UNIVERSIDADE FEDERAL DE SANTA CATARINA

CENTRO TECNOLÓGICO

DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Integração de uma Etapa de Análise Exploratória de Dados em um Sistema de Análise de Risco para Operações de Completação e Perfuração de Poços de Petróleo

Dyego Wrubel Santin

Florianópolis

2007

Dyego Wrubel Santin

Integração de uma Etapa de Análise Exploratória de Dados em um Sistema de Análise de Risco para Operações de Completação e Perfuração de Poços de Petróleo

Proposta de Projeto para o Trabalho de Conclusão de Curso em Ciências da Computação da Universidade Federal de Santa Catarina.

Orientador:
Mauro Roisenberg, Dr.

Universidade Federal de Santa Catarina
Centro Tecnológico
Departamento de Informática e Estatística

Florianópolis

2007

Monografia de Projeto Final de Graduação sob o título *"Integração de uma Etapa de Análise Exploratória de Dados em um Sistema de Análise de Risco para Operações de Completação e Perfuração de Poços de Petróleo"*, defendida por Dyego Wrubel Santin e aprovada em 2007, em Florianópolis, Estado de Santa Catarina, pela banca examinadora constituída pelos professores:

Prof. Mauro Roisenberg, Dr.
Orientador

Prof. Dalton Francisco de Andrade, Dr.
Universidade Federal de Santa Catarina

Profa. Silvia Modesto Nassar, Dra.
Universidade Federal de Santa Catarina

Agradecimentos

Aos meus queridos pais Domingos Santin e Albany Santin.

Em especial, ao Projeto E&P Risk e toda equipe pela grande experiência.

Ao Professor Dr. Mauro Roisenberg pela oportunidade.

Aos Professores Dr. Dalton Francisco de Andrade e Dra. Silvia Modesto Nassar pelo apoio.

Aos grandes companheiros e amigos da turma 022 que de alguma forma contribuiram para a realização deste trabalho.

Sumário

Lista de Figuras

1 Introdução

A análise de riscos e o gerenciamento em processos de exploração de petróleo, atualmente são áreas em crescente expansão. As companhias petrolíferas estão usufruindo de princípios de análise de riscos em combinação com novas tecnologias, para aumentar sua capacidade de exploração.

As operações de perfuração e completação de poços de petróleo são muito complexas, com um alto grau de imprecisão e vários fatores de riscos associados. Os maiores gastos estão relacionados com operações de sondas para intervenções em poços e manutenções corretivas. Para se ter uma idéia, os custos de extração no terceiro trimestre de 2006 atingiram 6,64US$/bbl (Petrobras).

Portanto, a complexidade destas operações é um dos grandes responsáveis pelos altos custos de produção. Desta maneira, surge a seguinte questão: **É possível construir um sistema computacional de simulação e análise de risco para estas operações? E além disso, capaz de fazer uma estimativa do tempo total para realizar uma determinada operação?**

Atualmente, muitas técnicas de Inteligência Computacional são empregadas com sucesso no desenvolvimento de sistemas inteligentes de previsão, suporte à decisão, controle, otimização, modelagem, classificação e reconhecimento de padrões em geral, aplicados em diversos setores: energético, industrial, econômico, financeiro, comercial, entre outros.

Diante disso, a solução proposta para as questões acima chama-se E&P Risk. O E&P Risk é um projeto desenvolvido pelos integrantes do laboratório PerformanceLab da UFSC (Universidade Federal de Santa Catarina) em parceria com o centro de pesquisa da Petrobras/Cenpes.

O E&P Risk está dividido em vários módulos:

- Módulo Monte Carlo;
- Módulo Bayesiano;

- Módulo de Teste de Aderência;
- **Módulo Neural.**

Diante deste contexto, o principal objetivo dessa proposta é apresentar novas abordagens e técnicas de análise exploratória de dados, que possam ser empregadas no E&P Risk, mais especificamente, no módulo neural. O Módulo neural foi introduzido na versão III do E&P Risk e teve continuação na versão IV.

1.1 Tema

A grande imprevisibilidade presente nas operações de perfuração e completação de poços de petróleo, devido a uma série de riscos e fatores, tais como o conhecimento limitado sobre as características geológicas, as dificuldades técnicas e ainda, o comportamento imprevisível de operadores humanos, torna imprescindível o uso de alguma técnica ou ferramenta para auxiliar na estimativa do tempo total gasto com essas operações.

1.2 Delimitação do Tema

A proposta é utilizar técnicas de análise exploratória de dados com a finalidade de auxiliar na análise de riscos em operações de perfuração e completação de poços de petróleo, visto que, estas técnicas vêm desempenhando um papel fundamental na descoberta de novos conhecimentos e informações válidas, implícitas em bases de dados.

1.3 Objetivo Geral

Desenvolver um novo submódulo de **Análise Exploratória** dos dados para integrar o processo de descoberta do tempo total, conforme pode ser visto na figura 1:

A justificativa é que um dos pontos críticos para os modelos de redes neurais é a qualidade dos dados de entrada, ou seja, *quanto maior a qualidade dos dados mais confiável será o resultado.*

Além disso, o módulo neural no E&P Risk III tava limitado a uma série de restrições impostas pela maneira como os dados eram importados.

1. Exige-se que o arquivo de entrada para os dados esteja no formato texto com a extensão (*.txt);

2. O caractere delimitador para os campos dos registros deve ser o ponto-e-vírgula (;);

3. O identificador dos campos de strings são as aspas duplas (“campo”);

4. Todo registro que possuir um campo não preenchido será descartado, não possibilitando ao usuário nenhum tipo de tratamento para estes campos.

A figura 1 ilustra todo o processo para a estimativa do tempo total no módulo neural proposto para o E&P Risk IV, que vai desde a importação dos dados que servirão de entrada para a etapa de análise exploratória, que por sua vez irá alimentar o algoritmo minerador que vai fazer a estimativa do tempo total.

