambiente amigável orientado a componentes e integrado ao banco de dados.

Para comparar a taxa de acerto das regras geradas pelo AGD, adotou-se, como referencial, uma variação do algoritmo C4.5, o algoritmo J4.8. Utilizou-se

uma ferramenta de domínio público conhecida como Weka, contendo o algoritmo J4.8, conforme relatado na seção 4.2.2.1. O J4.8 foi executado para cada atributo meta em cada experimento.

Na seção 7.5 apresentam-se os resultados do primeiro experimento, contendo: as IG's, as regras descobertas, uma análise comparativa (entre o AGD e o J4.8) da taxa de acerto na validação cruzada, uma análise comparativa da compreensibilidade do conhecimento obtido entre os dois algoritmos e resultados subjetivos do interesse nas regras descobertas obtidas através de entrevistas com prováveis usuários.

De forma semelhante, na seção 7.6, apresentam-se os resultados do segundo experimento.

Finalmente, na seção 7.7, encontra-se uma discussão sobre os dois experimentos descritos, seguido pelas considerações finais do capítulo na seção 7.8.

Na seção 7.2, a seguir, apresenta-se uma descrição geral dos dois experimentos realizados.

7.2 Descrição Geral dos Experimentos

Foram realizados dois experimentos a partir de impressões gerais de dois usuários diferentes.

No primeiro experimento utilizaram-se IG's fornecidas pelo co-orientador desta tese, o Prof. Dr. Alex Alves Freitas, pesquisador cadastrado no CNPq. Este primeiro experimento foi realizado a título de avaliação preliminar do sistema, e não refletiu condições ideais de experimentação, tendo em vista que idealmente as IG's deveriam ser fornecidas por um usuário não envolvido nesta pesquisa. Em todo caso, o experimento foi útil para validação do algoritmo. Além disso, a avaliação das regras descobertas foi feita por dois usuários não envolvidos na pesquisa. Mais precisamente, a avaliação do interesse nas regras descobertas foi efetuada pelo Pró-Reitor de Pesquisa e Pós-Graduação, Prof. Dr. Flavio Bortolozzi, e pelo Coordenador de Pesquisa, Prof. Dr. Manoel Camillo Penna Neto, ambos da PUC-PR.

No segundo experimento utilizaram-se IG's fornecidas pelo Pró-Reitor de Pesquisa e Pós-Graduação da UEM (Universidade Estadual de Maringá – PR), Dr. Gilberto Cezar Pavanelli, o qual também avaliou o interesse nas regras descobertas. Este experimento foi considerado mais relevante por ter o mesmo usuário como fornecedor das IG's e avaliador das regras descobertas, o que simula melhor a verdadeira interação do usuário como o futuro sistema completamente implementado.

Com relação aos atributos selecionados, uma diferença entre os dois experimentos é que o atributo GRANDE_AREA foi utilizado apenas no segundo experimento. Os demais atributos considerados foram empregados em ambos os experimentos. A razão pela qual o atributo GRANDE_AREA não foi utilizado no primeiro experimento é simplesmente o fato que esse atributo não estava disponível no momento em que o experimento foi realizado. Além disso, no primeiro experimento não foram utilizados os atributos metas como previsores, enquanto que no segundo experimento todos os atributos selecionados foram utilizados como previsores.

As regras descobertas foram avaliadas quanto aos critérios: acerto, compreensibilidade e interesse, dando-se ênfase ao último critério. O acerto foi medido através de Validação Cruzada (ver seção 6.8), comparando-se o acerto do algoritmo J4.8 com o acerto do AGD. A compreensibilidade foi avaliada comparando-se a quantidade e o tamanho das regras.

O interesse nas regras foi avaliado apenas no AGD, uma vez que apenas este considera este critério no processo de descoberta de conhecimento. O J4.8 não foi projetado para descobrir regras relevantes. Este critério foi avaliado de forma subjetiva pelo usuário através de entrevistas.

A seguir são apresentados os atributos candidatos selecionados com a descrição de seu significado e a forma como foram obtidos.

7.3 Seleção de Atributos

Antes de selecionar os atributos relevantes para a tarefa pretendida, convém avaliar a consistência do banco de dados analisado. Para isto realizou-se um cruzamento entre os dados divulgados pelo CNPq, através de sua

página na Internet, e o banco de dados utilizado, confirmando a consistência dos dados utilizados.

Constatou-se que o banco de dados utilizado, obtido através de uma ferramenta de migração da Oracle, corresponde exatamente à base de dados do CNPq utilizada para publicar estatísticas do Diretório.

O total de grupos de pesquisa da região sul cadastrados é de 2.317 e o total de pesquisadores desta região cadastrados é de 10.378 pessoas.

Na seqüência é apresentada uma descrição das tabelas do banco de dados utilizadas. Foram obtidas junto ao Grupo Stela 39 tabelas das quais apenas 8 foram utilizadas nos experimentos. As tabelas utilizadas foram as seguintes: EN_ENDERECO_GR, EN_FORMACAO, EN_PESQUISADOR_GR, EN_RECURSO_HUMANO, RE_GRUPO_PESQUISADOR, RE_RH_IDIOMA, RE_TOT_PRODUCAO_PESQUISADOR_GR e EN_AREA_PESQUISADOR.

De forma indireta (apenas p/ consulta), objetivando conhecer o significado de alguns códigos utilizados no banco de dados, foram também utilizadas as tabelas: EN_TIPO_PRODUCAO, contendo a descrição da produção codificada, e EN_MUNICIPIO, contendo o nome completo de cada município.

A partir das tabelas originais do banco de dados foram criadas as seguintes tabelas de apoio: FORMACAO_ANO e FORMACAO_NIVEL.

Para facilitar a construção de alguns atributos, foram criadas 19 visões, onde cada visão possui um campo chave confidencial que permite montar o registro completo de um pesquisador através de "joins".

Para selecionar os atributos relevantes foi necessário primeiro escolher a unidade de análise desejada. Havia pelos menos duas possibilidades: "grupo" ou "pesquisador". A relação entre estas duas unidades é do tipo (1 para n), impedindo a aplicação do algoritmo de MD diretamente no banco de dados considerando as duas unidades de análise simultaneamente.

A unidade de análise "grupo" exige a agregação dos dados de todos os pesquisadores pertencentes a um grupo de pesquisa para gerar informação agregada em grupos de pesquisa (exemplo, número total de publicações de todos pesquisadores do grupo).

Foram realizados experimentos utilizando "pesquisador" como unidade de análise, colhendo dados apenas sobre pesquisadores individuais, ignorando-se os dados sobre grupos. Futuramente o algoritmo poderá ser adaptado para realizar experimentos com a unidade de análise grupo.

Um atributo pode conter restrições particulares que reduzem o número total de registros considerados. Por exemplo, registros com valores de atributo ausentes foram eliminados, provocando a redução do conjunto de dados.

Com a restrição dos dados à região sul e a eliminação de registros com valores ausentes há uma redução significativa na quantidade de dados. No entanto, o tamanho relativamente pequeno do conjunto de dados resultante não foi problema uma vez que a quantidade de dados ainda é significativa (não houve "overfitting"). Além disso, o número de atributos foi reduzido através do uso de conhecimento sobre o domínio da aplicação.

Quando se realiza a conjunção das restrições o número de registros considerados reduz-se ainda mais, uma vez que a ocorrência de uma restrição de um atributo em um registro normalmente não coincide com a ocorrência de outra restrição de outro atributo no mesmo registro.

Como exemplo, considerando-se pesquisadores distintos e apenas o atributo IDADE (eliminando-se os registros com data de nascimento em branco) obtém-se 10.217 registros, mas quando se acrescenta o atributo STA_POS_DOUTORADO contendo a situação do pós-doutorado (eliminando-se também os registros com este campo em branco) o número de registros considerados se reduz para 9.919 pesquisadores com registros completos.

Procurou-se selecionar o maior número possível de atributos, aqueles que tivessem um número significativo de registros completos e que fossem potenciais descritores do perfil do pesquisador ou que demonstrasse alguma produção do mesmo. A Tabela 6 apresenta o nome dos atributos selecionados e o nome dos seus respectivos atributos originais e seus domínios.

Após introduzir todos os atributos selecionados obteve-se 5.894 registros completos (sem o atributo GRANDE_AREA), utilizados no primeiro experimento, redução causada principalmente pelo uso de atributos do Currículo juntamente com atributos do Diretório, pois vários pesquisadores

cadastrados no Diretório não estavam cadastrados no Currículo Lattes e vice-versa. Este problema será solucionado na versão 5.0 do Diretório onde todo pesquisador participante de um grupo deverá obrigatoriamente estar cadastrado no Currículo Lattes.

Tabela 6: Atributos originais e atributos construídos.

Atributo Original	Domínio	Tabela Original	Atributo Construído
tpo_nacionalidade	'B', 'E'	EN_PESQUISADOR_GR	NACIONALIDADE
sgl_pais	'AFS', 'ANG', 'BRA', etc,	EN_PESQUISADOR_GR	ORIGEM
cod_sexo	'M', 'F'	EN_PESQUISADOR_GR	SEXO
sgl_uf	todos estados	EN_PESQUISADOR_GR	UF
cod_municipio	7427 a 8973	EN_ENDEREÇO_GR	CIDADE
nvl_escrita	'P', 'R', 'B'	RE_RH_IDIOMA	ESCRITA_INGLES
sta_lider_grupo	'0', '1', '2'	RE_GRUPO_PESQUISADOR	LIDER_GRUPO
cod_area_conhec	*10100008 a 90100000*	*EN_AREA_PESQUISA DOR*	*GRANDE_AREA*
cod_nivel_form	1-9; A-C; X-Z	EN_PESQUISADOR_GR	NIVEL_FORMACAO
ano_obten_form	1900 a 2000	EN_FORMACAO	ANOS_FORMADO
dta_nascimento	xx/xx/1912 a yy/yy/1984	EN_PESQUISADOR_GR	IDADE
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	TRAB_TECNICO
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	CURSOS_MINISTRADOS
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	DISSERT_ORIENT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	TESES_ORIENT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	MONOGR_ORIENT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	TRAB_GRAD_ORIENT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	INIC_CIENT_ORIENT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	ARTIGOS_PER_NAC
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	ARTIGOS_PER_INT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	CAP_LIVROS_NAC
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	CAP_LIVROS_INT
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	ED_PUBLIC_LIVRO_NAC
nro_total cod_tipo_producao	0 a 150	re_tot_producao_pesquisador_gr	ED_PUBLIC_LIVRO_INT

Para o segundo experimento acrescentou-se o atributo GRANDE_AREA causando a redução para 5.690 registros. A Tabela 7 apresenta um resumo dos atributos selecionados e seus respectivos tipos, domínios, discretização, parâmetros dos conjuntos difusos (p1 a p4) - conforme definido na Figura 47 da seção 7.3.4 - e descrição.

Tabela 7: Atributos candidatos selecionados.

	ATRIBUTOS	Tipo	Domínio	p1 .. p4	Discretização	DESCRIÇÃO
1	NACIONALIDADE	Cat P	'B', 'E'			Nacionalidade do pesquisador
2	ORIGEM (contin. de nasc.)	Cat P	'1' a '7'			País onde o pesquisador nasceu → agregado em Continente
3	SEXO	Cat P	'M', 'F'			
4	UF (do grupo do pesquisador)	Cat P	'PR', 'SC' 'RS'			Estado onde está sediado o grupo ao qual o pesquisador pertence
5	CIDADE (do grupo do pesquisad.)	Cat P	'1' a '20'			Principais cidades. As cidades com menos pesquisa foram agrupadas em OUTRAS/estado.
6	ESCRITA_INGLES	Cat P	'P', 'R', 'B'			Nível de ESCRITA (Pouco, Razoável, Bem) em INGLÊS
7	LIDER_GRUPO	Cat P	'0', '1', '2'			Indica se é apenas participante (0), líder (1) ou co-líder (2)
8	GRANDE_AREA	Cat P	'1' a '8'			Indica a principal grande área na qual o pesquisador trabalha **Utilizado apenas no segundo experimento.**
9	NIVEL_FORMACAO	Num P	1 a 6	1,3,3,5	[(1,2) (3) (4,5,6)]	Última formação: Graduaç ... Pós-doutorado, livre_docênc.
10	ANOS_FORMADO (última titulação)	Num P	1 a 85	5,15,15,25	[(1..10) (11..20) (21..85)]	Há quanto tempo concluiu a última formação.
11	IDADE (do pesquisador)	Num P	17 a 89	25,40,50,65	[(17..33) (34..57) (58..89)]	Tempo de vida acumulado.
11	TRAB_TECNICO	Num P	0 a 69	0,1,5,10	[(0) (1..7) (8..69)]	Trabalhos técnicos de assessoria, consultoria, etc.
13	CURSOS_MINISTRADOS	Num P	0 a 40	0,1,3,10	[(0) (1,6) (7..40)]	Cursos de curta duração (extensão, especialização, etc.)
14	DISSERT_ORIENT	Num P	0 a 32	0,1,5,11	[(0) (1..8) (9..32)]	Quant. de dissertações orient. e concluídas no período 97-99
15	TESES_ORIENT	Num P	0 a 11	0,1,3,5	[(0) (1,4) (5..11)]	Quant. de teses orientadas e concluídas no período 97-99
16	MONOGR_ORIENT	Num P	0 a 36	0,1,3,5	[(0) (1,4) (5..36)]	Quant. de monografias orient. e concluídas no período 97-99
17	TRAB_GRAD_ORIENT	Num P	0 a 44	0,1,3,7	[(0) (1..5) (6..44)]	Quant. de trab_grad. orientados e concluídos período 97-99
18	INIC_CIENT_ORIENT	Num P	0 a 30	0,1,3,7	[(0) (1..5) (6..30)]	Quant. de trab_inic_cient. orient. e concluídos período 97-99
19	ARTIGOS_PER_NAC	Num M	0 a 34	0,2,6,14	[(0)(1..10) (11..34)]	Artigos publicados em periódicos nacionais.
20	ARTIGOS_PER_INT	Num M	0 a 56	0,1,3,7	[(0) (1..5) (6..56)]	Idem, internacionais.
21	CAP_LIVROS_NAC	Num M	0 a 39	0,1,2,8	[(0) (1..5) (6..39)]	Capítulos de livros publicados nacionais.
22	CAP_LIVROS_INT	Num M	0 a 15	0,1,1,3	[(0) (null) (1..15)]	Idem, internacionais.
23	ED_PUBLIC_LIVRO_NAC	Num M	0 a 10	0,2,4,6	[(0) (null) (1..10)]	Livro publicado e/ou organizado (edição) nacional.
24	ED_PUBLIC_LIVRO_INT	Num M	0 a 3	0,0,0,1	[(0) (null) (1..3)]	Idem, internacional

Cat P = Categórico Previsor; Num P = Numérico Previsor; Num M = Numérico Meta.

7.3.1 Atributos de Descrição do Pesquisador

A seguir são apresentadas considerações sobre a forma como alguns atributos de descrição do pesquisador foram preparados. Alguns atributos tiveram seus valores agrupados para reduzir o problema de fragmentação e incentivar a descoberta de regras com maior generalidade. As informações básicas de cada atributo encontram-se na Tabela 6 e Tabela 7.

O atributo NACIONALIDADE pode assumir um entre dois valores: 'B' (Brasileiro) ou 'E' (Estrangeiro). Criou-se uma nova tabela (PESQ), introduzindo-se este atributo, além do atributo chave que é utilizado como indíce que permite realizar a junção das tabelas origens e para agilizar o acesso à tabela final.

O atributo ORIGEM corresponde ao país de origem onde o pesquisador nasceu. Devido à alta fragmentação do seu atributo original (sgl_pais), criou-se o atributo ORIGEM como uma generalização do atributo "sgl_pais", agregando os países em continentes e separando-se os pesquisadores nascidos no Brasil.

Cada continente recebeu um número de 1 a 6, e o Brasil (número 7) foi codificado separadamente por incluir a maioria dos pesquisadores, sendo:
1 = África; 2 = América Central; 3 = América do Norte; 4 = Asia/Oceania; 5 = Europa; 6 = América do Sul (excluído o Brasil); 7 = Brasil.

Os registros correspondentes à Oceania foram agregados ao continente asiático por conter apenas uma ocorrência no grupo de dados considerados.

O atributo "UF" indica o estado onde está sediado o grupo de pesquisa, ao qual faz parte o pesquisador, representado pelas siglas 'PR', 'SC' e 'RS', correspondentes aos estados do Paraná, Santa Catarina e Rio Grande do Sul respectivamente.

O nome da cidade correspondente a cada código do atributo CIDADE encontra-se na tabela EN_MUNICIPIO.

O código original de representação deste atributo podia assumir um número muito grande de valores (7427 a 8973), compreendendo 66 cidades da região sul do Brasil. Cada valor cobre poucos exemplos, com pouco poder de

generalização. Para contornar este problema realizou-se uma agregação de valores baseada na concentração de pesquisadores de cada cidade.

No estado do Paraná, as cidades de Curitiba, Guarapuava, Londrina, Maringá e Ponta Grossa foram mantidas sem alteração, enquanto que as demais cidades do Paraná foram agregadas em um único valor de atributo denominado 'outras_pr'.

De forma análoga, no estado do Rio G. do Sul foram mantidas as cidades de Caxias do Sul, Ijui, Passo Fundo, Pelotas, Porto Alegre, Rio Grande, Sta. Cruz do Sul, Santa Maria e São Leopoldo, e as demais foram agrupadas no valor de atributo 'outras_rs'.

No estado de Santa Catarina, as cidades de Blumenau, Florianópolis e Itajai foram mantidas, e as demais cidades deste estado foram agregadas no valor de atributo 'outras_sc'.

A escolha das cidades deu-se baseado no número de ocorrências no banco de dados. Durante esta análise ocorreu a primeira descoberta interessante, apesar de ter sido feita de forma manual (não automática) e ser uma informação relativamente pontual, em vez de uma regra de previsão. Joinville, a maior cidade de SC, tem uma participação incipiente (0,6%) no número de pesquisadores do estado, enquanto que, por exemplo Blumenau, participa com 9%. Isto pode ser considerado conhecimento relevante, uma vez que se esperava o contrário.

Cabe lembrar, porém, que esse resultado se refere apenas à amostra de 5.894 pesquisadores selecionados para os experimentos, conforme explicado anteriormente.

O atributo ESCRITA_INGLES foi construído a partir do currículo Lattes e corresponde ao nível de escrita do pesquisador na língua inglesa especificamente. O seu domínio é (P)ouco, (R)azoavelmente e (B)em.

Existem quatro atributos correspondentes a domínio no idioma em nível de: leitura, conversação, escrita e compreensão, bem como dezenas de idiomas diferentes. Se fossem considerados os quatro atributos relacionados a um idioma, obter-se-ia uma proporção relativamente alta de atributos sobre idioma. Seria um "bias" favorecendo a ocorrência de idiomas nas regras, só

pelo fato de haver muita informação sobre idiomas. Assim, considerou-se apenas um dos atributos: nível de escrita. Este atributo é mais forte que nível de leitura uma vez que, para ter um bom nível de escrita, um bom nível de leitura geralmente é necessário.

Pelo mesmo motivo, decidiu-se colocar informação sobre apenas uma língua estrangeira, inglês, já que o número de artigos científicos publicados em inglês é muito maior do que o de artigos publicados em qualquer outra lingua (ocidental). Logo, a utilização deste atributo é plenamente justificável.

Este atributo exigiu o relacionamento entre a tabela RE_GRUPO_PESQUISADOR, do Diretório, e a tabela EN_RECURSO_HUMANO, do Currículo Lattes, que fazem parte de bancos de dados diferentes (projetos independentes). Isso provocou uma perda significativa de dados.