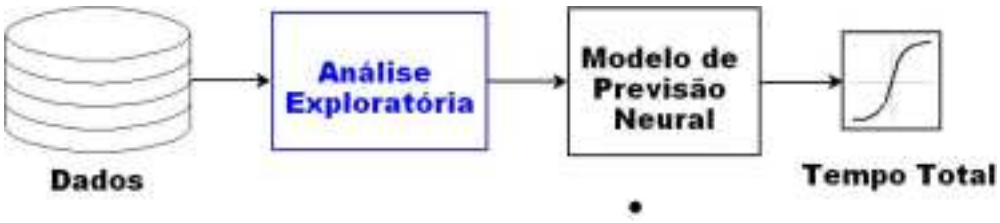

Figura 1: Etapa para preparação dos dados no módulo neural do E&P Risk IV

1.4 Objetivos Específicos

- Desenvolver uma interface gráfica que permita ao usuário ter uma visão geral dos dados, através de estatísticas gerais e gráficos de barras e além disso, interagir com ferramentas de manipulação e preparação dos dados;

- Implantar uma nova etapa de análise exploratória e preparação dos dados no processo de estimativa do tempo total realizado pelo modelo de previsão neural;

- Implementar ferramentas para a realização de tarefas de limpeza dos dados, assim como ferramentas para auxiliar a identificação de ruídos, valores corrompidos, inconsistências e redundâncias;

- Permitir a manipulação dos dados através de ferramentas de edição e de visualização gráfica dos dados;

- Permitir a seleção determinística dos atributos para que o usuário possa escolher apenas os atributos relevantes ao domínio em questão.
- Desenvolver uma ferramenta para auxiliar na tarefa de tratamento dos campos não preenchidos ("missings") dos dados.
- Utilizar e compreender ferramentas de consultas a base de dados, tais como a linguagem SQL.

1.5 Motivação

O crescente aumento das bases de dados devido à grande quantidade de informações gerada nos dias atuais, está tornando cada vez mais importante, o uso de ferramentas e técnicas automatizadas para extrair conhecimento dessas grandes coleções de dados. Dentre as técnicas mais utilizadas, temos a análise exploratória dos dados.

Além disso, é muito importante estimar o tempo total de uma operação de perfuração de um poço de petróleo, principalmente, devido ao alto custo deste tipo de operação.

1.6 Estrutura do Trabalho

O Trabalho está definido da seguinte maneira:

O capítulo 2 faz uma abordagem do processo de descoberta de conhecimento KDD, assim como o esquema que foi definido para o E&P Risk IV. O capítulo 3 diz respeito ao E&P Risk IV propriamente dito e o funcionamento do "wizard" de importação de dados. O capítulo 4 traz as técnicas de preparação de dados e análise exploratória empregadas para as mais diversas tarefas. O capítulo 5 trata a metologia de trabalho adotada assim como detalhes de implementação. Finamente, o capítulo 6 é dedicado às conclusões e trabalhos futuros do projeto.

2 Descoberta de Conhecimento - KDD

A área de Descoberta de Conhecimento em Banco de Dados (KDD - Knowledge Discovery in Databases) surgiu da necessidade de técnicas e soluções para análise de grandes bases de dados. O KDD é um processo amplo, que inclui várias etapas. Estas etapas podem ser classificadas de maneira geral como: Análise Exploratória de Dados e Mineração de Dados.

O objetivo da **análise exploratória** de dados é examinar a estrutura subjacente dos dados e aprender sobre os relacionamentos sistemáticos entre muitas variáveis. A análise exploratória de dados inclui um conjunto de ferramentas gráficas e descritivas, para explorar os dados, como pré-requisito para uma análise de dados mais formal, e como parte da construção de modelos.

Na etapa de **Mineração de Dados** (Data Mining) um conjunto de técnicas são empregadas para a extração de padrões de interesse e para a criação de modelos para estimação e predição de valores. Já entre as tarefas da etapa de Mineração, podemos encontrar técnicas de Análise de Agrupamentos, Análise de Regressão, Avaliação e Visualização, entre outras.

Mais detalhadamente, as etapas do processo completo de KDD são as seguintes (SILVA, 2004):

1. Limpeza dos dados: etapa para a eliminição de ruídos, dados inconsistentes ou corrompidos. Além disso, correção de possíveis erros e imputação ou eliminição de valores nulos e redundantes.

2. Integração dos dados: etapa de combinação de diferentes fontes de dados para a formação de um único repositório de dados.

3. Seleção: etapa para a seleção dos atributos que são relevantes para o objetivo da

tarefa. Por exemplo, o usuário pode decidir que informações como “Nome do poço” não são importantes para determinar o tempo total de perfuração de um novo poço.

4. Transformação dos dados: etapa para a transformação dos dados em um formato apropriado para a execução dos algoritmos de mineração.

5. Mineração: etapa essencial no processo de KDD, consiste na aplicação de técnicas inteligentes com o propósito de extrair os padrões e informações de interesse. Pode-se dizer que transforma dados em informações.

6. Avaliação ou Pós-processamento: etapa para identificação dos melhores padrões de acordo com algum critério adotado, ou seja, é um processo de refinamento dos resultados.

7. Visualização dos Resultados: etapa para a representação e interpretação do conhecimento minerado.

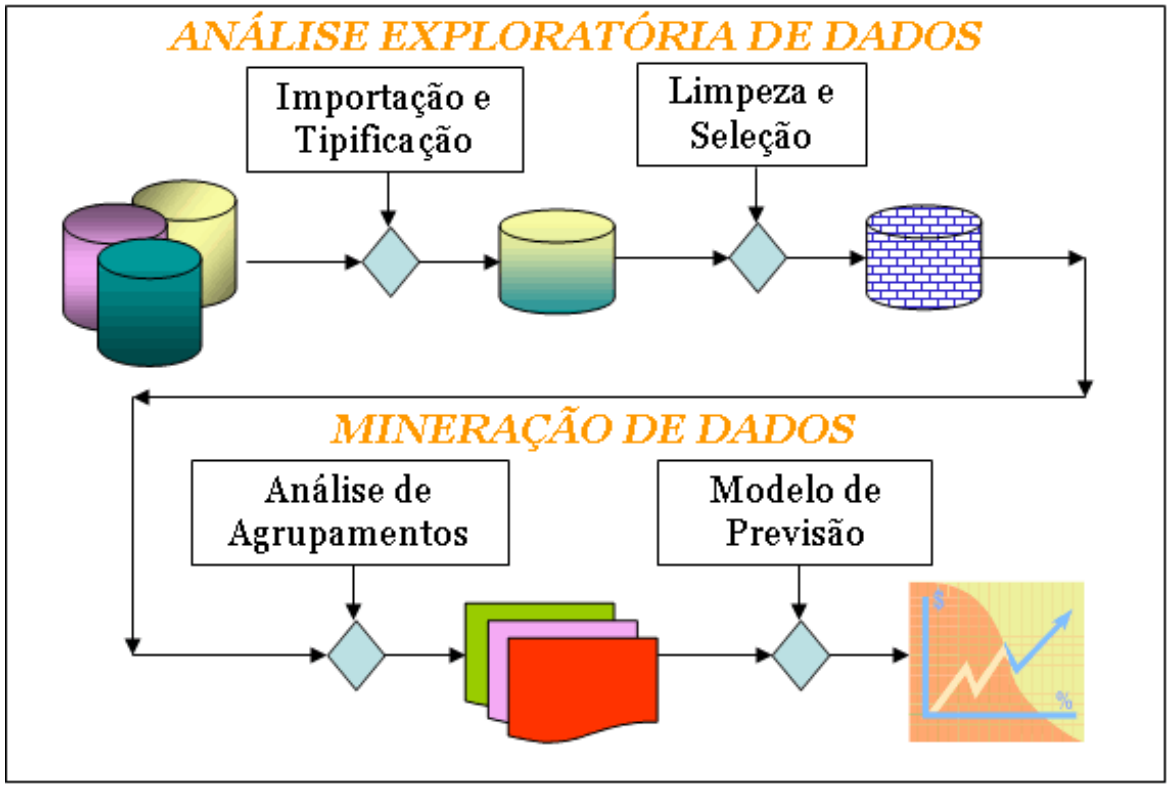

Figura 2: Etapas do Processo de Criação do Modelo de Previsão

2.1 Fundamentação da Tipificação

2.1.1 Tipificação

O processo de tipificação automático de variáveis implementado no E&P Risk IV, é uma ferramenta para auxiliar o usuário na definição dos tipos das variáveis. Definir os tipos das variáveis pode ficar complicado para um usuário não técnico que não tem conhecimento da semântica das mesmas, ou seja, não é um especialista do domínio. Além disso, o tipo da variável é uma informação importante para o submódulo neural do sistema.

O sistema proposto é uma adaptação e simplificação do processo de tipificação implementado no SETIP. (SANTOS, 2001) De acordo com a natureza da variável, o seu tipo pode ser classificado dentre as seguintes categorias: qualitativa, quantitativa, booleana ou do tipo data.

Para as variáveis do tipo qualitativa, booleana e data, a identificação é simples, uma vez que esta informação pode ser descoberta através dos próprios valores e formatos dos dados. No entanto, a complexidade aumenta quando os valores são numéricos, pois surge a seguinte questão:

*Os valores **numéricos** da variável possuem uma semântica de qualidades da variável, ou seja, a variável é qualitativa?*

Por exemplo, uma determinada variável qualitativa, "Grau de Dificuldade de Perfuração", pode representar uma escala da seguinte maneira:

- 1 para representar um grau alto;
- 0,5 para representar um grau médio;
- 0 para represe ntar um grau baixo;

A proposta é baseada na busca dessa informação em uma base de conhecimento de casos através da **política do vizinho mais próximo**, ou seja, encontra-se por similaridade o caso da base mais "parecido" com o novo caso indefinido. Dessa maneira, o tipo do novo caso é determinado de acordo com o tipo do mais similar a este novo caso. Este esquema pode ser visto na figura 3.

Um caso é definido através dos seguintes parâmetros (SANTOS, 2001):

- Média dos valores absolutos elevados ao expoente -1, para $X_i \neq 0$, com $i = 1, 2, ..., n$.

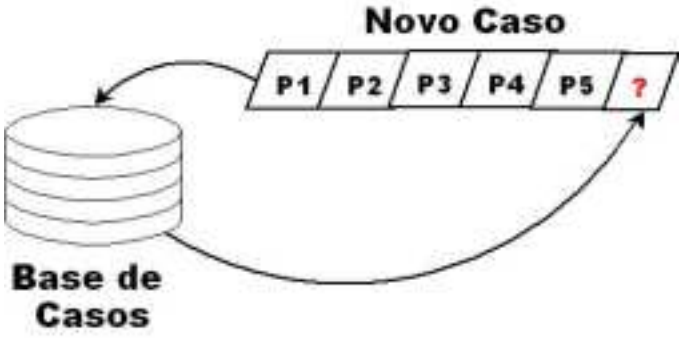

Figura 3: Diagrama do Esquema de Tipificação

Onde X_i representa os valores válidos e n a quantidade de valores válidos.

$$MedExpo = \frac{\sum_{i=1}^{n} |X_i|^{-1}}{n} \tag{2.1}$$

- Razão entre a quantidade de valores que se repetem nos dados R e o número de valores válidos n.

$$RazaoRep = \frac{R}{n} \tag{2.2}$$

- Média ponderada adaptada dos valores válidos para $max \neq min$.

$$MedPond = \frac{\varrho^{-\left(\frac{\sum_{i=1}^{n} \frac{1}{2X_i}}{n}\right)}}{(max - min)} \tag{2.3}$$

- Soma (dividida de forma a obter um valor entre zero e um) dos valores, na suposição de formarem uma Progressão Aritmética (PA), para $max + min \neq 0$ e $Dif \neq 0$, onde Dif representa o número de diferentes valores que os dados possuem.