O atributo GRANDE_AREA apresentou dificuldades na sua geração a partir do Diretório devido ao fato de cada grupo possuir pesquisadores de diversas áreas e até mesmo um mesmo pesquisador atuar em mais de uma área. Para contornar este problema construiu-se este atributo a partir do currículo Lattes selecionando-se a principal área de atuação de cada pesquisador, tendo seu domínio codificado conforme mostra a Tabela 8.

Tabela 8: Domínio do atributo GRANDE_AREA.

Domínio	Atributo GRANDE_AREA	N.º Regs. Brasil	N.º Regs. Região Sul
1	Ciências Exatas e da Terra	8.603	1.722
2	Ciências Biológicas	6.948	1.308
3	Engenharias	6.789	1.260
4	Ciências da Saúde	8.534	1.535
5	Ciências Agrárias	6.880	1.497
6	Ciências Sociais Aplicadas	4.457	1.137
7	Ciências Humanas	8.452	2.167
8	Lingüística, Letras e Artes	2.242	517
Total		52.905	11.143

O atributo NIVEL_FORMACAO foi considerado como um atributo contínuo. O seu domínio original é composto por 15 valores, a saber: 1 = Graduação; 2 = Aperfeiçoamento/Especialização; 3 = Mestrado; 4 = Doutorado; 5 = Pós-

Doutorado; 6 = Livre Docência; 7 = Curso Técnico; 8 = Extensão; 9 = Mestrado profissionalizante; A = Primeiro Grau Incompleto; B = Primeiro Grau Completo; C = Segundo Grau Completo; X = Aperfeiçoamento; Y = Outros; Z = Não Informado.

Entretanto, apenas os valores '1' a '6' foram considerados e transformados em numéricos (1 a 6), os quais representam as formações da maioria dos pesquisadores. Logo, os valores originais 7, 8, 9, A, B, C, X, Y e Z foram ignorados por descreverem formação através de cursos técnicos, cursos de extensão, mestrado profissionalizante, primeiro e segundo grau e outros irrelevantes para descrever formação de pesquisadores.

A tabela EN_FORMACAO contém o nível de formação em vários cursos realizados pelo pesquisador. Criou-se uma tabela, denominada FORMACAO_NIVEL, a qual seleciona apenas a última formação (MAX_NIVEL_FORM) de cada pesquisador, para auxiliar a formação do atributo "nivel_formação".

O atributo ANOS_FORMADO foi derivado do atributo "max_ano_form" obtido da tabela FORMACAO_ANO a qual foi povoada a partir do atributo "ano_obten_form" da tabela EN_FORMACAO. Considerou-se apenas o ano da formação mais recente (max_ano_form) do pesquisador.

Apesar de haver alguns registros com este atributo contendo o valor '1900', considerou-se como domínio '1916' a '2000' (ano da última atualização), já que não se tem conhecimento de pesquisadores centenários. Logo, os registros com este atributo com valor '1900' foram removidos.

Decidiu-se, também, converter o atributo "ano_obten_form" para número de anos desde a última formação (ANOS_FORMADO), uma vez que uma diferença de alguns anos (*e.g.*, 5 anos) na escala de milhares (0 a 2000) é uma diferença minúscula, mas esta mesma diferença na escala de anos desde a última formação (0 a 85) é bastante significativa.

7.3.2 Atributos Previsores de Produção do Pesquisador

O banco de dados do Diretório 4.0 possui uma tabela intitulada RE_TOT_PRODUCAO_PESQUISADOR_GR que incorpora indicadores

totalizadores para toda produção técnica e científica de todos os pesquisadores cadastrados no CNPq, incluindo-se publicações de artigos em revistas nacionais e internacionais, publicação de livros, produções técnicas diversas, orientações de trabalhos diversos, etc. A seguir são descritos os atributos selecionados construídos a partir desta tabela.

O atributo TRAB_TECNICO indica a quantidade de trabalhos técnicos produzidos. O atributo CURSOS_MINISTRADOS indica a quantidade de cursos de curta duração ministrados. O atributo DISSERT_ORIENT indica o número de orientações de dissertações de mestrado concluídas. O atributo TESES_ORIENT indica o número de orientações de teses de doutorado concluídas. O atributo MONOGR_ORIENT indica o número de orientações de monografias de especialização concluídas.

O atributo TRAB_GRAD_ORIENT indica o número de orientações de trabalhos de graduação concluídos e o atributo INIC_CIENT_ORIENT indica o número de orientações de trabalhos de iniciação científica concluídos.

7.3.3 Atributos Metas de Produção do Pesquisador

Dentre os atributos candidatos selecionados, consideraram-se potenciais candidatos a metas os principais atributos que se referem à produção dos pesquisadores. Os atributos escolhidos como meta foram: ARTIGOS_PER_NAC (nº de artigos publicados em periódicos nacionais), ARTIGOS_PER_INT (nº de artigos publicados em periódicos internacionais), CAP_LIVROS_NAC (capítulos de livros nacionais publicados), CAP_LIVROS_INT (capítulos de livros internacionais publicados), ED_PUBLIC_LIVRO_NAC (nº de edições/publicações de livros nacionais) e ED_PUBLIC_LIVRO_INT (nº de edições/publicações de livros internacionais). Os demais atributos (categóricos e difusos) foram utilizados apenas como atributos previsores.

O atributo ARTIGOS_PER_NAC se refere ao número de artigos completos publicados em periódicos nacionais por cada pesquisador. Para facilitar a sua construção, criou-se a visão V_ARTIGOS_PER_COMP_NAC que seleciona o tipo de produção e restringe aos trabalhos de âmbito nacional.

O atributo foi extraído, obtendo-se o mesmo número de registros válidos para os demais atributos, já que uma junção externa de tabelas considerou como zero o valor deste atributo para os pesquisadores que não tinham publicações deste tipo cadastradas no período considerado (1997 a 1999). Efetuou-se a discretização dos valores deste atributo, considerando a freqüência relativa para cada classe (baixo, médio, alto), conforme segue:

Tabela 9: Discretização do atributo ART_PER_NAC.

ValorAtrib	Freqüência	Valor Discretizado
0	2783	baixo
1 a 10	2963	médio
11 a 34	148	alto
TOTAL	5.894	

Na Tabela 9, a primeira coluna indica os valores possíveis para o atributo ARTIGOS_PER_NAC, a segunda coluna mostra a quantidade de registros contendo *ValorAtrib* neste atributo e a última coluna mostra o termo lingüístico utilizado. Cabe ressaltar que, no caso de atributos metas aparecendo no conseqüente de uma regra, os valores 'baixo', 'médio' e 'alto' são considerados como valores discretos "crisp", não difusos. Apenas quando um atributo contínuo ocorre no antecedente de uma regra ele é "fuzzificado", passando a ter valores difusos.

Os atributos metas ARTIGOS_PER_INT e CAP_LIVROS_NAC sofreram discretização de forma análoga.

O atributo CAP_LIVROS_INT foi discretizado em apenas duas classes, devido à baixa freqüência de valores maiores que zero, conforme a Tabela 10.

Tabela 10: Discretização do atributo CAP_LIVROS_INT.

ValorAtrib	Freqüência	Valor Discretizado
0	5503	baixo
1 ou mais	391	alto
TOTAL	5.894	

Os atributos ED_PUBLIC_LIVRO_NAC e ED_PUBLIC_LIVRO_INT também foram discretizados em apenas duas classes pelo mesmo motivo. Eles foram construídos a partir dos atributos "ed_livro_Int", "ed_livro_nac",

"public_livros_nac" e "public_livros_int". Todos estes 4 atributos são muito relevantes para avaliar a produção científica, mas são difíceis de serem previstos dado que poucos pesquisadores possuem valor 'alto' (maior que zero) para estes atributos. Por exemplo, apenas 38 teriam valor 'alto' para "ed_livro_int" enquanto 5856 teriam valor baixo para esse atributo. Isso criaria um sério problema de classes desbalanceadas. Para evitar este problema (cuja resolução não é o foco desta tese), optou-se pela simples abordagem de criar dois novos atributos metas combinando os 4 atributos, ou seja, efetuou-se a soma.

Em geral procurou-se discretizar em 3 valores. A razão para se discretizar alguns atributos metas em 2 valores, em vez de em 3, é que nos atributos em questão há poucos registros com valor de atributo maior do que 0. Assim, a discretização em 3 valores criaria 2 valores (classes) com uma freqüência muito baixa e, portanto, muito difíceis de serem previstos.

7.3.4 Fuzzificação dos Atributos Numéricos

Para viabilizar a aplicação da parte difusa do algoritmo proposto os atributos previsores devem ser fuzzificados definindo-se os conjuntos difusos associados a cada um deste atributos. Esta especificação pode ser baseada na distribuição relativa dos dados entre as classes ou o usuário pode especificar a Função de Pertinência (FP) associada com cada termo linguístico (baixo, médio, alto) de cada atributo previsor, baseado na Figura 47. O formato dos trapézios são especificados através dos parâmetros p1 a p4, conforme descrito na seção 6.7.

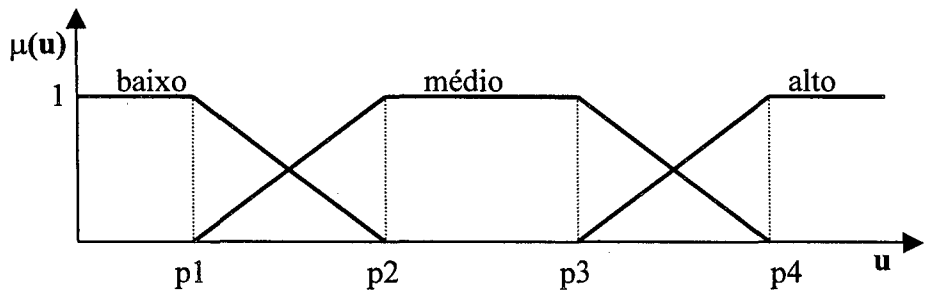


Figura 47: Funções de Pertinência Trapezoidais.

Considere-se:

u = Valor numérico do atributo;

$\mu(u)$ = Grau de Pertinência (valor contínuo) $\in[0, 1]$;

Deve-se construir FP's que resultem em uma distribuição balanceada de trapézios e semi-trapézios, ou seja:

$\mu_{baixo}(u) + \mu_{medio}(u) + \mu_{alto}(u) = 1$.

Considerando que os formatos para as FP's se resumem em retas horizontais, retas ascendentes (Figura 48) e retas descendentes (Figura 49), pode-se facilmente generalizar a implementação dos trapézios, da seguinte forma:

Retas horizontais → $\mu(u) = 1$;

Retas ascendentes → $\mu(u)$ = (u - Início)/(Limite – Inicio);

Retas descendentes → $\mu(u)$ = (Limite - u)/ (Limite – Inicio);

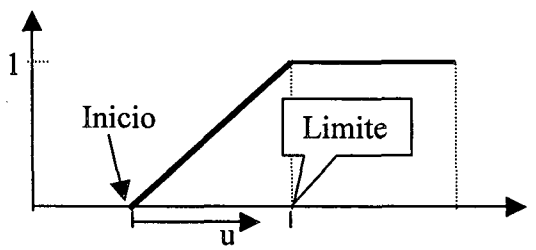


Figura 48: Reta Ascendente.

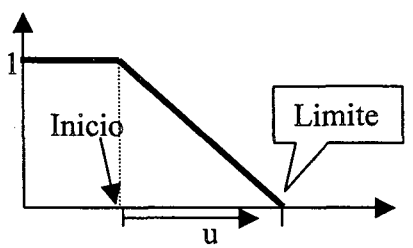


Figura 49: Reta Descendente.

Tomando como base a discretização fundamentada na freqüência relativa de cada classe e o bom senso, definiu-se FP's para cada atributo numérico, conforme mostra a quinta coluna (p1 .. p4) da Tabela 7 da seção 7.3.

7.4 Análise Preliminar de Resultados

Após selecionar todos os atributos relevantes para a tarefa pretendida obteve-se uma tabela com 23 atributos, sendo 7 categóricos previsores e 16 numéricos (10 previsores e 6 metas), contendo 5.894 registros completos sobre pesquisadores da região sul do Brasil.

A população inicial de cada execução do AG é gerada aleatoriamente, ou seja, cada condição dentro de cada regra (atributos e valores) é gerada de forma aleatória conforme a hora do relógio do sistema. Foi realizado um experimento utilizando as IG's como parte da população inicial supondo que isto poderia melhorar os resultados, mas os resultados não confirmaram esta hipótese e a conclusão em que se chegou é que este procedimento apenas diminui a aleatoriedade, não trazendo qualquer benefício.

Foram realizadas várias execuções preliminares do programa AGD, considerando quatro tipos de experimentos:

- Meta escolhido de forma determinística (veja seção 6.6) onde todos atributos podem ser previsores;
- Meta escolhido de forma determinística onde os atributos metas não podem ser previsores;
- Meta fixo (veja seção 6.6) onde todos atributos podem ser previsores;
- Meta fixo onde os atributos metas não podem ser previsores.

Os experimentos executando o AGD para todos os metas ao mesmo tempo (de modo que o meta previsto por uma regra é escolhido de forma determinística) forneceram resultados inferiores quando comparado com o AGD executado uma vez para cada meta fixo. Com isto decidiu-se utilizar apenas meta fixo nos experimento relatados nesta tese.

Na fase de teste, os primeiros experimentos foram realizados utilizando o método "Hold Out" (seção 4.3.1.1) para avaliação das regras descobertas. No entanto, após escolher o J4.8 para comparação de resultados, decidiu-se trocar Hold Out por Validação Cruzada, mesmo método utilizado por *default* no J4.8. Além disso, a validação cruzada permite utilizar todo o conjunto de dados para

treinamento e teste, mantendo o conceito de separação dos dados conforme descrito na seção 4.3.1.2.

Para realizar a validação cruzada, utilizaram-se todos os dados, divididos em dez partes, para estimar a taxa de acerto em cada classe. Para calcular o acerto nos dados de teste adaptou-se a Equação 6.1, descrita na seção 6.8, retirando o fator –1/2, obtendo-se a Equação 7.1.

$$\text{Acerto} = SC/(SC + SE) \qquad (7.1)$$

Na Equação 7.1, o numerador representa o número de registros que satisfazem o antecedente e o conseqüente da regra, e o denominador representa o número total de registros que satisfazem o antecedente da regra.

Cabe ressaltar que a Equação 7.1 é a forma original da Equação 6.1 (veja seção 4.3.2). O fator –1/2, existente na Equação 6.1, é utilizado apenas no cálculo do acerto nos dados de treinamento para penalizar as regras de baixa cobertura e auxiliar o processo evolutivo a evitar regras muito especializadas. Na fase de teste não se deve favorecer regra alguma, mas avaliar o real acerto nos dados de teste, razão pela qual não se utilizou o fator –1/2 na Equação 7.1.

Um problema que surgiu durante os experimentos preliminares foi como decidir quais regras, dentre todas aquelas descobertas pelo sistema (presentes na populacao final), deveriam ser efetivamente consideradas como "regras descobertas" e, portanto, mostradas ao usuário. Considerando que em cada geração é copiada por elitismo as duas melhores regras (maior fitness) da geração anterior, decidiu-se mostrar apenas a melhor regra da população final de cada classe (cada valor de cada atributo meta). Com isto foram obtidas 15 regras correspondentes às 15 classes consideradas nos experimentos – lembre-se que há três atributos metas com três valores e três atributos metas com dois valores, ou seja, 3x3 + 3x2 = 15 classes.

Em suma, para todos os experimentos com o AGD relatados, as regras selecionadas para serem mostradas para o usuário formam um conjunto de regras constituído pela melhor regra de cada par atributo meta/valor, descobertas a partir do conjunto completo de dados, com interesse maior que zero e acerto (nos dados de treinamento) maior que o máximo entre [0.5, Freqüência relativa da classe].

7.5 Primeiro Experimento

Neste experimento os metas não foram considerados como previsores. Este foi um experimento menos importante que o segundo (seção 7.6). Neste primeiro experimento, não se considerou o atributo GRANDE_AREA: a pessoa que formulou as IG's não sabia da existência deste atributo na época. Além disso, a pessoa que forneceu as IG's não avaliou os resultados. A avaliação foi realizada por outras pessoas, conforme descrito na seção 7.2.

Quadro 12: Impressões gerais (Prof. Alex).

IG[1]: SE ESCRITA_INGLES = B, TESES_ORIENT = alto
ENTÃO ARTIGOS_PER_INT = alto
IG[2]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO ARTIGOS_PER_INT = alto
IG[3]: SE ESCRITA_INGLES = P, DISSERT_ORIENT = baixo
ENTÃO ARTIGOS_PER_INT = baixo
IG[4]: SE ESCRITA_INGLES = B, TESES_ORIENT = alto
ENTÃO CAP_LIVROS_INT = alto
IG[5]: SE TESES_ORIENT = alto ENTÃO ARTIGOS_PER_NAC = alto
IG[6]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO ARTIGOS_PER_NAC = alto
IG[7]: SE DISSERT_ORIENT = baixo ENTÃO ARTIGOS_PER_NAC = baixo
IG[8]: SE NIVEL_FORMACAO = baixo ENTÃO ARTIGOS_PER_NAC = baixo
IG[9]: SE TESES_ORIENT = alto ENTÃO CAP_LIVROS_NAC = alto
IG[10]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO CAP_LIVROS_INT = alto
IG[11]: SE ESCRITA_INGLES = P, DISSERT_ORIENT = baixo
ENTÃO CAP_LIVROS_INT = baixo
IG[12]: SE NIVEL_FORMACAO = baixo ENTÃO CAP_LIVROS_INT = baixo
IG[13]: SE NIVEL_FORMACAO = médio ENTÃO ARTIGOS_PER_INT = baixo
IG[14]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO CAP_LIVROS_NAC = alto
IG[15]: SE DISSERT_ORIENT = baixo ENTÃO CAP_LIVROS_NAC = baixo
IG[16]: SE NIVEL_FORMACAO = baixo ENTÃO ED_PUBLIC_LIVRO_NAC = baixo
IG[17]: SE ESCRITA_INGLES = B, TESES_ORIENT = alto
ENTÃO ED_PUBLIC_LIVRO_INT = alto
IG[18]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO ED_PUBLIC_LIVRO_INT = alto
IG[19]: SE ESCRITA_INGLES = P, DISSERT_ORIENT = baixo
ENTÃO ED_PUBLIC_LIVRO_INT = baixo
IG[20]: SE TESES_ORIENT = alto ENTÃO ED_PUBLIC_LIVRO_NAC = alto
IG[21]: SE NIVEL_FORMACAO = baixo ENTÃO CAP_LIVROS_NAC = baixo
IG[22]: SE NIVEL_FORMACAO = alto, ANOS_FORMADO = alto
ENTÃO ED_PUBLIC_LIVRO_NAC = alto
IG[23]: SE DISSERT_ORIENT = baixo ENTÃO ED_PUBLIC_LIVRO_NAC = baixo
IG[24]: SE NIVEL_FORMACAO = baixo ENTÃO ED_PUBLIC_LIVRO_INT = baixo

No Quadro 12 apresentam-se as IG's (Impressões Gerais) especificadas para este experimento, sugeridas pelo Prof. Alex A. Freitas, professor pesquisador na PUC-PR. Conforme explicado na seção 6.9.1, as IG's são subjetivas.

Neste primeiro experimento não se utilizou os metas como previsores, enquanto que no segundo experimento (relatado na seção 7.6) todos os metas são considerados como possíveis previsores.