$$SomaPA = \frac{2}{(max + min)\, Dif} \tag{2.4}$$

- Razão (dividida de forma a obter um valor entre zero e um) dos valores, na suposição de formarem uma Progressão Aritmética (PA), para $max + min \neq 0$ e $Dif \neq 0$, onde $DDifi$ representa o número de valores que não se repetem nos dados.

$$RazaoPA = \frac{2\left(\sum_{i=1}^{n} \left(DDif_{i+1} - DDif_i\right)\right)}{(max + min)\, Dif} \tag{2.5}$$

A base de casos do esquema é ilustrada na tabela abaixo:

Caso	$MedExpo$	$RazaoRep$	$MedPond$	$SomaPA$	$RazaoPA$	Tipo
0	0,16273	0,31111	0,00741	0,02144	0,12500	0,0
1	0,83706	0,21429	0,03571	0,10967	0,95333	1,0
2	0,46578	0,25000	0,08333	0,19806	0,96667	1,0
3	0,05490	0,03445	0,00055	0,01544	0,64000	0,0
4	0,44000	0,66667	0,33333	0,40126	0,98000	1,0

2.2 Trabalhos Relacionados

2.2.1 Weka

O Weka (THE UNIVERSITY OF WAIKATO,) é um software de mineração de dados desenvolvido na linguagem de programação Java (Sun), com código aberto e licença GPL (General Public License), que apresenta uma coleção de algoritmos de aprendizado para a realização de tarefas de mineração de dados. Possui ferramentas para pré-processamento, classificação, regressão, clusterização, regras de associação e visualização dos dados.

2.2.2 SETIP - Sistema Especialista para Tipificar Dados de uma Pesquisa

O SETIP (SANTOS, 2001) é um sistema especialista integrado ao SEstat (sistema utilizado no ensino da estatística), que classifica variáveis de dados quanto ao seu tipo, qualitativa ou quantitativa.

3 E&P Risk IV

3.1 Introdução

O Sistema E&P Risk IV é um ambiente computacional voltado para a simulação e análise de risco em operações de perfuração e completação de poços de petróleo. Foi desenvolvido pelos integrantes dos laboratórios da UFSC (Universidade Federal de Santa Catarina), PerformanceLab, LEA e L3C sob encomenda da Petrobras/Cenpes.

3.1.1 Análise Exploratória dos Dados

A etapa de Análise Exploratória de Dados inclui tarefas que manipulam e preparam os dados para a etapa de mineração, tais como: limpeza, tipificação, integração dos dados, seleção de atributos e transformação de dados.

3.1.2 Mineração dos Dados

No caso do Sistema E&P Risk IV, o Sub-Módulo de Redes Neurais Artificiais provê um conjunto de técnicas e ferramentas destinadas a construir modelos para previsão de valores de tempos de intervenção em operações complexas de engenharia de poços a partir de informações históricas armazenadas em bancos de dados da companhia.

3.2 Importação e Esquema de Organização dos Dados

3.2.1 Organização dos Dados

Os dados são organizados e mantidos em uma estrutura interna de armazenamento, desta maneira, qualquer alteração que seja aplicada aos dados da estrutura, não será propagada aos dados originais na fonte.

3.2.2 Fontes de Dados Suportadas

O E&P Risk IV possui uma fonte de dados bastante diversificada e flexível, suportando os seguintes formatos:

3.2.2.1 Arquivos no formato texto com a extensão "txt" ou "csv"

Este formato é uma representação tabular dos dados. Portanto, o arquivo possui um conjunto de "n" linhas (registros), sendo cada registro composto por "m" colunas (valores dos campos) separados por um caractere delimitador. Diante disso, o número de variáveis de dados é determinado pelo número de colunas "m", ou seja, cada coluna representa uma variável distinta.

Além disso, a primeira linha pode ser um registro especial que contém os nomes das variáveis. Caso contrário, as variáveis assumirão nomes padrões definidos automaticamente pelo sistema. As demais linhas possuem apenas dados.

Exemplos de caracteres delimitadores que podem ser utilizados: o ponto-e-vírgula, a vírgula, caracteres de espaçamento (tabulação ou espaço em branco) e ainda, qualquer outro caractere que possa ser determinado pelo usuário.

Observações: É necessário a integridade do arquivo, ou seja, que todos os registros possuam o mesmo número de campos delimitados. Um campo **"missing"** será identificado pela ocorrência de dois delimitadores consecutivos. Além disso, é necessário a correspondência entre o caractere delimitador do arquivo com o indicado pelo usuário durante a etapa de importação dos dados.

3.2.2.2 Dados via provedor OLE DB

Neste caso, o acesso aos dados será feito através de um provedor OLE DB especificado pelo usuário

O que é OLE DB?

OLE DB extende por *Object Linking and Embedding for Databases*. Tecnologia desenvolvida pela Microsoft e usada para se ter acesso a diferentes fontes de informações ou bases de dados, de maneira uniforme. É dividida em consumidores e provedores. Os consumidores são os aplicativos que necessitam do acesso aos dados e os provedores são os componentes de software que fornecem uma interface para o acesso.

Especificando o provedor OLE DB no E&P Risk IV

Durante a etapa de importação, é usada a aba "Provider" do "Data Link Properties" para selecionar o provedor OLE DB apropriado para a base de dados, dentre uma lista de todos os provedores detectados no computador.

Por exemplo, as bases de dados do Microsoft Access (*.mdb), podem ser acessadas através do provedor Microsoft Jet 4.0.