7.5.1 Acerto das Regras Descobertas no 1º Experimento

Antes de relatar e analisar os resultados nos dados de teste cabem algumas considerações sobre o algoritmo proposto e o algoritmo utilizado como referência para comparação, o J4.8. As regras são obtidas pelo AGD considerando tanto a taxa de acerto quanto o interesse nas regras, enquanto que o J4.8 considera apenas a taxa de acerto na montagem da árvore de decisão, ignorando o interesse. Isto requer uma análise da natureza do J4.8 e do AGD através de alguns questionamentos:

O que o J4.8 deve descobrir? Regras de alta taxa de acerto.

O que o AG deve descobrir? Regras com alta taxa de acerto e relevantes (inesperadas).

Portanto, para realizar uma comparação justa do resultado dos dois algoritmos modificou-se a função de fitness do AGD (apenas nestes experimentos) para considerar apenas a taxa de acerto, e não o interesse da regra sendo avaliada.

Isso permitiu uma comparação direta da qualidade do conhecimento extraído entre os dois algoritmos, conforme mostra a Tabela 11. Nesta tabela apresenta-se: a freqüência relativa de cada classe em todos os registros, a percentagem de acerto com o J4.8, a percentagem de acerto e a cobertura com o AGD para cada classe. Os experimentos foram feitos com validação cruzada. Cabe ressaltar que tanto a taxa de acerto quanto a cobertura mostradas na tabela foram computadas sobre os dados de teste.

Os resultados da quinta coluna da Tabela 11 se referem à média ($\bar{a}$) e desvio padrão (*dp*) dos acertos em todas as partições calculados conforme as

equações descritas na seção 4.3.1.2. Os resultados da última coluna se referem à média e ao desvio padrão das coberturas calculadas utilizando as mesmas equações utilizadas para o cálculo do acerto, fazendo x_i = Cobertura em cada partição.

Um valor em negrito na quarta ou quinta coluna da Tabela 11 indica que a taxa de acerto do J4.8 ou do AGD, respectivamente, foi a maior.

Tabela 11: Taxa de acerto na validação cruzada – 1º experimento.

Conseqüente			J4.8 Qualidade	AGD Apenas Qualidade	
Atrib. Meta	**Classe**	**Freq. Relat. (%)**	**Acerto (%)**	**ACERTO (ã% ±DP)**	**Cobertura (ã ±DP)**
ARTIGOS_ PER_NAC	baixo	47.2	57.3	**76.7 ±13.2**	1.75 ±0.4
	médio	50.3	**57.5**	54.0 ±15.8	0.51 ±0.2
	alto	2.5	**14.3**	10.0 ±10.0	0.03 ±0.0
ARTIGOS_ PER_INT	baixo	64.7	74.0	**85.6 ±10.1**	1.58 ±0.3
	médio	29.2	**46.7**	18.1 ±7.6	1.06 ±0.4
	alto	6.0	**30.0**	26.7 ±12.7	0.59 ±0.2
CAP_LIVROS_ NAC	baixo	77.2	79.6	**98.6 ±1.4**	3.08 ±0.7
	médio	20.9	39.6	**48.9 ±16.3**	0.36 ±0.1
	alto	2.0	**22.8**	10.0 ±10.0	0.35 ±0.1
CAP_LIVROS_ INT	baixo	93.4	93.4	**98.5 ±1.1**	18.88 ±2.0
	alto	6.6	0.0	**10.0 ±10.0**	0.21 ±0.2
ED_PUBLIC_ LIVRO_NAC	baixo	83.7	84.0	**89.1 ±9.9**	2.38 ±0.4
	alto	16.3	**33.1**	0.0 ±0.0	0.17 ±0.1
ED_PUBLIC_ LIVRO_INT	baixo	97.9	97.9	**99.2 ±0.5**	28.91 ±5.3
	alto	2.1	0.0	**10.0 ±10.0**	0.32 ±0.1

Para realizar a validação cruzada o conjunto dos dados foi dividido em 10 subconjuntos iguais. Para isto foi criada uma rotina que faz a divisão do conjunto de dados de forma aleatória.

Fazendo uma comparação simples das taxas de acerto entre o J4.8 e o AGD tem-se apenas seis classes com taxas superiores no J4.8, enquanto no AGD tem-se nove classes com acerto superior. Há que se considerar que o desvio padrão do acerto em algumas classes torna a diferença insignificante, mas mesmo assim tem-se uma maior quantidade de classes na qual o AGD obteve acerto superior comparado ao J4.8.

Observa-se que o AGD supera o J4.8 sempre que a freqüência da classe é alta, enquanto que o J4.8 obteve melhor desempenho nas classes de baixa freqüência. Entretanto, em alguns casos críticos (atributos CAP_LIVROS_INT

e ED_PUBLIC_LIVRO_INT) o J4.8 não conseguiu classificar adequadamente, limitando-se a prever para todos os registros a classe da maioria – ou seja, obtendo taxa de acerto zero para classe da minoria.

Os baixos valores obtidos para Cobertura revelam a natureza de pequenos disjuntos dos padrões, além de representar a cobertura média de apenas um décimo (conjunto de teste) do conjunto dos dados.

7.5.2 Compreensibilidade das Regras Descobertas

A compreensibilidade do conhecimento obtido pode ser avaliada pelo número de condições em cada regra ou pelo tamanho da árvore de decisão. Quanto menor a regra, ou menor o número de nós na árvore, mais fácil a sua compreensão.

Tabela 12: Simplicidade sintática (metas não previsores).

	Atributo Meta	J4.8 N.º de Nós	Classe	AGD N.º Condições
1	ARTIGOS_ PER_NAC	872	baixo	3
2			médio	5
3			alto	4
4	ARTIGOS_ PER_INT	820	baixo	2
5			médio	3
6			alto	5
7	CAP_LIVROS_ NAC	388	baixo	4
8			médio	4
9			alto	4
10	CAP_LIVROS_ INT	01	baixo	2
11			alto	5
12	ED_PUBLIC_ LIVRO_NAC	162	baixo	1
13			alto	2
14	ED_PUBLIC_ LIVRO_INT	01	baixo	3
15			alto	4

A Tabela 12 revela a simplicidade sintática das regras descobertas pelo AGD e a complexidade sintática da árvore gerada pelo J4.8. Os dados da última coluna desta tabela foram obtidos executando o AGD sobre todos os dados disponíveis, mas utilizando Fitness = Qualidade, ignorando o interesse nas regras obtidas, permitindo uma comparação direta da compreensibilidade sintática das regras extraídas pelos dois algoritmos. Deve-se lembrar que cada regra gerada pelo AGD pode conter no máximo cinco condições.

Com estes resultados fica demonstrado que, em geral, o J4.8 fornece uma árvore de decisão, para cada atributo meta, com grande número de nós, além de gerar regras com muitas condições, dificultando a localização de regras relevantes e, portanto, reduzindo a compreensibilidade. As exceções são as árvores para os atributos metas CAP_LIVROS_INT e ED_PUBLIC_LIRO_INT, as quais são árvores degeneradas, contendo um único nó que prevê a classe da maioria para todos os registros.

No outro extremo, os resultados da Tabela 12 revelam que o AGD descobre conhecimento mais compreensível devido ao baixo número de regras (apenas 15 regras) e ao baixo número de condições em cada regra. Ele descobre apenas a melhor regra para cada classe (cada par atributo meta/valor), garantindo sempre um número reduzido de regras, facilitando a análise por parte do usuário, enquanto que a árvore do J4.8 gerou regras com até mais de 15 condições.

Porém, uma comparação entre a compreensibilidade das regras geradas pelo J4.8 e AGD não é inteiramente justa, pois o J4.8 tenta gerar regras cobrindo todo o espaço de dados, enquanto o AGD gera um número menor de regras cobrindo uma pequena parte do espaço de dados. Uma alternativa, para melhorar os resultados com o J4.8, é realizar a poda da árvore e também restringir o número de condições em cada regra.

7.5.3 Regras Descobertas no 1º Experimento

A fim de descobrir as regras finais a serem mostradas ao usuário, o programa AGD foi executado com todos os dados, sem o atributo GRANDE_AREA e sem utilizar os atributos metas como previsores, perfazendo um total de 15 chamadas à rotina do AG (uma para cada meta/valor) contendo uma população de 100 indivíduos que evoluem 60 gerações em cada chamada. O tempo total de processamento foi de aproximadamente 7 (sete) minutos utilizando um microcomputador Pentium III de 866 Mhz e 256 MB de memória.

No ANEXO B são apresentadas as regras extraídas pelo AGD no primeiro experimento, acompanhadas do grau de interesse medido pelo sistema,

número da IG fornecida pelo usuário, freqüência relativa da classe no conjunto total de dados e a cobertura no conjunto total de dados. As regras descobertas podem ser classificados em dois grupos: regras com cobertura muito baixa (pequenos disjuntos) e regras com boa cobertura.

As regras 1, 5, 6, 8, 9, 10, 11 e 13 foram consideradas "pequenos disjuntos" (primeiro grupo) por apresentarem cobertura inferior a 5.7 registros (0.1%) do conjunto total dos dados. Este critério foi escolhido empiricamente ao se verificar que algumas regras descobertas apresentavam cobertura muito abaixo deste percentual, enquanto que outras regras apresentavam cobertura bastante acima deste percentual. Isso induziu a escolha deste número como um parâmetro de referência, baseado no tamanho do conjunto de dados, para identificação relativa de "pequenos disjuntos" no contexto desta aplicação.

As regras do tipo pequenos disjuntos são potencialmente relevantes por representarem exceções (outliers), podendo revelar a existência de pesquisadores com características especiais ou indicar a existência de erro nos dados.

As demais regras (2, 3, 4, 7, 12, 14 e 15) possuem cobertura superior a 0.1% dos dados e foram incluídas no segundo grupo.

As regras descobertas podem ser analisadas isoladamente ou por comparação com as IG's. Analisando algumas regras isoladamente observa-se o seguinte:

A regra 1, por exemplo, revela alguns poucos pesquisadores que deveriam ter alta produção devido à alta titulação obtida há mais de 20 anos, mas na verdade não publicaram qualquer artigo em periódicos nacionais, caracterizando uma exceção. No entanto, como esta regra possui baixa cobertura ela representa exatamente uma exceção, logo não é muito surpreendente.

Por outro lado, a regra 4 revela que os pesquisadores em geral (regra com alta cobertura) com nível de formação alto obtido há mais de 20 anos possuem baixa publicação de artigos em periódicos internacionais quando sua habilidade de escrever em inglês é baixa. Este conhecimento pode ser considerado relevante e servir de subsídio, por exemplo, para se implantar um programa de

custeio de cursos de inglês para estes pesquisadores e aproveitar a boa formação já adquirida por eles e aumentar a quantidade de publicações internacionais.

A análise comparando cada regra descoberta com a correspondente IG contrariada, realizada através de entrevistas com usuários, é apresentada na próxima seção.

7.5.4 Interesse nas Regras Descobertas no 1º Experimento

Os dois grupos de regras descobertas, consideradas relevantes, foram mostrados aos usuários entrevistados para avaliação em dois formulários. Nestes formulários, cada usuário respondeu entre três conceitos diferentes (baixo interesse, interesse médio ou alto interesse) para cada regra descoberta pelo AGD.

Veja no ANEXO C – Entrevistas para Avaliação das Regras, respostas da avaliação das regras com cobertura baixa (Formulário A1) e respostas da avaliação das regras com cobertura alta (Formulário A2) obtidas na entrevista 1 com o Prof. Dr. Flavio Bortolozzi.

Tabela 13: Interesse (regras de baixa cobertura) - Prof. Flavio.

Regra	Interesse	
	Sistema	**Usuário**
1	0.67	*baixo*
5	0.25	baixo
6	0.33	*alto*
8	0.33	médio
9	0.25	*alto*
10	0.67	*baixo*
11	0.50	baixo
13	0.25	*alto*

Analisando a avaliação feita pelo Prof. Flávio sobre as regras de baixa cobertura observa-se (Tabela 13) uma discrepância (regras 1, 6, 9, 10 e 13) no interesse quando comparado aos valores do sistema. Isto provavelmente se deve ao fato das IG's terem sido fornecidas por usuário diferente do avaliador,

situação artificial uma vez que na prática o mesmo usuário deverá fornecer as IG's para logo em seguida obter as regras que contradizem suas hipóteses.

Tabela 14: Interesse (regras de cobertura alta) - Prof. Flavio.

Regra	Interesse	
	Sistema	Usuário
2	0.50	médio
3	0.33	*alto*
4	0.67	alto
7	0.67	alto
12	0.67	alto
14	0.67	alto
15	0.33	*alto*

Na avaliação feita pelo Prof. Flávio sobre as regras de cobertura alta os valores estão mais coerentes uma vez que estes resultados permitem uma avaliação mais intuitiva, mas ainda há algumas discrepâncias (regras 3 e 15) quando comparado aos valores do sistema, conforme mostra a Tabela 14.

Cabe ressaltar que, durante a avaliação das regras, o Prof. Flávio comentou que as regras deveriam conter algum atributo relacionado à grande área do pesquisador (exatas, humanas, etc.). Isso confirma a importância do atributo GRANDE_AREA, que foi usado no segundo experimento.

Veja no anexo "Entrevista 2 do Experimento 1" respostas de avaliação das regras com cobertura baixa (Formulário B1) e respostas de avaliação das regras com cobertura alta (Formulário B2) obtidas na entrevista com o Prof. Dr. Manoel Camillo Penna Neto.

Tabela 15: Interesse (regras de baixa cobertura) - Prof. Manoel.

Regra	Interesse	
	Sistema	Usuário
1	0.67	médio
5	0.25	médio
6	0.33	*alto*
8	0.33	baixo
9	0.25	*alto*
10	0.67	alto
11	0.50	médio
13	0.25	médio

A avaliação do Interesse feita pelo Prof. Manoel, resumida na Tabela 15, sobre as regras de baixa cobertura, bem como das regras de alta cobertura (Tabela 16), foi mais coerente quando comparado aos valores fornecidos pelo sistema, mas ainda apresentou alguma distorção (regras 3, 6, 7 e 9).

Tabela 16: Interesse (regras de cobertura alta) - Prof. Manoel.

	Interesse	
Regra	**Sistema**	**Usuário**
2	0.50	médio
3	0.33	*alto*
4	0.67	alto
7	0.67	*baixo*
12	0.67	médio
14	0.67	alto
15	0.33	médio

Comparando as respostas entre os dois entrevistados, a discrepância é relativamente pequena nas regras de alta cobertura, mas é bastante significativa nas regras de baixa cobertura.

7.6 Segundo Experimento

No segundo experimento um único usuário foi considerado e os metas são utilizados como possíveis previsores. Este foi um experimento mais completo incluindo o atributo GRANDE_AREA, além de utilizar o mesmo usuário na formulação das IG's e na avaliação das regras obtidas, que é de fato a maneira recomendada para uso do sistema na prática.

A inclusão do atributo GRANDE_AREA provocou a redução de 204 registros do primeiro experimento, que utilizou 5894 registros. Com a inclusão deste novo atributo obteve-se uma tabela com 24 atributos, sendo 8 categóricos previsores e 16 numéricos (10 previsores e 6 metas), contendo 5.690 registros completos sobre pesquisadores da região sul do Brasil.

Neste experimento as IG's foram obtidas através de uma entrevista pessoal com o Pró-Reitor de Pesquisas e Pós-Graduação da UEM, Dr. Gilberto Cezar Pavanelli. Durante esta entrevista descobriu-se que o atributo GRANDE_AREA seria fundamental para captar o conhecimento do entrevistado. Esta

constatação motivou a inclusão deste atributo na realização do segundo experimento.

Quadro 13: Impressões gerais (Prof. Pavanelli).

IG[1]: SE GRANDE_AREA = 5 ENTÃO ARTIGOS_PER_NAC = alto
IG[2]: SE GRANDE_AREA = 8 ENTÃO ED_PUBLIC_LIVRO_INT = baixo
IG[3]: SE GRANDE_AREA = 1 ENTÃO CAP_LIVROS_INT = alto
IG[4]: SE GRANDE_AREA = 2 ENTÃO ED_PUBLIC_LIVRO_INT = alto
IG[5]: SE GRANDE_AREA = 3 ENTÃO ED_PUBLIC_LIVRO_NAC = alto
IG[6]: SE GRANDE_AREA = 4 ENTÃO CAP_LIVROS_INT = baixo
IG[7]: SE GRANDE_AREA = 6 ENTÃO ED_PUBLIC_LIVRO_NAC = alto
IG[8]: SE GRANDE_AREA = 7 ENTÃO CAP_LIVROS_NAC = alto
IG[9]: SE GRANDE_AREA = 7, ARTIGOS_PER_INT = alto -->
ENTÃO ED_PUBLIC_LIVRO_NAC = alto
IG[10]: SE GRANDE_AREA = 5 ENTÃO ARTIGOS_PER_INT = baixo
IG[11]: SE LIDER_GRUPO = 1 ENTÃO ED_PUBLIC_LIVRO_INT = alto
IG[12]: SE LIDER_GRUPO = 0 ENTÃO ED_PUBLIC_LIVRO_NAC = baixo
IG[13]: SE IDADE = baixo ENTÃO CAP_LIVROS_NAC = baixo
IG[14]: SE IDADE = médio ENTÃO CAP_LIVROS_INT = alto
IG[15]: SE IDADE = alto ENTÃO ARTIGOS_PER_INT = médio
IG[16]: SE INIC_CIENT_ORIENT = baixo ENTÃO ARTIGOS_PER_NAC = baixo
IG[17]: SE INIC_CIENT_ORIENT = alto ENTÃO ARTIGOS_PER_NAC = alto
IG[18]: SE ARTIGOS_PER_NAC = baixo ENTÃO CAP_LIVROS_NAC = baixo
IG[19]: SE CAP_LIVROS_INT = alto ENTÃO ARTIGOS_PER_INT = alto
IG[20]: SE ED_PUBLIC_LIVRO_NAC = alto ENTÃO ARTIGOS_PER_NAC = baixo

A importância do atributo GRANDE_AREA ficou evidenciada após a obtenção das IG's fornecidas: 50% delas contém este atributo, conforme mostra a Quadro 13.

7.6.1 Acerto das Regras Descobertas no 2º Experimento

Conforme discutido na seção 7.5.1, utilizou-se validação cruzada para estimar o acerto em dados de teste separados do treinamento.

Assim como no primeiro experimento, o AGD foi executado com uma função de fitness considerando apenas o acerto (e não o interesse) da regra sendo avaliada, permitindo uma comparação direta da precisão preditiva do conhecimento extraído entre os dois algoritmos (J4.8 x AGD), conforme mostra a Tabela 17.

Fazendo-se uma comparação simples das taxas de acerto entre o J4.8 e o AGD neste experimento tem-se aproximadamente um empate técnico.

Tabela 17: Taxa de acerto na validação cruzada – 2º experimento.

Conseqüente			J4.8 Qualidade	AGD Apenas Qualidade	
Atrib. Meta	**Classe**	**Freq. Relat. (%)**	**Acerto (%)**	ACERTO (ā% ±DP)	**Cobertura (ā ±DP)**
ARTIGOS_PER_NAC	baixo	46.9	**64.9**	58.8 ±14.0	1.6 ±0.4
	médio	50.6	**63.9**	60.4 ±13.3	1.4 ±0.3
	alto	2.5	**9.1**	0.0 ±0.0	0.2 ±0.1
ARTIGOS_PER_INT	baixo	64.2	76.6	**90.7 ±5.4**	4.1 ±1.2
	médio	29.7	**45.3**	40.0 ±16.3	0.3 ±0.1
	alto	6.1	**32.2**	25.0 ±13.4	0.5 ±0.2
CAP_LIVROS_NAC	baixo	76.9	82.2	**95.2 ±3.0**	4.4 ±1.3
	médio	21.2	45.3	**56.7 ±15.8**	0.9 ±0.3
	alto	1.9	**27.4**	25.0 ±13.4	0.3 ±0.2
CAP_LIVROS_INT	baixo	93.2	93.4	**98.4 ±0.7**	24.0 ±2.9
	alto	6.8	**51.7**	14.3 ±10.4	0.3 ±0.2
ED_PUBLIC_LIVRO_NAC	baixo	83.5	86.0	**89.5 ±9.9**	7.2 ±3.1
	alto	16.5	54.7	**56.9 ±14.0**	0.6 ±0.1
ED_PUBLIC_LIVRO_INT	baixo	97.9	97.9	**98.9 ±0.5**	38.3 ±4.5
	alto	2.1	0.0	0.0 ±0.0	0.2 ±0.1

Observa-se que o AGD ganha do J4.8 sempre que a freqüência da classe é alta, enquanto que o J4.8 forneceu melhores resultados nas classes de baixa freqüência. No entanto, assim como no primeiro experimento, em casos críticos (*e.g.:* atributo ED_PUBLIC_LIVRO_INT) o J4.8 não consegue classificar adequadamente, limitando-se a prever a classe da maioria para todos os registros.

7.6.2 Regras Descobertas no 2º Experimento

A fim de descobrir as regras finais a serem mostradas ao usuário, o programa AGD foi executado com todos os dados, incluindo o atributo GRANDE_AREA e utilizando os metas como previsores, perfazendo um total de 15 chamadas à rotina do AG (uma para cada meta/valor) contendo uma população de 100 indivíduos que evoluem 60 gerações em cada chamada. O tempo total de processamento foi de aproximadamente 6 minutos e 40 segundos.

No ANEXO B são apresentadas as regras extraídas pelo AGD no segundo experimento acompanhadas do grau de interesse medido pelo sistema, número

da IG fornecida pelo usuário, freqüência relativa da classe no conjunto total de dados e a cobertura no conjunto total de dados.

Os resultados obtidos podem ser divididos em dois grupos: regras com cobertura muito baixa (pequenos disjuntos) e regras com boa cobertura. Os dois grupos de regras foram mostrados para o usuário em dois formulários, o qual respondeu conforme mostra o ANEXO C – Entrevistas para Avaliação das Regras.

As regras 1, 3, 6, 7, 9, 11 e 15 (FORMULÁRIO C1) foram consideradas "pequenos disjuntos" por apresentarem cobertura inferior a 0.1% do conjunto total dos dados. As regras 2, 4, 5, 8, 10, 12, 13 e 14 (FORMULÁRIO C2) possuem cobertura superior a 0.1% dos dados e foram agrupadas separadamente. As principais regras descobertas neste experimento estão resumidas no Quadro 14 e no Quadro 15.

Quadro 14: Regras com cobertura baixa.

REGRA 1:
SE GRANDE_AREA = Ciências Agrárias
E NIVEL_FORMACAO = baixo (Graduação ou Especialização)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

REGRA 7:
SE GRANDE_AREA = Ciências Humanas
E NIVEL_FORMACAO = baixo (Graduação ou Especialização)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = baixo (0)

Quadro 15: Regras com boa cobertura.

REGRA 4:
SE IDADE = alto (>57)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = baixo (0)

REGRA 12:
SE GRANDE_AREA = Engenharias
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

REGRA 14:
SE GRANDE_AREA = Ciências Biológicas
ENTÃO Nº_DE_LIVROS_INTERNAC_EDITADOS/PUBLICADOS = baixo (0)

Os dois formulários (C1 e C2) foram apresentados ao Pró-Reitor, Dr. Pavanelli, em uma entrevista onde, inicialmente, mostraram-se as suas IG's (que ele mesmo propôs em entrevista realizada um mês antes) para que ele pudesse se recordar do que havia declarado como hipóteses. Isto foi feito com o objetivo de simular uma sensação real de interação com o sistema onde o usuário primeiro fornece o seu conhecimento para o AGD para, logo em seguida, obter conhecimento novo em contradição as suas IG's.

Em seguida mostrou-se cada regra nova acompanhada da IG que foi contrariada pela regra. O objetivo deste procedimento foi provocar a sensação de surpresa, que o sistema deverá proporcionar, ao negar o que o usuário afirmou ser óbvio e conseqüentemente valorizar o conhecimento novo e surpreendente obtido.

7.6.3 Interesse nas Regras Descobertas no 2º Experimento

A primeira constatação, de caráter geral, foi que o entrevistado se interessou mais por regras com alta cobertura do que por regras que representavam exceções (outliers).

Outra constatação geral foi que o entrevistado demonstrava dificuldade em entender regras que continham muitas condições no antecedente e reduzia o interesse na mesma à medida em que o tamanho da regra aumenta, por se distanciar muito da IG (as IG's em geral são bem curtas).

Durante a apresentação das regras do Formulário C1 (regras de baixa cobertura) verificou-se o seguinte: a análise feita sobre regras de baixa cobertura é diferente da análise feita no caso de regras com alta cobertura.

Quando uma regra surpreende, a baixa cobertura apenas confirma que esta regra representa uma simples exceção, e isto não é relevante. Quando uma regra representa algo evidente (ou óbvio) que se espera da maioria, a baixa cobertura causa interesse por indicar que apenas alguns (menos de seis) pesquisadores correspondem às características discriminadas na regra, enquanto que o esperado era que um número significativo dos mesmos correspondessem àquelas características, causando impacto e, portanto, surpresa (alto interesse).

Tabela 18: Interesse (regras de baixa cobertura) – Prof. Pavanelli.

Regra	Interesse	
	Sistema	Usuário
1	0.50	alto
3	0.33	médio
6	0.20	baixo
7	0.50	alto
9	0.25	baixo
11	0.33	médio
15	0.33	baixo

Para avaliar o quanto os valores calculados para o grau de interesse se aproximam da realidade, apresenta-se, na Tabela 18, uma comparação entre o interesse calculado pelo sistema e o interesse real do usuário por cada regra com baixa cobertura. Esta tabela demonstra maior coerência na avaliação, comparado com os valores da Tabela 13 e Tabela 15, confirmando a hipótese, levantada no primeiro experimento, de que quando o mesmo usuário formula as IG's e avalia as regras descobertas a tendência é haver maior aproximação entre o grau de interesse calculado pelo sistema e a avaliação do usuário.

Veja na Tabela 19 a mesma comparação, considerando as regras com cobertura alta.

Tabela 19: Interesse (alta cobertura) – Prof. Pavanelli.

Regra	Interesse	
	Sistema	Usuário
2	0.50	médio
4	0.33	*alto*
5	0.20	médio
8	0.50	baixo
10	0.25	médio
12	0.33	*alto*
13	0.33	baixo
14	1.0	alto

7.7 Discussão Comparativa entre os Experimentos

Considerando os resultados da validação cruzada apresentados nas seções 7.5.1 e 7.6.1, conclui-se que o AGD fornece taxas de acerto nos dados de teste

equivalentes ao J4.8. No primeiro experimento o AGD apresentou taxas de acerto levemente melhores em várias classes, enquanto que no segundo experimento o J4.8 apresentou taxas de acerto um pouco superiores em algumas classes.

Entretanto, o J4.8 não foi projetado para extrair conhecimento relevante enquanto que o AGD foi concebido com a meta específica de extrair apenas conhecimento relevante, além de correto (foi verificado empiricamente que a inclusão do interesse no cálculo da função de Fitness não muda de forma significativa a taxa de acerto do AGD).

Na maioria dos atributos metas, o J4.8 forneceu uma árvore imensa, dificultando a compreensibilidade do conhecimento, enquanto que o AGD fornece apenas o conhecimento estimado como sendo relevante para o usuário, contendo apenas uma regra para cada meta/valor, que neste estudo de caso corresponde a 15 classes.

Comparando os resultados da validação cruzada do segundo experimento com os resultados do primeiro experimento, verifica-se que houve uma melhora na taxa de acerto em quase todas as classes (com exceção de uma) no segundo experimento, tanto no AG como no J4.8. Este ganho, nos resultados do segundo experimento, se deve à utilização dos atributos metas como previsores e principalmente à presença do atributo GRANDE_AREA no segundo experimento.

A utilização dos metas como possíveis previsores ajuda a caracterizar melhor o experimento como modelagem de dependência onde isto pode ocorrer.

A utilização dos metas como previsores melhorou a taxa de acerto em geral devido provavelmente à grande correlação naturalmente existente entre os atributos de produção científica representados principalmente pelos atributos escolhidos como metas. Por outro lado, a inclusão do atributo GRANDE_AREA trouxe benefício significativo à taxa de acerto devido à grande disparidade no perfil dos pesquisadores de diferentes áreas, facilitando a identificação de nichos de instâncias que cobrem perfis específicos de algumas áreas.

Comparando o interesse nas regras obtidas pelo AGD entre os dois experimentos conclui-se que ambos fornecem regras de interesse (veja Tabela 20) confirmando a missão do AGD em encontrar regras relevantes.

As regras obtidas no primeiro experimento apresentaram grau de interesse levemente superior, provavelmente devido à característica das IG's, fornecidas por um pesquisador que naturalmente deve possuir menos visão de relacionamentos entre os atributos quando comparado com um Pró-Reitor de Pesquisa (segundo experimento). É mais fácil descobrir conhecimento que contradiz IG mais simples (talvez até errada) do que contradizer IG que represente conhecimento mais correto no contexto da aplicação.

Tabela 20: Comparando interesse e cobertura.

Conseqüente		Metas NÃO Previsores		Metas Previsores	
Atrib. Meta	**Classe**	**Interesse**	**Cobertura**	Interesse	**Cobertura**
ARTIGOS_PER_NAC	baixo	0.67	2.1	0.67	**2.5**
	médio	0.50	**49.0**	0.50	26.0
	alto	0.33	**20.3**	0.33	1.5
ARTIGOS_PER_INT	baixo	**0.67**	38.5	0.50	**550.4**
	médio	0.25	5.5	**0.50**	**92.0**
	alto	**0.33**	4.5	0.25	**5.0**
CAP_LIVROS_NAC	baixo	**0.67**	15.7	0.50	**344.5**
	médio	**0.33**	4.2	0.25	**60.0**
	alto	**0.25**	1.4	0.00	**2.5**
CAP_LIVROS_INT	baixo	**0.67**	2.2	0.50	**3092.3**
	alto	**0.50**	1.0	0.33	1.0
ED_PUBLIC_LIVRO_NAC	baixo	0.67	126.1	**1.00**	**735.0**
	alto	0.25	2.4	**0.50**	**53.0**
ED_PUBLIC_LIVRO_INT	baixo	**0.67**	36.9	0.50	**325.0**
	alto	0.33	**17.4**	**0.50**	1.0

Os resultados de compreensibilidade sintática apresentados na seção 7.5.2, correspondentes ao primeiro experimento, foram conclusivos razão pela qual no segundo experimento não foi apresentada qualquer avaliação da comprensibilidade sintática das regras obtidas.

Comparando a cobertura entre os dois experimentos, observa-se uma grande vantagem no segundo experimento onde a cobertura em geral é superior (veja Tabela 20, cujos resultados se referem a execuções do AGD sobre o conjunto total de dados). Este fato se deve provavelmente à

característica das IG's do segundo experimento, fornecidas por um Pró-Reitor de Pesquisa que possui maior visão do relacionamento entre os diversos atributos da aplicação, ocasionando a inclusão do atributo GRANDE_AREA que permite fazer distinção entre as características dos pesquisadores de diferentes áreas. Por exemplo, os pesquisadores da área de exatas normalmente publicam mais no exterior do que os pesquisadores da área de letras.

Portanto, o principal mérito, na distinção entre os dois experimentos, fica com as IG's, que proporcionaram as principais mudanças nos resultados alcançados, demonstrando a importância da interação do usuário com o AGD.

7.8 Considerações Finais

Neste capítulo apresentaram-se resultados de implementação de um algoritmo de mineração de dados (denominado AGD), baseado em algoritmos genéticos, capaz de descobrir conhecimento correto (válido), compreensível e surpreendente e que pode auxiliar à tomada de decisões no domínio da gestão e planejamento do fomento da ciência e tecnologia da região sul do Brasil.

As taxas de acerto obtidas pelo J4.8 nos dois experimentos (Tabela 11 e Tabela 17) demonstraram a importância da escolha dos atributos, tanto da sua natureza quanto da sua utilização como meta ou previsor. A inserção do atributo GRANDE_AREA e a utilização dos atributos metas como previsores fizeram aumentar a taxa de acerto do J4.8 no segundo experimento em mais de 13% em média.

A análise da compreensibilidade sintática das regras, apresentada no primeiro experimento, permite concluir que o J4.8 gera conhecimento complicado para o usuário devido ao grande número de regras e ao grande número de condições em algumas regras, enquanto que o AGD fornece poucas regras de fácil compreensão com no máximo cinco condições em cada regra. Mesmo assim, apesar das regras obtidas pelo AGD serem bastante intuitivas, inicialmente o usuário apresentou dificuldades em entender algumas regras, principalmente aquelas que tinham mais de três condições no antecedente.

Existe uma relação direta entre a simplicidade da regra (poucas condições) e o grau de interesse quando o critério utilizado para medir o interesse é a contradição entre os conseqüentes de uma regra e de uma IG. Se as IG's fornecidas pelo usuário possuírem poucas condições, as regras com muitas condições terão baixo valor de similaridade do antecedente e, portanto, um baixo grau de interesse.

Os resultados apresentados na Tabela 20 demonstram a influência das impressões gerais no processo de busca das melhores regras. Quanto melhor é o conhecimento fornecido (coerência na especificação) para o AGD melhores serão as regras obtidas pelo sistema (maior cobertura). Nos dois experimentos 40% das IG's não aparecem como contradição das regras finais, mas nada pode ser afirmado sobre a contribuição que elas deram ao processo evolutivo.

As entrevistas e a conseqüente interação com o usuário do segundo experimento conduziram à especificação de uma metodologia para se chegar à avaliação subjetiva dos resultados obtidos. Esta metodologia pode ser resumida no seguinte:

- escolher um usuário potencial que tenha conhecimento do domínio da aplicação e que tenha interesse no futuro sistema a ser implementado;
- selecionar os atributos candidatos juntamente com este usuário;
- após o projeto e implementação do protótipo, preparar um questionário que induza respostas capazes de gerar as IG's;
- marcar uma entrevista informal e apresentar este questionário na forma de diálogo procurando dirimir possíveis dúvidas do usuário quanto ao significado das perguntas e à modalidade das respostas;
- transformar as respostas do questionário em regras caracterizando as impressões gerais do usuário;
- marcar outra entrevista, apresentar as IG's ao usuário para confirmação e possíveis ajustes;
- executar o protótipo e extrair regras com o apoio das impressões gerais;
- marcar outra entrevista, mostrar primeiro as IG's para que o usuário se lembre das suas hipóteses. Em seguida, mostrar cada regra descoberta

acompanhada da IG contrariada para que o usuário classifique o grau de interesse na nova regra descoberta;

- Comparar a avaliação do usuário com o grau de interesse calculado;
- Com base nos resultados da comparação, efetuar ajustes no protótipo e repetir esta metodologia.

Na última entrevista é importante explicar ao usuário o significado de grau de interesse esclarecendo que as regras descobertas estão respaldadas por dados reais do banco de dados e que o esperado é que elas venham a contradizer as suas hipóteses.

As conclusões gerais obtidas com esta tese, bem como sugestões de pesquisas futuras, encontram-se no próximo capítulo.

8 CONCLUSÕES E SUGESTÕES

Esta tese possui dois enfoques principais: um em pesquisa básica e outro em pesquisa aplicada. A pesquisa básica se concentra no desenvolvimento de uma técnica de mineração de dados capaz de extrair especificamente conhecimento relevante. A aplicação está situada no âmbito da gestão de C&T tendo como estudo de caso a região sul do Brasil. Em suma, desenvolveu-se um protótipo contendo uma técnica capaz de extrair conhecimento novo e relevante com resultados de aplicação na área de C&T.

Neste capítulo são apresentadas as principais conclusões a respeito das técnicas de MD investigadas e também a respeito dos resultados experimentais obtidos utilizando dados sobre pesquisadores cadastrados no CNPq.

Na seção 8.1 são apresentadas as conclusões sobre a validade do AGD, tendo como referencial o algoritmo J4.8, destacando os resultados alcançados nos dois experimentos relatados no Capítulo 7, em especial na busca por conhecimento relevante no contexto de C&T, baseados em uma abordagem subjetiva. Além disso, apresenta conclusões sobre a combinação híbrida de AG com a lógica difusa, a compreensibilidade do conhecimento obtido e a confirmação da hipótese: é possível extrair conhecimento novo e surpreendente ao planejamento e gestão de C&T a partir dos bancos de dados já existentes e disponíveis.

Na seção 8.2 apresentam-se sugestões para trabalhos futuros que darão continuidade a esta pesquisa, tanto na descoberta de conhecimento novo no contexto da aplicação, como na construção e aprimoramento de técnicas e métodos desenvolvidos pela comunidade científica da área de mineração de dados.

8.1 Conclusões

O Diretório dos grupos de pesquisa no Brasil, descrito na seção 2.6.1, tem sido bastante explorado contando, inclusive, com um aplicativo, disponível para uso através da Internet (www.cnpq.br), que fornece diversas informações sobre

grupos de pesquisa, permitindo a montagem de tabelas com atributos escolhidos pelo usuário. Mesmo assim, o AGD conseguiu extrair conhecimento oculto e relevante baseado em algumas hipóteses (impressões gerais) de dois pesquisadores, sendo um deles um pró-reitor de pesquisa, um dos tipos de usuário para o qual o sistema foi desenvolvido. Há a possibilidade de mais conhecimento ser descoberto dependendo de novas IG's que podem ser elaboradas por outros usuários ou pelos mesmos. Uma pessoa pode estar interessada em coisas diferentes em momentos diferentes favorecendo a criação de novas IG's.

No contexto da aplicação, o algoritmo proposto mantem a confidencialidade das informações. Devido a própria natureza do processo de KDD, que procura conhecimento oculto no universo da informação e não em seus registros individuais, ele garante a execução do algoritmo sem violar o princípio da privacidade.

Apesar das comparações entre o algoritmo AGD e o J4.8, apresentados no Capítulo 7, não foi o foco desta tese mostrar qual algoritmo descobre regras melhores. Comparações entre diversos algoritmos já são tratadas largamente na literatura, conforme descrito na seção 4.2.2.

As comparações foram necessárias para avaliar a eficácia do AGD comparado com o acerto obtido com o J4.8, que é um algoritmo bastante conhecido pela comunidade de mineração de dados, servindo como referencial.

Os acertos obtidos com o AGD foram aproximadamente equivalentes, na média, aos acertos obtidos com o J4.8 na aplicação considerada, levando-se em consideração as 15 classes a serem previstas. O AGD descobriu regras com maior acerto em algumas classes, enquanto o J4.8 descobriu regras com maior acerto em outras classes. Assim, o foco das conclusões desta tese são a compreensibilidade e interesse do conhecimento descoberto.

A principal contribuição desta tese foi o desenvolvimento de um algoritmo voltado especificamente para descoberta de conhecimento RELEVANTE na área de C&T, e isto é inédito além de relevante. Como produto, implementou-se um protótipo, permitindo a avaliação do futuro sistema de computador

denominado AGD, que interage com o usuário utilizando o seu conhecimento prévio para gerar conhecimento novo, permitindo ao mesmo rever seus conhecimentos e ainda possibilitando a confirmação ou refutação de suas hipóteses. Não se tem informação de sistema semelhante no contexto de C&T, sendo, portanto, inédito.