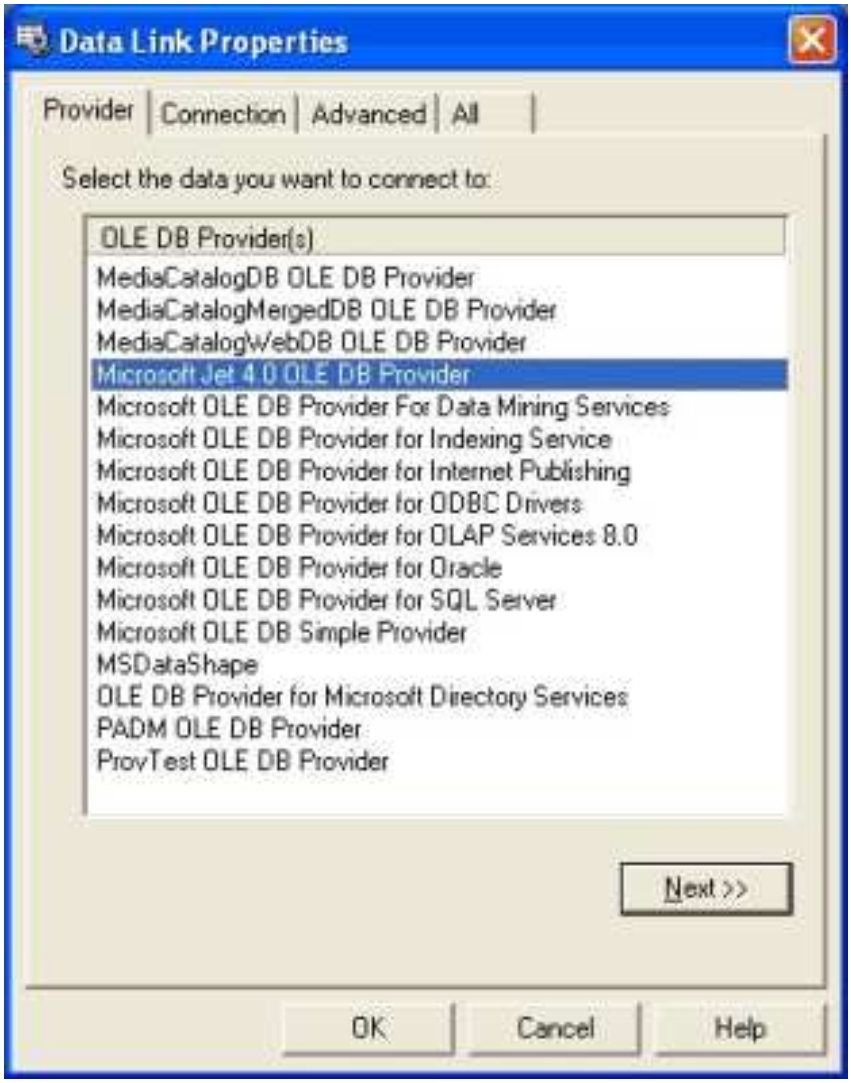

Figura 4: Selecionando o provedor OLE DB

Especificando a conexão para o acesso

É na aba "Connection" que são definidas as propriedades de conexão requeridas para fazer o acesso a um determinado provedor. A conexão pode ser testada para verificar as propriedades.

Figura 5: Propriedades da conexão

3.2.2.3 Microsoft Excel 97/2000/2002

Esta opção espera um arquivo com a extensão "xls" contendo internamente uma ou mais planilhas. No último caso, o usuário selecionará a planilha desejada para a importação dos seus dados.

3.2.3 Configurações Regionais

O E&P Risk IV adota as configurações regionais definidas no sistema operacional para os padrões numéricos, como o separador de decimal, milhar, entre outros.

Deste modo, os dados originais usados para a importação deverão obedecer o esquema de configuração local do sistema. Por exemplo, caso o sistema esteja configurado para

usar as configurações regionais do Brasil, assumi-se como separador de decimal a vírgula (“,”) e como separador de milhar o ponto (“.”).

Além disso, deve-se descartar a possibilidade de usar o mesmo caractere, por exemplo a vírgula, tanto para separador de decimal como para delimitador entre os valores dos dados.

3.2.4 Importação dos Dados

A importação dos dados é realizada em forma de um “Wizard”. Deste modo, uma sequência pré-determinada de passos será executada de acordo com a opção de fonte de dados escolhida pelo usuário.

A seguir é apresentado os passos de um procedimento padrão para importação de dados à partir de um aquivo texto com a extensão CSV.

3.2.4.1 Importando os dados de um arquivo CSV

1. O primeiro passo é selecionar o tipo da fonte dos dados. Figura(6).

Figura 6: Selecionado a fonte dos dados

2. O passo seguinte é determinar o nome do arquivo que contém os dados, o caractere delimitador dos campos dos registros e ainda, indicar se o primeiro registro contém ou não os nomes dos atributos (Figura 7).