Em um caso extremo, onde todo conhecimento fornecido pelo usuário for absolutamente correto e não houver qualquer conhecimento oculto ou exceção no banco de dados relacionado com algum dos atributos que aparecem nas IG's, o AGD não terá o que descobrir e não fornecerá qualquer regra de interesse. No outro extremo, se todo conhecimento fornecido pelo usuário for errado ou contraditório com o conhecimento oculto no banco de dados, então o AGD deverá fornecer regras com alto grau de interesse. Entretanto, dificilmente algum usuário se situará em algum destes extremos com todas as suas IG's, permitindo a conclusão de que o AGD, em geral, deverá ser útil para a maioria de seus usuários potenciais.

Na abordagem utilizada para busca de regras de interesse o usuário de antemão fornece o que sabe (impressões gerais) e, conforme demonstrado nos itens 4.5.3, a descoberta de regras que contradizem o conhecimento anterior do usuário é considerada de interesse. Nesta abordagem, a contribuição desta tese foi adaptar uma forma de medir qualidade de conhecimento, descrita na seção 6.9.4, que mostra o quanto (grau de interesse) a regra contraria as hipóteses do usuário.

Os resultados obtidos confirmaram a validade da adaptação destas medidas de interesse, gerando regras que contrariam (em determinado grau) as impressões gerais do usuário. Um exemplo claro é a regra 12, descoberta no segundo experimento, que contraria em 100% a IG 5 onde um usuário acreditava que os pesquisadores da área de engenharias tinham alto índice de edição/publicação de livros nacionais enquanto que na realidade o AGD descobriu o contrário, ou seja, os pesquisadores desta área possuem baixa produção neste indicador.

A avaliação das regras descobertas feita por três usuários potenciais, dois de uma universidade particular (um Pró-Reitor de Pesquisa e o outro

Coordenador de Pesquisa) e um de uma universidaide pública (Pró-Reitor de Pesquisa), comprovaram a tese de que o AGD descobre regras de interesse. De todas as avaliações obtidas, 45% das regras foram classificadas como sendo de alto interesse, 31% foram classificadas como sendo de médio interesse e apenas 24% foram consideradas de baixo interesse. Em várias regras descobertas pelo AGD o grau de interesse calculado em geral está compatível com a avaliação subjetiva dos usuários, principalmente quando é o mesmo que fornece as IG's.

Os resultados sobre compreensibilidade sintática apresentados na seção 7.5.2 foram importantes para demostrar que o AGD sempre descobre conhecimento simples (poucas regras, com poucas condições por regra), enquanto que em vários casos o J4.8 descobre centenas de regras embutidas em árvores de decisão de difícil interpretação por parte do usuário. O algoritmo J4.8, na aplicação considerada, gerou mais de 50 páginas de resultados.

Outra contribuição da técnica proposta está em descobrir regras de interesse de forma direta, percorrendo caminhos de busca que ignoram o óbvio e direcionando a busca para descobrir apenas regras relevantes.

De acordo com King *et al.* (1995), um problema grave, quando se usa AG's, é o ajuste de parâmetros. No entanto, a experiência obtida durante a implementação e testes do AGD não confirmaram esta afirmativa, fato demonstrado pela generalidade do ajuste de parâmetros para as quinze classes diferentes definidas nesta aplicação.

Apesar do protótipo implementado resolver a tarefa de modelagem de dependência através da abordagem de classificação, onde vários atributos metas são considerados durante sua execução exigindo a repetição da rotina de evolução genética, o AGD não exigiu muito tempo de processamento nesta aplicação (cerca de 1 minuto para cada 1000 registros).

A utilização da lógica difusa simplificou a representação do conhecimento através de uma generalização de dados numéricos em termos lingüísticos, sem perder o sentido da informação. Esta abordagem se resumiu na utilização de uma linguagem difusa que é mais natural no âmbito do usuário planejador que não deve se preocupar com detalhes (intervalos numéricos específicos) e sim

com tendências e previsões abrangentes (alto/baixo, pouco/muito), e isto parece inédito neste contexto de aplicação.

A abordagem por modelagem de dependência se mostrou viável, apesar do maior custo computacional, compensada pela sua maior abrangência, dado que permite ao usuário obter conhecimento mais diversificado envolvendo a previsão de diversos atributos metas.

A confirmação do interesse nesta pesquisa, bem como nos resultados obtidos, está na decisão em se utilizar estes estudos e resultados como base para o desenvolvimento de um projeto maior, denominado InterSul. Este é um projeto inter-institucional e multidisciplinar, custeado pelo CNPq através do programa de fomento denominado Plano Sul, que inclui a participação de pesquisadores e bolsistas de três Universidades: UEM, PUC-PR e UFSC (Universidade Federal de Santa Catarina).

Esta tese apresenta uma contribuição não só para este projeto (InterSul), mas também para os Pró-Reitores de Pesquisa das Universidades, para as agências de fomento a C&T da região sul e para a comunidade científica da área de MD que estuda a combinação híbrida de técnicas de IA para aproveitar as melhores características de cada uma delas.

A existência de um sistema como o AGD abre espaço para uma verdadeira mineração dos dados sobre C&T já existentes e ainda é favorecido pela chegada de novos dados periodicamente, com a formação de novas versões do Diretório de Grupos de Pesquisa que em breve terá seus dados atualizados continuamente. Além disso, o CNPq já conta com os currículos atualizados diariamente e está integrando os dados da CAPES também. Com todo este volume de dados que está se aglutinando, o sistema AGD é uma alternativa plausível de sair do protótipo para ser implementada e aprimorada como ferramente de auxílio ao planejamento e gestão de C&T.

8.2 Trabalhos Futuros

Para o protótipo implementado foram selecionados vários atributos, alguns de descrição do pesquisador e outros indicando sua produção, e todos foram

utilizados nas execuções do programa. Em experimentos futuros deve-se oportunizar o usuário a eliminar atributos que ele considerar irrelevantes, bastando para isso desativar estes atributos no banco de dados. Outra possibilidade, que exige alguma alteração no programa, é incluir outros atributos que o usuário julgar relevante mas que ainda não foram usados pelo AGD. As duas abordagens deverão existir no sistema final a ser implementado no projeto InterSul.

A versão atual do AGD retorna ao usuário apenas a melhor regra para cada classe. No futuro deve-se avaliar a possibilidade de selecionar mais de uma regra para cada classe da população final para ser mostrada para o usuário.

Constatou-se que quando uma regra descoberta tem mais condições no antecedente do que o número de condições existentes na IG o entrevistado reduz de forma significativa o interesse na regra. Apesar deste fato estar sendo levado em consideração no cálculo do grau de interesse, talvez seja necessário no futuro reavaliar este cálculo no sentido de diminuir o grau de interesse de forma mais acentuada neste caso.

A dificuldade do usuário compreender regras com muitas condições, verificada nas entrevistas do segundo experimento, sugere a possibilidade de se restringir o número máximo de condições em cada regra para três, mas isto deverá ser confirmado empiricamente no futuro.

Em experimentos futuros, uma alternativa é exigir que a regra cubra um número mínimo de registros (ex.: cinco). Se a cobertura mínima não é satisfeita pode-se podar a regra, removendo-se alguma(s) de suas condições, garantindo maior generalidade da regra.

Os resultados preliminares tentando descobrir regras prevendo todos os metas ao mesmo tempo no processo evolutivo demonstraram ser piores do que utilizando meta fixo. No entanto, uma investigacao mais detalhada dos prós e contras dessas duas abordagens deverá ser retomada no futuro.

Partindo da hipótese de que quanto melhor o conhecimento fornecido para o sistema melhor será o conhecimento obtido e, considerando a influência das

IG's, relatadas na seção 7.7, sugere-se realizar futuramente experimentos onde o conhecimento obtido pelo AGD seja realimentado como IG's.

Há três casos possíveis para se avaliar o interesse na regra, conforme descrito na seção 4.5.3. Nesta tese foi considerado apenas o caso do conseqüente inesperado (seção 6.9.4) ficando os demais casos para pesquisas futuras. Uma implementação trivial é o caso da confirmação de hipóteses, que exige poucas alterações no AGD, aproveitando a rotina para cálculo do grau de similaridade já existente.

Como o AGD foi treinado com uma função de fitness contendo dois objetivos (acerto e interesse), em pesquisas futuras pretende-se utilizar AG's multi-objetivos trabalhando com dominância de Paretto, técnica capaz de encontrar a melhor combinação entre vários objetivos incomensuráveis e conflitantes. Um exemplo está na Figura 50, onde se deseja maximinar dois objetivos: acerto e interesse.

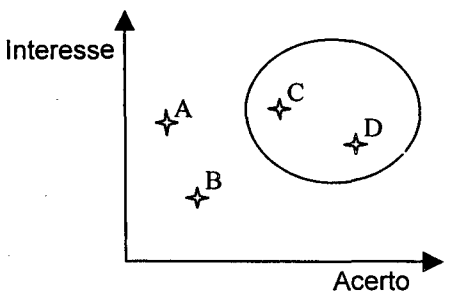


Figura 50: Dominância de Paretto

No exemplo da Figura 50, há uma competição entre os dois indivíduos (C e D) em destaque dentro do círculo, onde um possui acerto maior e outro possui interesse maior, situação de conflito tratada pela dominância de Paretto. Neste caso, nenhum dos dois pode ser considerado melhor que o outro, e ambos são ditos fazer parte do "conjunto de Paretto", contendo soluções não dominadas. Porém, tanto C quanto D são claramente melhores que os indivíduos A e B, já que C e D representam soluções melhores tanto no objetivo de maximizar interesse quanto no objetivo de maximizar acerto. Assim, diz-se que A e B são dominados por C e D.

Experimentos paralelos demostraram que algumas regras de interesse um pouco menor (segunda maior fitness) apresentavam acerto maior do que a

regra de maior fitness. Em pesquisas futuras deve-se empregar a técnica de AG's multi-objetivos para tentar contornar este problema e estudar formas de aproveitar estas regras também.

Na seqüência desta pesquisa, deve-se continuar a interação com os usuários entrevistados e ampliar a avaliação dos conhecimentos obtidos em termos da utilização dos mesmos como apoio à gestão de C&T. Poder-se-ia, assim, alcançar o topo da pirâmide mostrada na Figura 1 do Capítulo 1, transformando efetivamente o conhecimento descoberto em uma decisão inteligente.

O AGD não tem exclusividade em dados cientométricos e sim possui ênfase na participação direta do usuário no processo de extração de conhecimento. Assim, o método apresentado nesta tese é aplicável em qualquer domínio de conhecimento em que seja possível a aplicação de algoritmos de extração do conhecimento e em que a participação do usuário no processo é factível. Entre esses, pode-se citar os seguintes campos:

a) **Extração de conhecimento em CRM**. Analistas de mercado das organizações ou consultores de mercado (usuários) podem fornecer pistas sobre o comportamento do consumidor, o que viabiliza a aplicação do AGD em uma base de dados sobre as operações empresa-clientes, a fim de confirmar (ou negar) padrões de comportamento diante de novos produtos, promoções, ajustes de preços, etc.

b) **Extração de conhecimento em dados financeiros**. Analistas de mercado de ações possuem diversas impressões gerais sobre o comportamento do mercado. Algumas chegam a ser dogmas da profissão. O AGD pode, diante de uma massa de dados sobre o movimento de ações, procurar a confirmação ou, principalmente, identificar momentos em que as crenças dos analistas foram negadas pelo comportamento de mercado. As análises específicas, antes de representarem *outliers* de informação, podem indicar situações do mercado em que o comportamento usual deixou de ocorrer.

c) **Extração de conhecimento em bases de balanços e movimentos financeiros de empresas**. Uma das práticas dos consultores financeiros é

a análise de índices financeiros específicos de empresas (financial ratios). Há uma faixa de comportamento para cada índice, de acordo com as características da empresa e do mercado em que atuam. Desvios de valores podem indicar problemas. Com uma base para formação de índices e as impressões gerais de um analista de empresas, poder-ser-ia aplicar o AGD e, uma vez mais, indicar empresas que apresentam comportamento satisfatório, mesmo com indicadores que conflitam com o esperado ou ainda empresas que estão com problemas financeiros, embora mantenham os indicadores em faixas esperadas.

Esta lista poderia continuar, já que em todo o domínio de KDD em que há um usuário capaz de fornecer impressões gerais há condição de aplicação do método apresentado nesta tese.

9 FONTES BIBLIOGRÁFICAS

AGRAWAL, R. & SHAFER, J. C. Parallel mining of association rules. ***IEEE Transactions on Knowledge and Data Engineering,*** vol. 8, NO. 6, 1996.

AGRAWAL, R. & SRIKANT, R. Fast algorithms for mining association rules. ***Proc. of the 20th Int'l Conference on Very Large Databases***. Santiago, Chile, 1994.

-------. Mining generalized association rules in large relational tables. ***Proc. of 21st Int'l Conference on Very Large Databases***. Zurique, Suíça, 1995.

AGRAWAL, R. *et al.* Mining association rules between sets of items in large databases. ***Proc. of the ACM SIGMOD Conference***. Washington, DC, USA, p. 207-216, 1993.

ANDERSON, P. ART2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns. ***Applied Optics***, 151-162, 1990.

APTE, C.; WEISS, W. & GROUT, G. Predicting defects in disk drive manufacturing: a case study in high-dimensional classification. **Proc. IEEE 9th Conf. Artif. Intel. For Applications**, 212-218. Orlando, Florida, 1993.

BACK, T.; FOGEL, D. B. & MICHALEWICZ, Z. **Evolutionary Computation 1**: Basic Algorithms and Operators. Chapter 24: Blickle T. Tournament selection. Institute of Phisics Publishing, 2000.

BALA, J.; HUANG, J.; VAFAIE, H.; DEJONG, K. & WECHSLER, H. Hibrid learning using genetic algorithms and decision trees for pattern classification. **Proc. 14th Int. Joint Conf. AI (IJCAI-95),** 719-724, 1995.

BARANAUSKAS, José Augusto Extação Automática de Conhecimento por Múltiplos Indutores. USP – Instituto de Ciências Matemáticas e de Computação, São Carlos – SP. **Tese de Doutorado**, Cap. 3: Metodologia de Avaliação de Algoritmos, 2001.

BECHER, J.D.; BERKHIN, P. & FREEMAN, E. Automating Exploratory Data Analysis for Efficient Data Mining. **KDD 2000**, Boston MA USA, 424-429, 2000.

BOJARCZUK, C.C.; LOPES, H.S. & FREITAS, A.A. Genetic programming for knowledge discovery in chest pain diagnosis. **IEEE Engineering in Medicine and Biology Magazine**, 19(4) – special issue on data mining and knowledge discovery, 38-44, 2000.

BRACKETT, M. H. ***The data warehouse challenge: taming data chaos***. New York: John Wiley & Sons, 1996.

BRISOLLA, S. N. Indicadores para apoio à tomada de decisão. ***Ciência da Informação,*** Brasília, v. 27, n. 2, p. 221-225, maio/ago. 1998.

CAPES. Brasília. http://www.capes.gov.br/distribuição_de_arquivos/datacapes, 1999.

CARVALHO, D.R. & FREITAS, A.A. A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining. To appear in: ***Proc. Genetic and Evolutionary Computation (GECCO-2000),*** Las Vegas, NV, USA. July 2000.

CHEN, M-S., HAN, J. E YU, P. S. Data mining: an overview from a database perspective. ***IEEE Transactions on Knowledge and Data Engineering,*** v. 8, n. 6, p.886-883, 1996.

CHERKAUER, K.J. & SHAVLIK, J.W. Growing simpler decision trees to facilitate knowledge discovery. **Proc. 3rd Int. Conf. Knowledge Discovery & Data Mining**, 315-318.AAAI Press, 1997.

CHEUNG, D. W., NG , V. T. & FU, A. W. Efficient mining of association rules in distributed databases. ***IEEE Transactions on Knowledge and Data Engineering***, v.8, n. 6, p. 911-922, 1996.

CNPq. ***Construindo o futuro: propostas e realizações da gestão 95-98.*** Brasília: CNPq, 1998.

------. ***Diretório dos Grupos de Pesquisa no Brasil.*** **Versão 3.0**. Disponível na Internet. http//:www.cnpq.br/gpesq3. 31 mar.1999a.

------. ***Plataforma Lattes***. Brasília: 1999a. http://www.cnpq.br/lattes/index.htm. 1999b.

------. **Resenha Estatística do CNPq**. Brasilia: CNPq, 2001.

COUTINHO, L. G., FERRAZ, J. C. ***Estudo da Competitividade da Indústria Brasileira.*** Brasília: MCT, Campinas: Papirus, 1994.

COVER, T. M. & THOMAS, J. A. **Elements of Information Theory**. John Wiley & Sons, 1991.

DE JONG, K.A. et al. Using Genetic Algorithms for concept Learning. ***Machine Learning,*** p. 161-188, 1993.

DEB, Kalyanmoy et al. Learning to Avoid Moving Obstacles Optimally for Mobile Robots Using a Genetic-Fuzzy. In: Eiben, A. E. et al. (Eds.) ***Parallel Problem Solving from Nature (PPSN-V)***. Lecture Notes in Comp. Science 1498, pp. 583-592. Springer-Verlg, 1998.

DELGADO, Myriam et al. Modular and Hierarchical Evolutionary Design of Fuzzy Systems. In: Banzhaf, W. et al. (Eds.) ***Proc. Genetic and Evolutionary Computation Conference (GECCO'99)***, Vol. 1, Morgan Kaufmann Publishers, pp. 180-187, 1999.

DHAR, V.; CHOW, D. & PROVOST, F. Discovering Interesting Patterns for Investment Decision Making with GLOWER – A Genetic Learner Overlaid With Entropy Reduction. To appear in: ***Journal of Data Mining and Knowledge Discovery***, 2000.

DIAS, M. M.; MATTOS, M.M.; ROMÃO, W.; TODESCO, J.L. & PACHECO, R.C.S. Data Warehouse – Presente e Futuro. ***Revista Tecnológica*** 7:59-73, 1998.

DOMINGOS, P. Occam's Two Razors: The Sharp and the Blunt. ***Fourth Int. Conf. on Kd & DM***, p. 37-43, 1998.

FAYYAD, U.M. & UTHURUSAMY, R. Proceedings, ***First International Conference on Knowledge Discovery and Data Mining***. Menlo Park, Calif.: The AAAI Press, 1995.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G. & SMYTH, P. ***Advances in knowledge discovery & data mining***. Chapter 1: From data mining to knowledge discovery: an overview. AAAI/MIT, 1996a.

-------. Knowledge Discovery and Data Mining: Towards a Unifying Framework. ***Second International Conference on KD & DM***. Portland, Oregon, 1996b.

FAYYAD, U.M. Editorial. **Data Mining and Knowledge Discovery Journal**. 1 (1), 5-10, 1997.

FERTIG, C.S.; Freitas, A.A.; Arruda, L.V.R. and Kaestner, C. A Fuzzy Beam-Search Rule Induction Algorithm. Principles of Data Mining and Knowledge Discovery: Proc. ***3rd European Conf. (PKDD-99)*** Lecture Notes in Artificial Intelligence 1704, 341-347. Springer-Verlag, 1999.

FERNÉ, G. Science & Technology in the New World Order. ***Ciência e Tecnologia no Brasil: Uma Nova Política para um Mundo Global***[3], 1993.

FIDELIS, M. V.; LOPES, Heitor S. & FREITAS, Alex A. Um Algoritmo Genético para Descobrir Regras de Classificação em Data Mining. ***II ENIA***, julho de 1999.

FREITAS, A.A. ***Generic, Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems***. Ph.D. Thesis, University of Essex, UK, July 1997a.

[3] Este trabalho faz parte de um estudo realizado pela Escola de Administração de Empresas da Fundação Getúlio Vargas por solicitação do MCT e do Banco Mundial, dentro do PADCT II.

FREITAS, A.A. On objective measures of rule surprisingness. **Proc. 2nd European Symp.** On Principles of Data Mining and Knowledge Discovery (PKDD-98), Lecture Notes in Artificial Intelligence. 1510, 1-9, 1998.

-------. A genetic algorithm for generalized rule induction. In: R. Roy et al. Advances in Soft Computing - Engineering Design and Manufacturing, 340-353. (**Proc. WSC3, *3rd On-Line World Conference on Soft Computing***, hosted on the Internet, July 1998.) Springer-Verlag, 1999a.

-------. Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper. To appear in: ***SIGKDD Explorations.*** V.2, 2000a.

------. Evolutionary Algorithms for Data Mining and Knowledge Discovery. To appear in: Ghosh, A. & Tsutsui, S. (Eds.) ***Theory and Application of Evolutionary Computation – Recent Trends,*** 2000b.

------. Evolutionary Algorithms. To appear in: ***Handbook of Data Mining and Knowledge Discovery.*** Oxford University Press, 2001b.

------. A survey of evolutionary algorithms for data mining and knowledge discovery. To appear in: A. Ghosh and S. Tsutsui. (Eds.) **Advances in Evolutionary Computation**. Springer-Verlag, 2002a. (pre-print, unformatted version)

-------. **Data Mining and Knowledge Discovery with Evolutionary Algorithms**. (Forthcoming book). Berlin: Springer-Verlag, 2002b.

FREITAS, A.A. & LAVINGTON, S.H. **Mining Very Large Databases with Parallel Processing**. Kluwer, 1998.

GAGO, Pedro & BENTO, Carlos. A metric for selection of the most promising rules. ***Second European Conference, PKDD-98,*** p. 19-27, 1998.

GOLDBERG, D. E. ***Genetic algorithms in search, optimization, and machine learning.*** New York: Addison-Wesley Publishing Company, Inc., 1989.

GONÇALVES, A. L. ***Utilização de Técnicas de Mineração de Dados na Análise dos Grupos de Pesquisa no Brasil.*** Florianópolis. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-graduação em Engenharia de Produção, UFSC, 2000.

GRAETTINGER, T. Digging up $$$ with data mining – an executive's guide. http://www.tdan.com/i010ht01.htm ***Discovery Corps***, Inc., 1999.

GUIMARÃES, R. Avaliação e Fomento de C&T no Brasil: Propostas para os anos 90. ***MCT/CNPq***, 1994.

HAN, J. & KAMBER, M. ***Data Mining: Concepts and Techniques***. Morgan Kaufmann Publishers, 300 p., 2000.

HAND, David J. ***Construction and Assessment of Classification Rules***. John Wiley & Sons, 1997. (Capítulos 6, 7 e 8)

HANSELMAN, D., LITTLEFIELD, B. ***Matlab: versão do estudante: guia do usuário***. São Paulo: Makron Books, 1997.

HELENE, M. E. M. ***Ciência e Tecnologia de mão dadas com o poder***. 2a Edição, Editora Moderna, 1996.

HERRERA, F. *et al.* Tackling real-coded genetic algorithms: operators and tools for behavioral analysis, ***mimeo***, 1996.

HOLLAND, J.H. Escaping brittleness. In: Michalski, R.S.; Carbonell, J. & Mitchel, T. (Eds.), **Machine Learning**: An artificial intelligence approach (Vol. 2). San Mateo, CA: Morgan Kaufmann, 1986.

HOUCK, C.R., Joines, J. & KAY, M. A genetic algorithm for function optimization: a Matlab implementation. ***ACM Transactions on Mathematical Software***, Submitted 1996.

ISHIBUCHI, H. & NAKASHIMA, T. Designing Compact Fuzzy Rule-Based Systems with Default Hierarchies for Linguistic Approximation. ***CEC-99***, p. 2341-2348, 1999.

JANIKOW, C.Z. A knowledge-intensive genetic algorithm for supervised learning. ***Machine Learning***, 13, p. 189-228, 1993.

-------. A genetic algorithm for optimizing fuzzy decision trees. ***6th Int. Conf. GA***, p. 421-428, 1995.

JOHN, George H. et al. Irrelevant Features and the Subset Selection Problem. ***11th Int. Conf. ML***, p. 121-129, 1994.

JC E-Mail. FHC sanciona projetos aprovados no Congresso de quatro fundos setoriais para financiar pesquisas. ***Jornal da Ciência,*** No. 1590, 25/julho, 2000a.

KACPRZYK, J. & STRYKOWSKI, P. Linguistic Summaries of Sales Data at a Computer Retailer via Fuzzy Logic and a Genetic Algorithm. ***Congress of Evol. Comp. CEC-99***, p. 937-943, 1999.

KIM, YongSeog; STREET, W.N. & MENNCZER, F. Feature Selection in Unsupervised Learning via Evolutionary Search. **KDD 2000**, Boston MA USA, 365-369, 2000.

KING, R. D. et al. STATLOG: Comparison of classification algorithms on large real-world problems. ***Applied Art. Int.*** 9(3), pp. 289-333, May/June 1995.

KOHAVI, Ron & JOHN, George H. The Wrapper Approach. In: Liu. J. & Motoda, H. (Eds) **Feature Extraction, Construction and Selection: a data mining perspective**. Kluwer, 1998.

KONDO, Edson K. Desenvolvendo Indicadores Estratégicos em Ciência e Tecnologia: as Principais Questões. ***Ciência da Informação***, Brasília, v.27, n.2, p. 128-133, maio/ago. 1998.

KRIEGER, E.M. Science in Brazil – An Overview. ***Brazilian Academy of Sciences***, 1999.

KUHN, T. S. ***A estrutura das revoluções científicas***. Editora Perspectiva, 1982.

LEE, Mal-Rey Generating fuzzy rules by genetic method and its application. ***International Journal on Artificial Intelligence Tools***, Vol. 7, No. 4, pp. 399-413, 1998.

LIU, Bing & HSU, Wynne. Post-analysis of learned rules. ***AAAI-96***, p. 828-834, 1996.

LIU, B.; HSU, W. & CHEN, S. Using general impressions to analyze discovered classification rules. ***Third Int. Conf. on KD and DM, KDD-97***, p. 31-36, 1997.

LONGO, Mauricio B. & Smith Jr., R. **Delphi 4**. Rio de Janeiro: Brasport, 1999.

MACHADO, C. Como dar o tiro certo na hora de decidir. ***Informática Exame***, 1996. http://www2.ulo.com.br/info/arquivo/ie120/capa.html, 1998.

MACIAS-CHAPULA, C. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. ***Ciência da Informação***, Brasília, v.27, n.2, p. 134-140, maio/ago. 1998.

MACLEAN, M. *et al.* Identifying Research Priorities in Public-Sector Funding Agencies: Mapping Science Outputs onto User Needs. ***Technology Analysis and Strategic Management***, v. 10 (issue 2), 1998.

MAJOR, John A. & MANGANO, John J. Selecting among rules induced from a hurricane database. ***Knowledge Discovery in Databases Workshop, AAAI-93***, p. 28-44, 1993.

MARTINS, G. M., GALVÃO, G. O diretório dos grupos de pesquisa no Brasil: perspectivas de fomento e avaliação. ***Educação Brasileira***, v.16, n. 33, p. 11-29, 1994.

MCHUGH, L. Who's Doing What With Data Mining. Teradata Review Summer, http://www.teradatareview.com/summer99/mchugh.html, 1999.

MCT. ***Ciência e tecnologia no Brasil: uma nova política para um mundo globalizado.*** Brasília. http://www.mct.gov.br. 1994.

_____. ***Plano Plurianual 1996-99***. Brasília. http://www.mect.gov.br/documentos. 1996.

_____. Através do Conexão C&T. **Notícias do MCT** via E-mail, 7/dezembro/2000.

MEC divulga ranking entre Universidades. Revista ***Veja***. p. 35, 15-9-1999.

MENDEL, J. M. Fuzzy logic systems for engineering: a tutorial. ***Proceedings of the IEEE***. v.83, n.3, p. 345-377, 1995.

MICHALEWICZ, Z. ***Genetic Algorithms + Data Structures = Evolution Programs***. 3rd., Chapter 12: Machine Learning, Springer-Verlag, 1996.

MICHIE, D. et al. (Eds.) ***Machine Learning, Neural and Statistical Classification***, Chapters 1, 2 and 11, 1994.

MITCHELL, T. ***Machine Learning***. McGraw-Hill, 1997.

MORALES, A. T. Identificação Difusa de Sistemas: Proposta de um Modelo Adaptativo. Florianópolis. **Tese** (Doutorado em Engenharia de Produção) – Programa de Pós-graduação em Engenharia de Produção, UFSC, 1997.

MOTA, C.; FERREIRA, H. & ROSA, A. Independent and Simultaneous Evolution of Fuzzy Sleep Classifiers by Genetic Algorithms. ***GECCO-99***, p. 1622-1629, 1999.

NIGRO, M. O melhor caminho até seu cliente. ***Byte Brasil***, p. 44-66, janeiro de 1997.

NODA, E.; FREITAS, A. A. & LOPES, H. S. Discovering interesting prediction rules with a genetic algorithm. ***Proc. Congress on Evolutionary Computation (CEC-99)***, 1322-1329. Washington D.C., USA, July 1999.

OCDE. ***Manual de Frascati***: propuesta de norma práctica para encuestas de investigación y desarrollo experimental. Paris, 1993.

OCDE. The measurement of scientific and technological activities: manual on the measurement of human resources devoted to S&T, ***Canberra Manual***. Paris, 1995.

OCDE. ***Oslo Manual***: proposed guidelines for collecting and interpreting technological innovation data. Paris, 1997.

PÁIRCÉIR, R.; McCLEAN, S. & SCOTNEY, B. Discovery of Multi-level Rules and Exceptions from a Distributed Database. **KDD 2000**, Boston MA USA, 523-532, 2000.

PANDYA, A. S. & MACY, R. B. Pattern Recognition with Neural Networks. ***IEEE Press***, 1995. P.319-369.

PAZZANI, M.J. Knowledge discovery from data? **IEEE Intelligent Sustems**, 15(2), p.10-13, March/April 2000.

PENA-REYES, C. A. & SIPPER, M. Designing Breast Cancer Diagnostic Systems via a Hybrid Fuzzy-Genetic Methodology. ***Proc. Fuzzy-IEEE'99***, 1999.

PERNELL, C., THEMLIN, J-M, RENDERS, J-M. & ACHEROY, M. Optimization of fuzzy expert systems using genetic algorithms and neural networks. **IEEE Trans. on Fuzzy Systems**, v. 3, n. 3, p. 300-312, 1995.

PIATETSKY-SHAPIRO, G. & MATHEUS, C. The Interestingness of Deviations, in **Proceedings of KDD-94 workshop**, AAAI Press, 1994.

PPA - ***Plano Plurianual 2000-2003***: Mensagem ao Congresso Nacional - Brasília : Ministério do planejamento, orçamento e gestão, Secretaria de Planejamento e Avaliação, http://www.abrasil.gov.br/index.htm, 1999.

PUNCH, W. F.; GOODMAN, E. D.; PEI, M.; CHIA-SHUN, L.; HOVLAND, P. & ENBODY, R. Further research on feature selection and classification using genetic algorithms. Proc. **5th Int. Conf. Genetic Algorithms** (ICGA-93), 557-564, 1993.

QUINLAN, J.R. Generating production rules from decision trees. ***Proc. IJCAI-87***, p. 304-307, 1987.

------. **C4.5: Programs for Machine Learning.** Morgan Kaufmann, 1993.

REISDORPH, Kent **Sams Teach Yourself Delphi 4 in 21 Days.** Sams Publishing, Borland Press, 918 p., 1998.

RENDELL, Larry & CHO, Howard Empirical Learning as a Function of Concept Character. ***ML*** 5(3), pp. 267-298, 1990.

ROMÃO, Wesley et al. Uma visão geral sobre rede neural artificial com arquitetura ART2. ***Revista Tecnológica***, n.6, p. 59-71, 1999a.

ROMÃO, Wesley; NIEDERAUER, Carlos A.P.; MARTINS, Alejandro; MORALES, Aran Tcholakian & PACHECO, Roberto C.S. Algoritmos genéticos e conjuntos difusos aplicados ao controle de um processo térmico. ***Revista Tecnológica***, n.8, p. 7-21, 1999b.

ROMÃO, Wesley; NIEDERAUER, Carlos A.P.; MARTINS, Alejandro; MORALES, Aran Tcholakian; PACHECO, Roberto C.S. & BARCIA, Ricardo M. Extração de regras de associação em C&T: O algoritmo Apriori. ***XIX Encontro Nacional em Engenharia de Produção***, 1999c.

SAUL, A. M. & ABRAMOWICZ, M. Avaliação da pós-graduação: superamos os limites? ***Educação Brasileira.*** Brasília, v. 19, n. 38, 111-119. 1997.

SCHREIBER, August Th. *et al.* **Knowledge Engineering and Management: the CommonKADS methodology**. MIT Press, Chapter 1, 2000.

SCHWARTZMAN, Simon *et al.* ***Ciência e Tecnologia no Brasil: A capacitação brasileira para a pesquisa científica e tecnológica.*** V. 3, 420p. Rio de Janeiro: Editora Fundação Getulio Vargas, 1996.

SCHWARTZMAN, S. & CASTRO, C. M. ***Pesquisa universitária em questão***. Editora da UNICAMP, Ícone Editora, São Paulo, CNPq, 1986.

SHANAHAN, J. G. et al. Constructive Induction of Fuzzy Cartesian Granule Feature Models using Genetic Programming with applications. ***CEC-99***, p. 218-225, 1999.

SILBERSCHATZ, Avi & TUZHILIN, Alexander What Makes Patterns Interesting in Knowledge Discovery Systems. ***IEEE Transactions on Knowledge and data engineering***, Vol. 8, No. 6, pp. 970-974, December 1996.

SILVA, Edna Lúcia da & MENEZES, E.M. **Metodologia da pesquisa e elaboração de dissertação**. Florianópolis: Laboratório de Ensino a distância da UFSC, 118p., 2000.

SILVEIRA JR, A. & VIVACQUA, G.A. ***Planejamento Estratégico como Instrumento de Mudança Organizacional.*** Editora UnB, 1996.

SMITH, S.F. **A learning system based on genetic algorithms**. Doctoral dissertation, University of Pittsburgh, Department of Computer Science, 1980.

SPINAK, E. Indicadores Cienciométricos. ***Ciência da Informação,*** Brasília, v. 27, n. 2, p. 121-148, maio/ago., 1998.

SRIKANT, R. & AGRAWAL, R. Mining quantitative association rules in large relational tables. ***Proc. of ACM SIGMOD Conf. on Management of Data.*** Canadá, 1996.

STATSOFT Portugal Ltda. ***Data mining com o statistica***. Disponível na Internet. http://www.statsoftinc.com/portugal/datamin.html. 28 junho 1998.

SZMRECSÁNYI, T. Avaliação em Ciência e Tecnologia: Necessidade, Critérios e Procedimentos. ***Revista de Administração***, 1987.

THULSTRUP, E. W. A Qualidade da Pesquisa nos Países em Desenvolvimento. Título original: Improving the Quality of Research in Developing Country Universities. ***PHREE Background Paper Series***, 1992.

TIMES. Latin American Editon, 22/5/2000.

TURNEY, P. D. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligent Research. Two Crows Corporation. **Introduction to Data Mining and Knowledge Discovery**, 2nd edition, 369-409, 1995.

VELHO, L. Indicadores de C&T: Como medir a ciência? ***Revista Brasileira de Tecnologia***, Brasília, v.16(1), jan./fev. 1985.

_____. Avaliação Acadêmica. A hora e a vez do "baixo clero". ***Ciência e Cultura***. 41(10): 957-968, outubro 1989.

VOZNIKA, F. & MENDES, R. Garimpeiro: Mineração de Dados com Programação Genética e Lógica Fuzzy. ***Documentação Acadêmica***, PPGIA, PUC-PR, 2000.

WANG, Ke & SUNDARESH, S. Selecting Features By Vertical Compactness of Data. In: Liu. J. & Motoda, H. (Eds) **Feature Extraction, Construction and Selection: a data mining perspective**. Kluwer, 1998.

WEISS, Sholom M. & KULIKOWSKI, C. A. **Computer Systems That Learn.** Morgan Kaufmann, Chapter 2, p. 17-49, 1991.

WITTEN, Ian H. & FRANK, Eibe **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.** Chapter 8: Nuts and bolts: machine learning algoritnms in Java. Morgan Kaufmann Publishers, 2000.

XIONG, N. & LITZ, L. Generating Linguistic Fuzzy Rules for Pattern Classification with Genetic Algorithms. ***PKDD-99***, p. 574-579, 1999.

YANG, Jihoon & HONAVAR, Vasant Feature Subset Selection Using a Genetic Algorithm. In: Liu. J. & Motoda, H. (Eds) **Feature Extraction, Construction and Selection: a data mining perspective**. Kluwer, 1998.

ZADEH, L. A. Fuzzy Sets. **Information and Control**, 8, 338-353, 1965.

ZIMMERMANN, H-J. ***Fuzzy set theory.*** Boston: Kluwer Academic Publishers, 3ª ed. 1996.

10 ANEXOS

10.1 ANEXO A – Experimento Preliminar

Formalização da Tarefa de Associação

Seja um banco de dados contendo informações sobre os pesquisadores atuantes no Brasil. Almeja-se descobrir associações importantes entre esses dados, onde:

- $I = \{i_1, i_2, ..., i_m\}$ é um conjunto de literais, denominados itens. São as características e atributos dos pesquisadores. Por exemplo, I = {idade, sexo, ..., área de atuação, artigos publicados};
- T é um conjunto de certos itens de um pesquisador, tal que $T \subseteq I$;
- D é uma tabela representando todas as características e atributos dos pesquisadores; e
- X e Y são conjuntos de itens específicos dos pesquisadores, tais que $X \subseteq T$ e $Y \subseteq T$.

Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, onde $X \subset I$, $Y \subset I$ e $X \cap Y = \varnothing$. A regra $X \Rightarrow Y$ pertence a D com confiança c se $c\%$ dos registros em D que contêm X também contêm Y. A regra $X \Rightarrow Y$ tem suporte s em D se $s\%$ dos registros em D contém $X \cup Y$ (Agrawal & Srikant, 1994). Então, dado uma tabela D, o objetivo é descobrir as regras de associação relevantes.

O problema de descobrir todas as regras de associação, tal como formulado por Agrawal, pode ser decomposto em duas etapas:

- encontrar todos os conjuntos de itens (*itemsets*) que apresentam suporte maior que o suporte mínimo estabelecido pelo usuário. Os *itemsets* que atendem a este quesito são denominados *itemsets freqüentes*; e
- utilizar os *itemsets freqüentes* obtidos para gerar as regras de associação do banco de dados.

A maior parte do tempo de processamento da mineração de regras de

associação é determinada pela primeira etapa, a qual exige sucessivas buscas na base de dados.

Encontrando o conjunto dos *itemsets freqüentes*, as regras de associação correspondentes podem ser diretamente identificadas. Algoritmos que realizam a contagem eficiente dos *itemsets* são a chave para o sucesso dos métodos de mineração em grandes bancos de dados (Chen *et al.*, 1996).

O Algoritmo Apriori

O algoritmo *Apriori* é um dos algoritmos mais conhecidos quando o assunto é mineração de regras de associação em grandes bancos de dados centralizados. Ele encontra todos os conjuntos de itens freqüentes, denominados *itemsets freqüentes* (L_k).