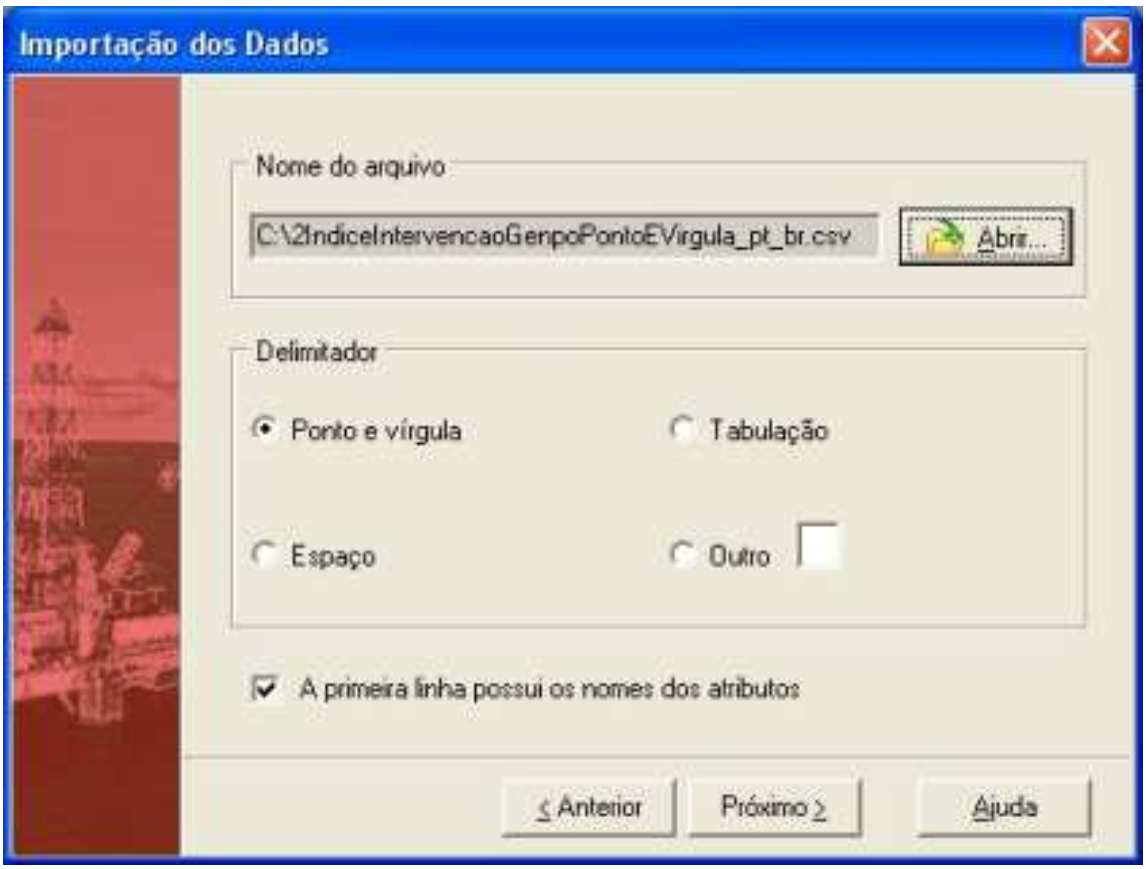

Figura 7: Determinando o nome do arquivo, o delimitador e se possui os nomes dos atributos

3. Após isso, deve-se fazer a seleção dos atributos que serão importados para a análise exploratória. (Figura 8).

   O avanço no procedimento de importação, exatamente neste ponto, implicará na execução do processo de tipificação automático das variáveis de dados que estão sendo obtidas para a etapa de análise.

   Visto que a tipificação é apenas uma ferramenta para auxiliar o especialista a determinar os tipos para as suas variáveis de entrada, o próximo passo consiste na visualização dos resultados da tipificação e ainda, possível alteração e correção para os tipos determinados.

4. Finalmente, o último passo é a visualização e edição dos tipos dos atributos. Figura(9).

Figura 8: Selecionado os atributos

Figura 9: Visualizando e editando os tipos dos atributos

4 Análise Exploratória dos Dados

A etapa que sucede a importação dos dados do sistema é chamada de **"análise exploratória dos dados"**. Nesta etapa, uma série de informações são gerenciadas através de uma interface para vizualização de estatísticas gerais, atributos, tipos dos atributos, gráficos, gráficos de barras e ainda, ferramentas para limpeza, seleção, transformação e edição dos dados.

As informações estão distribuídas em diferentes perspectivas da interface gráfica, conforme pode ser visto na figura 10:

1. **Informações Gerais:** exibe o nome da fonte dos dados, o total de registros lidos da base e o total de atributos.
2. **Atributos:** exibe uma lista com os nomes dos atributos e uma caixa de seleção ao lado de cada nome. Acima da lista existem três botões:
   - **Todos:** marca a caixa de seleção de todos os atributos.
   - **Remover:** remove os atributos que possuem a caixa de seleção marcada.
   - **Inverter:** marca a caixa de seleção que está desmarcada e vice-versa.
3. **Tipos:** exibe os tipos dos atributos previamente tipificados entre as quatro possíveis classes: quantitativos, qualitativos, tipo booleano e data.
4. **Atributo Selecionado (estatísticas):** apresenta algumas estatísticas de um determinado atributo selecionado. Dentre as estatísticas destacam-se:
   - **Distintos:** número de diferentes valores que os dados possuem para o atributo selecionado.
   - **Únicos:** número de valores do atributo que possuem uma única instância nos dados.

- **Missing:** número de registros dos dados para qual o valor do atributo não está preenchido.

Abaixo destas estatísticas, é exibida uma lista com informações sobre os valores do atributo de acordo com o seu tipo. Para um atributo quantitativo, a lista exibirá quatro novas estatísticas descrevendo a distribuição dos dados, sendo estas: o valor mínimo, máximo, média e o desvio padrão. Para os demais tipos, a lista exibirá cada possível valor do atributo com a sua correspondente frequência, ou seja, o número de vezes que o valor aparece nos registros.

5. **Gráficos de Barras (visualização primária):** apresenta o gráfico de barras de freqüência do atributo selecionado. Acima do gráfico de barras tem-se um componente "combobox" com os nomes dos atributos que podem ser selecionados. Ao seu lado direito, encontram-se informações (classe, freqüência) que serão exibidas quando o usuário clicar em uma barra do gráfico de barras.

Além disso, no canto inferior esquerdo da tela existem quatro outras funcionalidades:

1. **Salvar:** salva os dados em um arquivo texto (*.txt). O caractere delimitador de valores poderá ser o ponto-e-vírgula, ou então, o caractere de tabulação.
2. **Ir para o treinamento:** Passa para a etapa de treinamento da rede neural se não existirem valores missigns para os dados, caso contrário, o tratamento dos mesmos ainda deverá ser feito.
3. **Edição:** Exibe a janela de edição dos dados.
4. **Visualização:** Exibe a janela de vizualização dos dados.

4.1 Edição dos Dados

A tabela de edição de dados, que pode ser vista na Figura 11, é o mecanismo que o usuário possui para realizar algum tipo de tratamento ou transformação nos dados.

Ao clicar com o botão direito do mouse sobre a tabela de edição, o usuário acessa um "popup menu"com as principais funcionalidades para realizar a tarefa de edição dos dados, conforme pode ser visto na figura 11.