Ele pode trabalhar com um número grande de atributos, gerando várias alternativas combinatórias entre eles. O algoritmo Apriori realiza buscas sucessivas em toda a base de dados, mantendo um ótimo desempenho em termos de tempo de processamento (Agrawal & Srikant, 1994).

O algoritmo principal (Apriori) faz uso de duas funções: a função Apriori_gen, para gerar os candidatos e eliminar aqueles que não são freqüentes, e a função Genrules, utilizada para extrair as regras de associação. O algoritmo e as duas funções foram implementadas através do ambiente de simulação Matlab® (Hanselman & Littlefield, 1997).

O primeiro passo do algoritmo é realizar a contagem de ocorrências dos itens para determinar os *itemsets freqüentes* de tamanho unitário (1-*itemsets freqüentes*). Os passos posteriores, k, consistem de duas fases. Primeiro, os *itemsets freqüentes* L_{k-1}, encontrados no passo anterior (k-1) são utilizados para gerar os conjuntos de itens potencialmente freqüentes, os *itemsets candidatos* (C_k). O procedimento para geração de candidatos é descrito no parágrafo seguinte. Na seqüência, é realizada uma nova busca no banco de dados, contando-se o suporte de cada candidato em C_k.

Na geração dos *itemsets candidatos* tomam-se como argumento L_{k-1}, o conjunto de todos (k-1)-*itemsets freqüentes*. Utiliza-se a função *Apriori_gen*, que retorna um superconjunto de todos os *k-itemsets freqüentes.* A intuição por

trás desse procedimento é que se um *itemset* X tem suporte mínimo, todos os seus subconjuntos também terão (Agrawal & Shafer, 1996). A função, em um primeiro estágio, une L_{k-1} com L_{k-1}. No estágio seguinte, são eliminados os *itemsets* $c_k \in C_k$, desde que um dado (k-1)-*subset* de c_k não pertença a L_{k-1}.

O último passo é a descoberta das regras (função *Genrules*). A geração de regras, para qualquer *itemset freqüente*, significa encontrar todos os *subsets* de *l* contendo pelo menos dois itens. Assim, para todo e qualquer *subset a*, produz-se uma regra $a \Rightarrow (l - a)$ somente se a razão (suporte (*l*)/suporte(*a*)) é ao menos igual à confiança mínima estabelecida pelo usuário.

Para gerar regras com múltiplos conseqüentes, são considerados todos os *subsets*. Por exemplo, dado um *itemset ABCD*, considera-se primeiro o *subset ABC*, seguido de *AB*, etc. Se $ABC \Rightarrow D$ não atinge uma confiança suficiente (confiança < *minconf*), não é necessário verificar se $AB \Rightarrow CD$.

Exemplo de Extração de Regras

O exemplo a seguir, adaptado de Chen *et ai.* (1996) e de Agrawal & Srikant (1994), demonstra como funciona a extração de regras de associação através do algoritmo *Apriori*.

Tabela 21: Dados do fictício grupo de pesquisa GP.

Registro (pesquisador)	Itens Categóricos				
	A	*B*	*C*	*D*	*E*
501	1	0	0	1	0
502	0	1	1	0	1
503	1	0	1	0	1
504	0	1	1	0	1
505	0	1	1	0	0
freqüência	2	3	4	1	3

Legenda:
A = pesquisador estrangeiro; 0 = ausência do atributo; e
B = pesquisador brasileiro; 1 = presença do atributo.
C = sexo feminino;
D = sexo masculino; e
E = pesquisador doutor.

Supondo um banco de dados formado somente por um grupo de pesquisa, GP. Supondo, também, que este grupo seja composto por cinco

pesquisadores, associando-se a cada um deles cinco itens binários, conforme a Tabela 21.

Seja C_k o conjunto de *k-itemsets candidatos*, onde $k = 5$. Cada membro c_k deste conjunto tem dois campos: *itemset* e *contador de suporte*, representados, respectivamente, por *its* e *cs* no Quadro 16. Seja L_k o conjunto dos *k-itemsets freqüentes*. De modo análogo, cada membro do conjunto também possui *its* e *cs*.

O primeiro passo do algoritmo conta a freqüência que os itens ocorrem para determinar os 1-*itemsets freqüentes* (última linha da Tabela 21).

Posteriormente, obtém-se o conjunto de candidatos 1-*itemsets*, C_1, mostrado no Quadro 16. Assumindo um suporte mínimo igual a dois, ou seja, *minsup* = 40%, L_1 é composto pelos elementos de C_1 com suporte igual ou superior a 40%. No exemplo, somente o *itemset D* não atendeu a esta condição, ficando L_1 composto por {*A*}, {*B*}, {C} e {*E*}.

Quadro 16: Geração dos *itemsets candidatos* (*C*) e dos *freqüentes*(*L*)

C1		L1	
its	*cs*	*its*	*cs*
{*A*}	2	{*A*}	2
{*B*}	3	{*B*}	3
{*C*}	4	{*C*}	4
{*D*}	1	{*E*}	3
{*E*}	3		

C2		L2	
its	*cs*	*its*	*cs*
{*AB*}	0	{*BC*}	3
{*AC*}	1	{*BE*}	2
{*AE*}	1	{*CE*}	3
{*BC*}	3		
{*BE*}	2		
{*CE*}	3		

C3		L3	
is	*cs*	*its*	*cs*
{*BCE*}	2	{*BCE*}	2

Para descobrir o conjunto dos 2-*itemsets freqüentes*, de modo a continuar satisfazendo ao suporte mínimo, o *Apriori* usa a concatenação $L_1 * L_1$ para gerar o conjunto candidato C_2, que consiste de 2-*itemsets*. Por exemplo, {*C*} e {*E*} geram {*CE*}. Mais uma vez, cada ocorrência é computada. No caso {*CE*} ocorre três vezes em GP (registros 502, 503 e 504). L_2 é determinado com base no suporte de cada candidato de C_2. Agora são excluídos {*AB*}, {*AC*} e {*AE*}, pois têm suporte inferior ao mínimo estabelecido.

A geração de C_3 é obtida a partir de L_2 de uma maneira distinta. Os futuros *itemsets* candidatos devem manter uma ordem lexicográfica, tal que quando a

concatenação L_2* L_2 for realizada, o primeiro item de um *itemset* seja idêntico ao primeiro item do outro *itemset* e assim sucessivamente. Porém, o último item do *itemset* deve ser menor, lexicograficamente, que o último item do outro *itemset.* Esta regra pode ser representada como:

$$itemset\_p = \{p_1, p_2, ..., p_n\}, itemset\_q = \{q_1, q_2, ..., q_n\} \quad (10.1)$$

sendo necessário que:

$$p_1 = q_1, p_2 = q_2, ..., p_n < q_n \quad (10.2)$$

No Quadro 16, o *itemset* candidato {*BCE*}, em C_3, foi formado concatenando {*BC*} com {*BE*}, pois *B* = *B* e *C* < *E*. Este foi o único conjunto que pôde ser formado, pois não há outra concatenação que satisfaça a Equação 10.2. A concatenação {*BC*}*{*CE*}, por exemplo, não satisfaz a Equação 10.2, pois, lexicograficamente, $p_1 = B$ é menor que $q_1 = C$.

Tabela 22: Ocorrências dos conjuntos de atributos

Registro (pesquisador)	Itens Categóricos		
	B	*C*	*E*
501	0	0	0
502	1	1	1
503	0	1	1
504	1	1	1
505	1	1	0
Freqüência	3	4	3

O passo seguinte é descobrir as regras de associação. No caso do fictício grupo de pesquisa GP, supondo uma confiança mínima de 60% e mantendo o suporte mínimo em 40%, uma regra provável seria *BC* ⇒ *E*. Para ela, a confiança é igual a suporte(*BCE*)/suporte(*BC*), cujo resultado é $\frac{2}{3}$, ou 66%, satisfazendo a condição imposta (ver Tabela 22).

Para a regra *BC* ⇒ *E* ou (pesquisador brasileiro; sexo feminino) ⇒ (pesquisador doutor), seu suporte seria o percentual de ocorrências de *BCE* com relação ao total de pesquisadores do grupo, que resulta em 40% ($\frac{2}{5}$).

Então, esta é uma regra válida. Isto equivale a dizer que, das pesquisadoras brasileiras, 66% têm doutorado, muito embora estas brasileiras portadoras do título de doutor correspondam a apenas 40% dos indivíduos do grupo.

Outra provável regra seria *B* ⇒ *CE*. Para esta situação, o valor da confiança

seria idêntico, pois a razão suporte (*BCE*)/suporte (*B*) também é igual a $\frac{2}{3}$.

Resultados preliminares da aplicação do algoritmo Apriori aos dados do Diretório 3.0 podem ser observados na seção a seguir.

Aplicação do Algoritmo Apriori sobre o Diretório 3.0

Os dados apresentados no Capítulo 2 são organizados pelo CNPq. Uma das metas do CNPq é construir ferramentas para a aquisição e análise de dados mais aprofundados relativos ao sistema brasileiro de C&T. Esta expectativa motivou a aplicação de *MD* à versão 3.0 do Diretório, a qual estava disponível na época. Uma técnica de MD, chamada de Extração de Regras de Associação, foi avaliada (Romão *et al.*, 1999c) quando aplicada sobre esta versão do Diretório.

Calcado no objetivo de demonstrar o potencial do algoritmo *Apriori* na extração de regras para apoio à gestão de C&T, foi utilizado todo o universo de 34.040 pesquisadores cadastrados no Diretório 3.0, os quais foram escolhidos como unidades de análise.

A base de dados da versão 3.0 do Diretório, conforme consta na seção 2.6, contém informações sobre pesquisadores distribuídos em 8.632 grupos de pesquisa pertencentes a 181 instituições. Esses pesquisadores geraram 339.568 produtos, tais como artigos em periódicos, comunicações em eventos, livros, formação de recursos humanos, etc.

Para facilitar a implementação do algoritmo A*priori* e da função *Apriori-gen*, extraídos de Agrawal & Srikant (1995), efetuou-se uma adaptação da estrutura de dados substituindo a estrutura original, em forma de *hash tree*, por matrizes, mantendo a lógica e seqüência originais do algoritmo.

A metodologia empregada foi a seguinte:

- pré-processamento dos dados;
- aplicação do algoritmo *Apriori*; e
- análise e refinamento dos resultados.

O primeiro passo do pré-processamento consistiu em realizar a seleção de alguns itens, dos pesquisadores, considerados relevantes. Esses itens

são os atributos que caracterizam um pesquisador, tais como sua idade e sexo, e os indicadores de produção de C&T, como a quantidade de artigos publicados no período 1995-97.

Quadro 17: Itens selecionados para o algoritmo *Apriori*

ITENS CATEGÓRICOS
Nacionalidade (brasileira, estrangeira);
Idade ([>24], [25..29], [30..34], [35..39], [40..44], [45..49], [50..54], [55..59], [60..64], [≥65]);
Sexo (masculino, feminino);
Titulação Máxima (graduação, especialização/aperfeiçoamento, mestrado, doutorado);
Grande Área do Conhecimento (Ciências Agrárias, Ciências Biológicas, Ciências das Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias e Ciência da Computação, Lingüística, Letras e Artes);

ITENS QUANTITATIVOS
Artigos Publicados em Periódicos (0, [1..2], [3..5], [6..10], [11..20], [21..30], [31..40], [≥41]);
Dissertações Orientadas ([0], [1], [2], [3], [4..7], [8..12], [≥13]); e
Teses Orientadas ([0], [1], [2], [3], [4..5], [6..7], [8..9], [≥10]).

Esta seleção permitiu testar o algoritmo em um computador pessoal, através do *software* Matlab, o qual se ajusta muito bem às aplicações acadêmicas. Todos os itens foram extraídos dos módulos "Recursos Humanos" e "Produção Científica e Tecnológica", ambas pertencentes ao Diretório 3.0. Foram escolhidos cinco itens categóricos e três quantitativos, conforme relação apresentada no Quadro 17.

Os itens categóricos, tais como Nacionalidade e Sexo, foram representados na forma booleana. Por exemplo: O item Sexo foi representado pelos atributos booleanos Sexo_M e Sexo_F. A alternativa de atributos booleanos é uma forma de representar dados categóricos para possibilitar o uso do algoritmo Apriori "convencional".Os itens quantitativos foram estratificados em faixas. Exemplificando, o item Idade foi dividido em faixas com amplitude de 5 anos e cada faixa foi considerada como um item booleano. O mesmo foi efetuado com os itens Artigos Publicados em Periódicos, Dissertações e Teses Orientadas.

Este pré-processamento gerou um arquivo contendo 33.675 registros com

50 campos cada um, onde cada registro representa um pesquisador.

Esta abordagem, de transformação de atributos categóricos e quantitativos em itens booleanos, permitiu gerar uma tabela de dados onde a ausência de um item é representada pelo valor '0' e a presença pelo valor '1'. Assim, se um pesquisador é mulher, o item Sexo_F recebe o valor '1' e o item Sexo_M recebe o valor '0'.

Feito isto, a etapa seguinte é aplicar o algoritmo *Apriori* para encontrar, entre os itens booleanos, as regras de associação que sejam relevantes para o decisor. Isto é obtido através de delimitações (restrições) impostas pelo decisor na execução do algoritmo, através da estipulação de suporte e confiança mínimos. A metodologia empregada pode ser resumida nos seguintes passos:

- selecionar os atributos quantitativos e categóricos;
- transformar os atributos quantitativos em categóricos, dividindo em intervalos iguais;
- extrair os itens selecionados, através da linguagem SQL (*Structured Query Language*), gerando uma tabela booleana;
- calcular o suporte para cada item e para conjuntos de itens;
- encontrar todos os conjuntos de itens cujo suporte seja pelo menos igual ao suporte mínimo, obtendo o conjunto de itens freqüentes;
- a partir do conjunto de itens freqüentes, gerar as regras de associação que apresentem confiança pelo menos igual à confiança mínima; e
- escolher manualmente as regras relevantes.

No estudo, duas abordagens foram utilizadas. A primeira abordagem realizou várias simulações utilizando diversos valores de suporte e confiança mínimos, em busca dos itens mais freqüentes, conforme a proposta original do algoritmo. A segunda abordagem tratou dos itens menos freqüentes, acrescentando ao algoritmo um limite máximo para o suporte.

Na próxima seção apresentam-se alguns resultados com a implementação do algoritmo Apriori na tarefa de extração de regras de associação.

Extraindo Regras de Associação

A Tabela 23 resume o número de regras encontradas com a abordagem convencional, em função de suporte e confiança mínimos estabelecidos e considerando a busca por itens mais freqüentes. A primeira constatação é que o número de regras geradas é inversamente proporcional aos valores determinados para *minconf* e *minsup*. À medida que estes últimos decrescem, aumenta o número de regras. Esta constatação era esperada visto que, quanto menores os valores de *minsup* e *minconf*, maiores as chances de encontrar regras mais especializadas (no sentido de conterem um maior número de itens), o que tende a ser indesejável.

Tabela 23: Resultados do algoritmo Apriori

Simulação	Suporte mínimo (minsup)	Confiança mínima (minconf)	Nº de regras geradas
S1	70%	80%	5
S2	60%	50%	8
S3	50%	40%	29
S4	50%	20%	29

Para a simulação S1, o número reduzido de regras foi acompanhado de alguma redundância. Por exemplo, observe as regras a seguir:

- R1: (brasileiro, não orientou tese) $\Rightarrow$ (não orientou dissertação), c = 85%, s = 76,4%.
- R2: (brasileiro) $\Rightarrow$ (não orientou tese, não orientou dissertação), c = 80%, s = 76,4%

Com a redução de *minconf* e *minsup*, o aumento no número de regras resultou na seleção de outros itens. Com *minsup* = 60% e *minconf* = 50% (simulação S2), além de quatro regras idênticas as da simulação S1, outras regras selecionaram somente homens, ou somente doutores. Isto é coerente e indica que o algoritmo foi implementado corretamente, pois 58% dos pesquisadores inventariados pelo Diretório são homens e 55% são doutores (CNPq, 1998).

Para as simulações S3, S4 e S5, o aumento no número de regras foi

acompanhado do aumento de redundância. Porém, novas regras foram descobertas. Agora, as regras passaram a considerar mulheres (42% dos pesquisadores do Diretório são mulheres (CNPq, *op. cit.*)).

Mas, em se tratando de pesquisadores, é bastante provável que o decisor esteja interessado em descobrir o comportamento das minorias, posto que elas podem revelar pesquisadores com alta performance ou, por outro lado, com baixo rendimento. Este é o enfoque a seguir.

Modificando o Algoritmo Apriori

Para buscar regras que possam revelar as exceções do Diretório, adotou-se duas medidas. Primeiro, reduziu-se bastante o Suporte Mínimo. Segundo, introduziu-se uma restrição adicional ao suporte, que passou a ter um limite máximo (*maxsup*), além do limite mínimo.

Essa restrição adicional descarta regras com alto suporte, atuando como um filtro que efetivamente seleciona apenas regras com suporte relativamente baixo (entre *minsup* e *maxsup*).

A Tabela 24 mostra o número de regras geradas para três simulações, com a introdução de limite máximo ao suporte. As simulações geraram regras até certo ponto óbvias, apesar de destacar "minorias".

Tabela 24: Resultados impondo-se um suporte máximo

Simulação	Suporte mínimo (minsup)	Suporte máximo (maxsup)	Confiança mínima (minconf)	Nº de regras geradas
S6	1%	10%	15%	3
S7	1%	20%	25%	4
S8	3%	49%	20%	19

Duas das regras obtidas foram:

- Regra 1, S7: (título = graduação) $\Rightarrow$ (idade < 24 anos), s = 1,05%, c = 15,85%.
- Regra 3, S8: ([6..10] artigos produzidos) $\Rightarrow$ (titulou um mestre), s = 1,07%, c = 16,38%.

A interpretação de s na Regra 1 da simulação S7 é: 1,05% de todos os pesquisadores são graduados e possuem menos de 24 anos.

A interpretação de *c* é: 15,85% dos graduados têm menos de 24 anos, ou 84,15% desses pesquisadores têm 24 anos ou mais.

A conclusão é que há um pequeno contingente muito jovem envolvido com pesquisa. Porém, o decisor poderia ser levado a perguntar, por exemplo, qual é a faixa etária dos 84,15% restantes? Há pessoas nesta situação em idade avançada? Porque não realizaram um mestrado ou doutorado? Será que não estaria havendo uma certa confusão em registrar pessoal técnico como pesquisador? Ou seja, as duas regras, mais do que constatações normais, induzem a explorar em maior profundidade os dados.

Na simulação S8 foi possível encontrar algumas características dos pesquisadores do sexo feminino e dos mestres entre 30 e 35 anos, entre outras.

Analisando outras regras, observou-se que os pesquisadores das Ciências Exatas, Biológicas e Saúde publicaram até sete artigos no período; os das Ciências Agrárias e Humanas publicaram até cinco artigos, ao passo que os das Engenharias, Ciências Sociais e Lingüística, Letras e Artes publicaram no máximo dois artigos. Isto pode estar revelando o perfil dessas áreas no tocante à disseminação do conhecimento em periódicos científicos.

Por último, utilizados baixos valores para o suporte e a confiança, surgiram os pesquisadores estrangeiros, onde 3,18% produziram entre 11 e 20 artigos, 2,85% formaram entre 4 e 7 mestres e 4,88% têm mais de 64 anos. Ou seja, são grupos seletos (os estrangeiros são minoria) e de alta produtividade.

Conclusão

As simulações, utilizando dados correspondentes a todos os estados brasileiros, foram realizadas sobre alguns atributos e indicadores dos pesquisadores inventariados pelo Diretório. Regras de associação consistentes foram geradas e nichos específicos foram descobertos. Por um lado, algumas delas são previsíveis, por outro lado o algoritmo exige o aprofundamento do conhecimento por parte do decisor, onde a sua sensibilidade e experiência são fundamentais. Ele deve intervir estipulando limites para o suporte e a confiança.

Nesse experimento promoveram-se pequenas modificações no algoritmo *Apriori*, em especial a introdução de um limite superior para o suporte. Com isto, foi possível explorar com maior profundidade as informações de grupos seletos de pesquisadores. Entretanto, uma limitação ficou evidente: a necessidade de ajustes de modo a eliminar problemas como a redundância de regras.

Em resumo, diferentemente da proposta original do *Apriori*, concebido para extrair regras onde a alta confiança é essencial, para descobrir regras relevantes no contexto dos grupos de pesquisa a confiança da regra pode ser empregada com valores baixos, caracterizando uma descrição de exceções. Porquanto, importa saber não a organização da maioria, mas o perfil e o desempenho de alguns poucos pesquisadores.

Os resultados demonstraram que este tipo de técnica gera muitas regras, além de exigir sucessivos acessos ao banco de dados. Estes resultados motivaram a exploração de técnicas de Mineração de Dados capazes de fornecer informações mais resumidas e estratégicas para o apoio à tomada de decisão em C&T.

10.2 ANEXO B – Resultados

Primeiro Experimento

Regras Descobertas no Primeiro Experimento - Página ½

Regras Descobertas para IG's do Prof. Alex A. Freitas

EXPERIMENTO: Todos os dados, Metas NÃO são previsores, Sem o atributo GRANDE_AREA.

CLASSE REGRA
Interesse IG Freq_da_Classe Cobertura(SC+SE)

1 SE (CIDADE = 12) (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto)
ENTÃO (ARTIGOS_PER_NAC = baixo)
0,67 6 0,472 2,10

2 SE (TESES_ORIENT = alto)
ENTÃO (ARTIGOS_PER_NAC = médio)
0,50 5 0,503 49,00

3 SE (SEXO = M) (CURSOS_MINISTRADOS = alto) (DISSERT_ORIENT = baixo)
ENTÃO (ARTIGOS_PER_NAC = alto)
0,33 7 0,025 20,29

4 SE (ESCRITA_INGLES = P) (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto)
ENTÃO (ARTIGOS_PER_INT = baixo)
0,67 2 0,647 38,50

5 SE (ESCRITA_INGLES = B) (DISSERT_ORIENT = alto) (TESES_ORIENT = alto) (INIC_CIENT_ORIENT = baixo)
ENTÃO (ARTIGOS_PER_INT = médio)
0,25 1 0,292 5,50

6 SE (ORIGEM = 6) (DISSERT_ORIENT = baixo) (TESES_ORIENT = médio)
ENTÃO (ARTIGOS_PER_INT = alto)
0,33 3 0,060 4,50

7 SE (NACIONALIDADE = E) (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto)
ENTÃO (CAP_LIVROS_NAC = baixo)
0,67 14 0,772 15,70

8 SE (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto) (CURSOS_MINISTRADOS = alto)
ENTÃO (CAP_LIVROS_NAC = médio)
0,33 14 0,209 4,19

Continua na Página 2

Regras Descobertas no Primeiro Experimento - Página 2/2

META REGRA
Interesse IG FreqTotal Cobertura(SC+SE)

9 SE (CIDADE = 8) (TRAB_TECNICO = baixo) (CURSOS_MINISTRADOS = alto) (DISSERT_ORIENT = baixo)
ENTÃO (CAP_LIVROS_NAC = alto)
0,25 15 0,020 1,43

10 SE (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto) (MONOGR_ORIENT = alto)
ENTÃO (CAP_LIVROS_INT = baixo)
0,67 10 0,934 2,20

11 SE (NACIONALIDADE = E) (CIDADE = 16) (ESCRITA_INGLES = P) (DISSERT_ORIENT = baixo)
ENTÃO (CAP_LIVROS_INT = alto)
0,50 11 0,066 1,00

12 SE (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto) (DISSERT_ORIENT = baixo)
ENTÃO (ED_PUBLIC_LIVRO_NAC = baixo)
0,67 22 0,837 126,10

13 SE (CIDADE = 15) (CURSOS_MINISTRADOS = médio) (DISSERT_ORIENT = baixo) (MONOGR_ORIENT = alto)
ENTÃO (ED_PUBLIC_LIVRO_NAC = alto)
0,25 23 0,163 2,43

14 SE (CIDADE = 1) (NIVEL_FORMACAO = alto) (ANOS_FORMADO = alto)
ENTÃO (ED_PUBLIC_LIVRO_INT = baixo)
0,67 18 0,979 36,90

15 SE (ANOS_FORMADO = alto) (DISSERT_ORIENT = baixo) (TESES_ORIENT = médio)
ENTÃO (ED_PUBLIC_LIVRO_INT = alto)
0,33 19 0,021 17,40

Segundo Experimento

Regras Descobertas no Segundo Experimento - Página ½

Regras Descobertas para IG's do Prof. Pavanelli

EXPERIMENTO: TODOS OS DADOS, Com atributo GRANDE_AREA, Metas são previsores.

CLASSE REGRA
Interesse IG FreqüênciaTotal Cobertura(SC+SE)

1 SE (GRANDE_AREA = 5) (NIVEL_FORMACAO = baixo)
ENTÃO (ARTIGOS_PER_NAC = baixo)
0,50 1 0,469 2,50

2 SE (ED_PUBLIC_LIVRO_NAC = alto)
ENTÃO (ARTIGOS_PER_NAC = médio)
0,50 20 0,506 26,00

3 SE (TESES_ORIENT = médio) (TRAB_GRAD_ORIENT = médio)
(ED_PUBLIC_LIVRO_NAC = alto)
ENTÃO (ARTIGOS_PER_NAC = alto)
0,33 20 0,025 1,75

4 SE (IDADE = alto)
ENTÃO (ARTIGOS_PER_INT = baixo)
0,50 15 0,642 550,40

5 SE (CAP_LIVROS_INT = alto)
ENTÃO (ARTIGOS_PER_INT = médio)
0,50 19 0,297 92,00

6 SE (ESCRITA_INGLES = B) (LIDER_GRUPO = 1) (GRANDE_AREA = 5)
(DISSERT_ORIENT = medio) (ED_PUBLIC_LIVRO_INT = alto)
ENTÃO (ARTIGOS_PER_INT = alto)
0,20 10 0,061 2,50

7 SE (GRANDE_AREA = 7) (NIVEL_FORMACAO = baixo)
ENTÃO (CAP_LIVROS_NAC = baixo)
0,50 8 0,769 2,50

8 SE (GRANDE_AREA = 7) (INIC_CIENT_ORIENT = alto)
ENTÃO (CAP_LIVROS_NAC = médio)
0,25 8 0,212 26,25

9 SE (UF = RS) (ANOS_FORMADO = alto) (TRAB_TECNICO = alto)
(ARTIGOS_PER_NAC = baixo)
ENTÃO (CAP_LIVROS_NAC = alto)
0,25 18 0,019 3,70

Continua na Página 2

Regras Descobertas no Segundo Experimento - Página 2/2

META	REGRA Interesse	IG	FreqüênciaTotal	Cobertura(SC+SE)
10	SE (ESCRITA_INGLES = P) (IDADE = médio) ENTÃO (CAP_LIVROS_INT = baixo)			
	0,50	14	0,932	910,40
11	SE (NACIONALIDADE = E) (GRANDE_AREA = 4) (CAP_LIVROS_NAC = baixo) ENTÃO (CAP_LIVROS_INT = alto)			
	0,33	6	0,068	1,00
12	SE (GRANDE_AREA = 3) ENTÃO (ED_PUBLIC_LIVRO_NAC = baixo)			
	1,00	5	0,835	735,00
13	SE (LIDER_GRUPO = 0) (CAP_LIVROS_NAC = alto) ENTÃO (ED_PUBLIC_LIVRO_NAC = alto)			
	0,50	12	0,165	53,00
14	SE (GRANDE_AREA = 2) ENTÃO (ED_PUBLIC_LIVRO_INT = baixo)			
	1,00	4	0,979	887,00
15	SE (GRANDE_AREA = 8) (ARTIGOS_PER_INT = baixo) (CAP_LIVROS_INT = alto) ENTÃO (ED_PUBLIC_LIVRO_INT = alto)			
	0,33	2	0,021	2,50

10.3 ANEXO C – Entrevistas para Avaliação das Regras

10.3.1 Entrevista 1 do Experimento 1

FORMULÁRIO A1

FORMULÁRIO A1 - Página 1/2

REGRAS COM COBERTURA BAIXA

AVALIAÇÃO DE REGRAS FEITA POR:
Prof. Dr. Flavio Bortolozzi
Pró-Reitor de Pesquisa e Pós-Graduação - PUCPR

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

1 SE CIDADE = Ijui (RS)
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

5 SE PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = alto (Bom)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = alto (>8)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = alto (>3)
E Nº_DE_INICIAÇÕES_CIENTÍFICAS_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = médio (1 a 5)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

6 SE ORIGEM (Continente onde nasceu) = América do Sul (menos Brasil)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = médio (1 a 4)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = alto (>5)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

Continua na Página 2

FORMULÁRIO A1 – Página 2/2

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

8 SE NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = médio (1 a 5)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

9 SE CIDADE = Florianópolis
E Nº_DE_TRABALHOS_TÉCNICOS = baixo (0)
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>5)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS INTERNACIONAIS

10 SE NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_MONOGRAFIAS_ORIENTADAS = alto (>4)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = baixo (0)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

11 SE NACIONALIDADE = Estrangeira
E CIDADE = Rio Grande (RS)
E PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = baixo (Pouco)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS NACIONAIS EDITADOS/PUBLICADOS

13 SE CIDADE = Porto Alegre (RS)
E Nº_DE_CURSOS_MINISTRADOS = médio (1 a 6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_MONOGRAFIAS_ORIENTADAS = alto (>4)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

FORMULÁRIO A2

FORMULÁRIO A2 – Página 1/2

REGRAS COM BOA COBERTURA

AVALIAÇÃO DE REGRAS FEITA POR:
Prof. Dr. Flavio Bortolozzi
Pró-Reitor de Pesquisa e Pós-Graduação - PUCPR

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

2 SE Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = alto (>3)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = médio (1 a 10)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

3 SE SEXO = Masculino
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = alto (>10)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

4 SE PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = baixo (Pouco)
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

7 SE NACIONALIDADE = Estrangeira
E NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

Continua na Página 2

FORMULÁRIO A2 – Página 2/2

REGRAS PREVENDO: Nº DE LIVROS NACIONAIS EDITADOS/PUBLICADOS

12 SE NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS INTERNACIONAIS EDITADOS/PUBLICADOS

14 SE CIDADE = Curitiba
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

15 SE Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = médio (1 a 4)
ENTÃO Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

10.3.2 Entrevista 2 do Experimento 1

FORMULÁRIO B1

FORMULÁRIO B1 – Página 1/2

REGRAS COM COBERTURA BAIXA

AVALIAÇÃO DE REGRAS FEITA POR :
Prof. Dr. Manoel Camillo Penna Neto
Coordenador de Pesquisa – PUCPR

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

1 SE CIDADE = Ijui (RS)
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

5 SE PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = alto (Bom)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = alto (>8)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = alto (>3)
E Nº_DE_INICIAÇÕES_CIENTÍFICAS_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = médio (1 a 5)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

6 SE ORIGEM (Continente onde nasceu) = América do Sul (menos Brasil)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = médio (1 a 4)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = alto (>5)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

Continua na Página 2

FORMULÁRIO B1 – Página 2/2

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

8 SE NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = médio (1 a 5)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

9 SE CIDADE = Florianópolis
E Nº_DE_TRABALHOS_TÉCNICOS = baixo (0)
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>5)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS INTERNACIONAIS

10 SE NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_MONOGRAFIAS_ORIENTADAS = alto (>4)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

11 SE NACIONALIDADE = Estrangeira
E CIDADE = Rio Grande (RS)
E PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = baixo (Pouco)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS NACIONAIS EDITADOS/PUBLICADOS

13 SE CIDADE = Porto Alegre (RS)
E Nº_DE_CURSOS_MINISTRADOS = médio (1 a 6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_MONOGRAFIAS_ORIENTADAS = alto (>4)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse

FORMULÁRIO B2

REGRAS COM BOA COBERTURA

AVALIAÇÃO DE REGRAS FEITA POR:
Prof. Dr. Manoel Camillo Penna Neto
Coordenador de Pesquisa – PUCPR

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

2 SE Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = alto (>3)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = médio (1 a 10)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

3 SE SEXO = Masculino
E Nº_DE_CURSOS_MINISTRADOS = alto (>6)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = alto (>10)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

4 SE PROFICIÊNCIA_DE_ESCRITA_EM_INGLES = baixo (Pouco)
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

7 SE NACIONALIDADE = Estrangeira
E NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = baixo (0)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

Continua na Página 2

FORMULÁRIO B2 - Página 2/2

REGRAS PREVENDO: Nº DE LIVROS NACIONAIS EDITADOS/PUBLICADOS

12 SE NIVEL_FORMAÇÃO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS INTERNACIONAIS EDITADOS/PUBLICADOS

14 SE CIDADE = Curitiba
E NIVEL_FORMACAO = alto (doutorado, pós-dout. ou livre doc)
E Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
ENTÃO Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

15 SE Nº_DE_ANOS_DESDE_ÚLTIMA_TITULAÇÃO = alto (>20)
E Nº_DE_DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = baixo (0)
E Nº_DE_TESES_DE_DOUTORADO_ORIENTADAS = médio (1 a 4)
ENTÃO Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

10.3.3 Entrevista do Experimento 2

FORMULÁRIO C1

FORMULÁRIO C1 - Página 1/2

REGRAS COM COBERTURA BAIXA

AVALIAÇÃO DE REGRAS FEITA POR:
Prof. Dr. Gilberto Cezar Pavanelli
Pró-Reitor de Pesquisa e Pós-Graduação – UEM.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

IG[1]: SE GRANDE_AREA = 5 (Ciências Agrárias)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = alto (>10)

1 SE GRANDE_AREA = Ciências Agrárias
E NIVEL_FORMACAO = baixo (Graduação ou Aperfeiçoamento/Especialização)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (**X**) Alto Interesse.

IG[20]: SE Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

3 SE TESES_ORIENTADAS = médio (1 a 4)
E TRABALHOS_DE_GRADUAÇÃO_ORIENTADOS = médio (1 a 5)
E Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = alto (>10)

( ) Baixo Interesse; (**X**) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

IG[10]: SE GRANDE_AREA = 5 (Ciências Agrárias)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = baixo (0)

6 SE ESCRITA_INGLES = alto (Bom)
E LIDER_DE_GRUPO
E GRANDE_AREA = Ciências Agrárias
E DISSERTAÇÕES_DE_MESTRADO_ORIENTADAS = médio (1 a 8)
E Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = alto (>5)

(**X**) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

Continua na Página 02

FORMULÁRIO C1 – Página 2/2

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

IG[8]: SE GRANDE_AREA = 7 (Ciências Humanas)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>5)

7 SE GRANDE_AREA = 7 (Ciências Humanas)
E NIVEL_FORMACAO = baixo (Graduação ou Especialização)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = baixo (0)
( ) Baixo Interesse; ( ) Interesse Médio; (**X**) Alto Interesse.

IG[18]: SE Nº_DE_ARTIGOS_PERIÓDICOS_NACIONAIS = baixo (0)
ENTÃO CAPITULOS_DE_LIVROS_NACIONAIS = baixo (0)

9 SE UF = RS
E ANOS_FORMADO = alto (>20)
E TRAB_TECNICO = alto (>7)
E Nº_DE_ARTIGOS_PERIÓDICOS_NACIONAIS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>0)
(**X**) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS INTERNACIONAIS

IG[6]: SE GRANDE_AREA = 4 (Ciências da Saúde)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = baixo (0)

11 SE NACIONALIDADE = Estrangeira
E GRANDE_AREA = 4 (Ciências da Saúde)
E Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = baixo (0)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)
( ) Baixo Interesse; (**X**) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS INTERNACIONAIS EDITADOS/PUBLICADOS

IG[2]: SE GRANDE_AREA = 8 (Lingüística, Letras e Artes)
ENTÃO Nº_DE_LIVROS_INTERNAC_EDITADOS/PUBLICADOS = baixo (0)

15 SE GRANDE_AREA = 8 (Lingüística, Letras e Artes)
E Nº_DE_ARTIGOS_PERIÓDICOS_INTERNACIONAIS = baixo (0)
E Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)
ENTÃO Nº_DE_LIVROS_INTERNACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
(**X**) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

FORMULÁRIO C2

FORMULÁRIO C2 - Página 1/2

REGRAS COM BOA COBERTURA

AVALIAÇÃO DE REGRAS FEITA POR:
Prof. Dr. Gilberto Cezar Pavanelli
Pró-Reitor de Pesquisa e Pós-Graduação – UEM.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS NACIONAIS

IG[20]: SE Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = baixo (0)

2 SE Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)
ENTÃO NÚMERO_DE_ARTIGOS_EM_PERIÓDICOS_NACIONAIS = médio (1 a 10)

( ) Baixo Interesse; (**X**) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE ARTIGOS EM PERIÓDICOS INTERNACIONAIS

IG[15]: SE IDADE = alto (>57)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = médio (1 a 5)

4 SE IDADE = alto (>57)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (**X**) Alto Interesse.

IG[19]: SE Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = alto (>5)

5 SE Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)
ENTÃO Nº_DE_ARTIGOS_EM_PERIÓDICOS_INTERNACIONAIS = médio (1 a 5)

( ) Baixo Interesse; (**X**) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS NACIONAIS

IG[8]: SE GRANDE_AREA = 7 (Ciências Humanas)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>5)

8 SE GRANDE_AREA = 7 (Ciências Humanas)
E INICIAÇÃO_CIENTÍFICA_ORIENTADAS = alto (>5)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = médio (1 a 5)

(**X**) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

Continua na Página 2

FORMULÁRIO C2 – Página 2/2

REGRAS PREVENDO: Nº DE CAPÍTULOS DE LIVROS INTERNACIONAIS

IG[14]: SE IDADE = médio (34 a 57)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = alto (>0)

10 SE ESCRITA_INGLÊS = baixo (Pouco)
E IDADE = médio (34 a 57)
ENTÃO Nº_DE_CAPÍTULOS_DE_LIVROS_INTERNACIONAIS = baixo (0)

( ) Baixo Interesse; (**X**) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS NACIONAIS EDITADOS/PUBLICADOS

IG[5]: SE GRANDE_AREA = 3 (Engenharias)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

12 SE GRANDE_AREA = 3 (Engenharias)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (**X**) Alto Interesse.

IG[12]: SE NÃO_LIDER_DE_GRUPO
ENTÃO Nº_DE_LIVROS_NACION_EDITADOS/PUBLICADOS = baixo (0)

13 SE NÃO_LIDER_DE_GRUPO
E Nº_DE_CAPÍTULOS_DE_LIVROS_NACIONAIS = alto (>5)
ENTÃO Nº_DE_LIVROS_NACIONAIS_EDITADOS/PUBLICADOS = alto (>0)

(**X**) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

REGRAS PREVENDO: Nº DE LIVROS INTERNACIONAIS EDITADOS/PUBLICADOS

IG[4]: SE GRANDE_AREA = 2 (Ciências Biológicas)
ENTÃO Nº_DE_LIVROS_INTERNAC_EDITADOS/PUBLICADOS = alto (>0)

14 SE GRANDE_AREA = 2 (Ciências Biológicas)
ENTÃO Nº_DE_LIVROS_INTERNAC_EDITADOS/PUBLICADOS = baixo (0)

( ) Baixo Interesse; ( ) Interesse Médio; (**X**) Alto Interesse.