Dentre as principais funcionalidades da etapa de edição destacam-se:

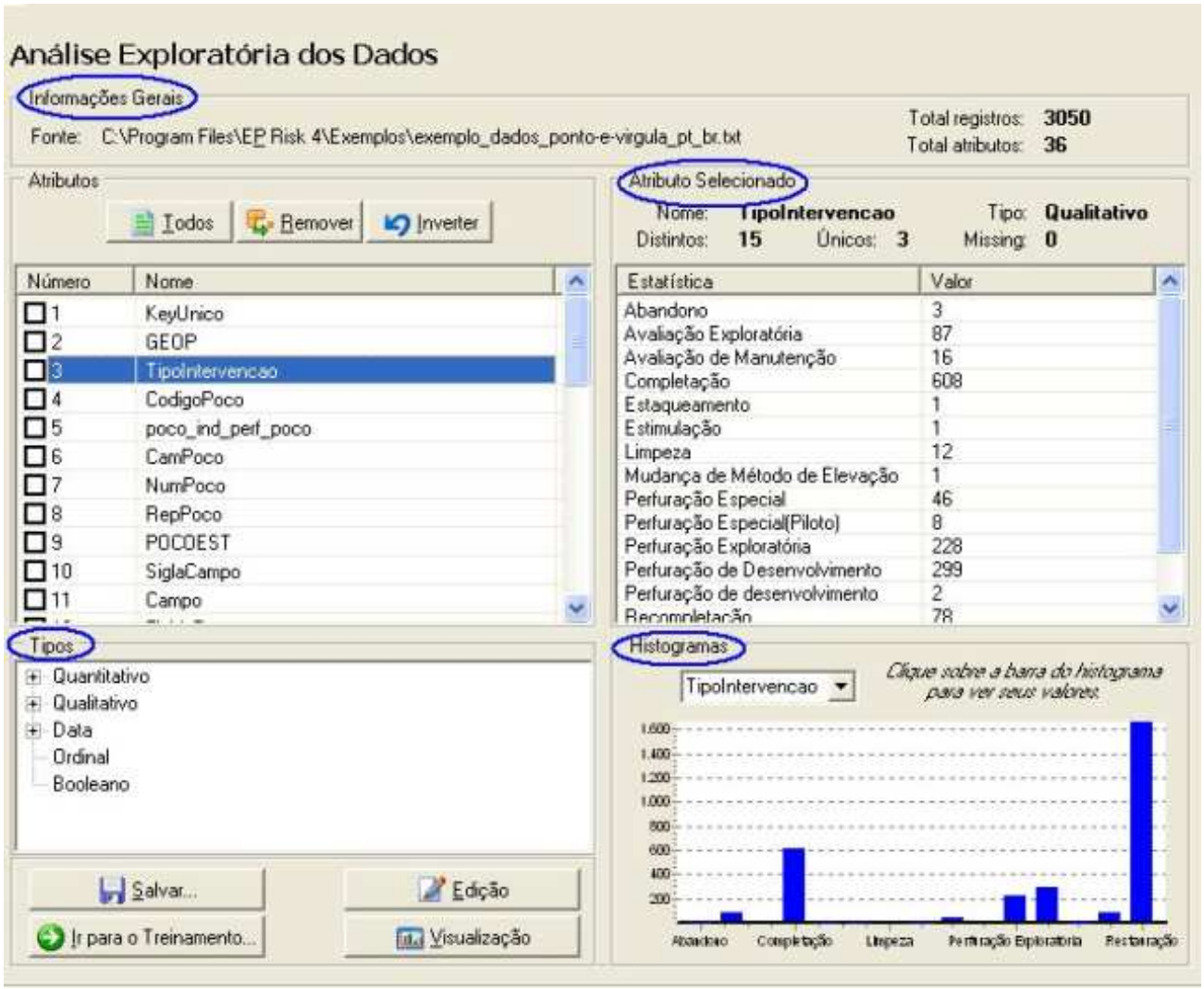

Figura 10: Análise Exploratória

1. **Renomear Atributo:** utilizada para renomear o nome do atributo.

2. **Substituir:** substituir valores dos atributos por novos valores.

   - Permitir a substituição de **todos** os valores do atributo por um novo valor escolhido. Para um atributo numérico por um novo valor qualquer (Figura 12). Para um atributo nominal por um dos valores nominais já existentes.
   - Permitir a substituição de todos os valores **"missings"** por um novo valor especificado. Para um atributo numérico o novo valor pode ser zero, o valor médio, ou um número qualquer. Para um atributo nominal, um dos valores nominais já existentes (Figura 13).

3. **Missings:** realizar o tratamento dos valores não preenchidos ("missings").

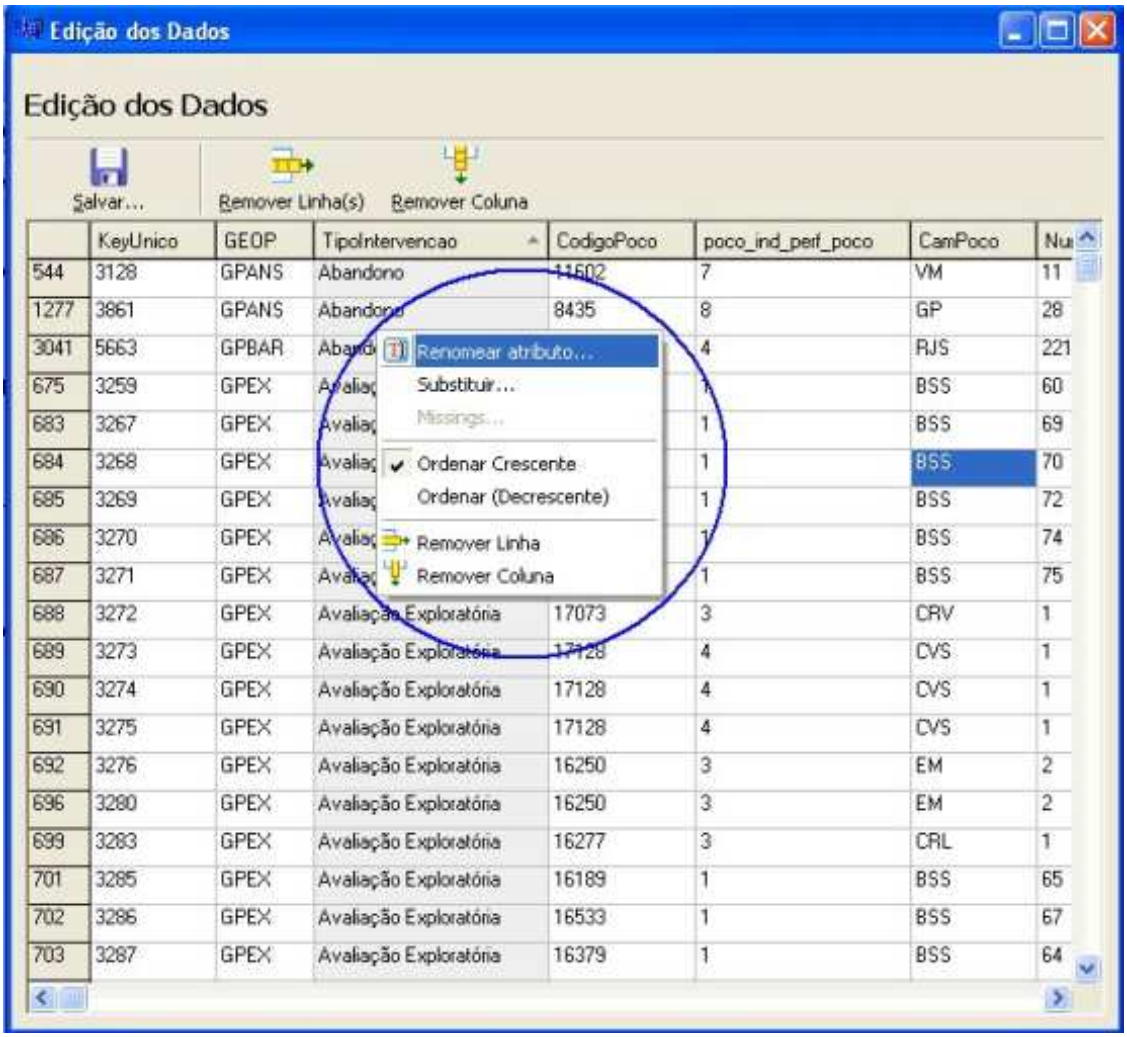

Figura 11: Edição dos dados

4. **Ordenar (Crescente):** reordenar a visualização de todos os registros por ordem crescente dos valores do atributo.

5. **Ordenar (Decrescente):** reordenar a visualização de todos os registros por ordem decrescente dos valores do atributo.

   A ordenação da coluna também é feita automaticamente ao clicar-se com o botão esquerdo do mouse sobre o nome da coluna. A coluna ordenada destaca-se das demais por possuir uma indicação (seta) ao lado do nome do atributo e pela sua cor de fundo diferenciada.

6. **Remover linha:** remover a linha da tabela.

7. **Remover coluna:** remover a coluna da tabela.

8. **Salvar:** permitir a exportação dos dados da tabela para dois possíveis formatos.

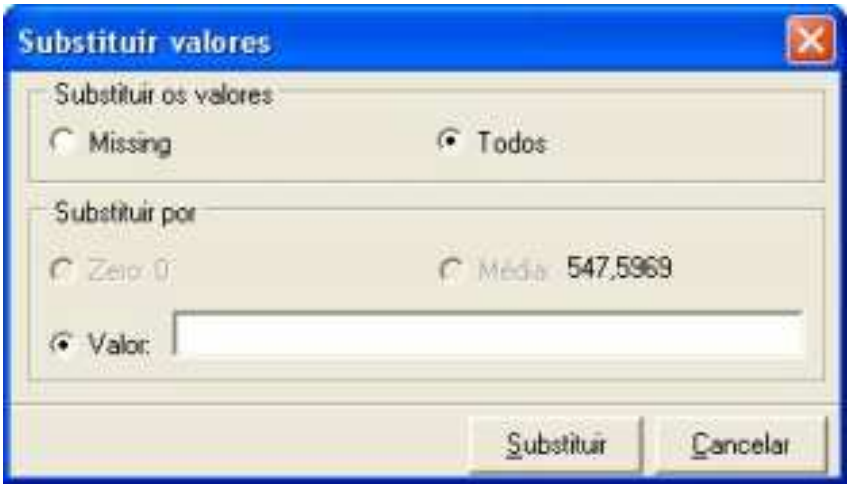

Figura 12: Substituição numérica

Figura 13: Substituição nominal

O primeiro para arquivo texto (*.txt) com o caractere de tabulação de delimitador e o segundo, também arquivo texto (*.csv) com o caractere de ponto-e-vírgula de delimitador.

Além disso, existem outras funcionalidades que podem ser usadas na tarefa de edição. Por exemplo, o valor de um campo pode ser alterado diretamente na tabela. Neste caso, para um atributo nominal através de uma lista dos possíveis valores que é exibida na tela através de um componente "combobox" e para um atributo numérico através do próprio valor numérico digitado.

Ainda, é possível remover da tabela várias linhas consecutivas selecionadas pelo usuário.

4.2 Visualização dos Dados

Ferramenta gráfica que permite fazer a vizualização dos dados através de gráficos de pontos e gráficos de barras. A janela possui dois modos de visualização dos dados: Gráficos de Pontos e Gráficos de Barras.

4.2.1 Gráficos de Pontos

O modo de visualização dos gráficos apresenta três partes principais:

1. **Eixos:** Estes campos definem quais atributos serão representados no eixo das abscissas (x) e das ordenadas (y) do gráfico de pontos.

   Conforme pode ser visto na figura 14 destacado em azul, **TipoIntervencao** e **TipoSonda** respectivamente.

2. **Gráficos de Pontos:** Mostra o conjunto de pontos de coordenadas (x, y) plotados da base de dados para os atributos definidas em eixos.

   O efeito de "zoom" no gráfico (+/-) é obtido selecionando-se a área do gráfico desejada com o botão esquerdo do mouse pressionado.

3. **Registros:** Mostra o(s) registros(s) correspondente(s) da base de dados de uma determinada coordenada (x, y) selecionada.

   Por exemplo, conforme porde ser visto na figura 14 destacado em preto, os registros exibidos correspondem aos casos da base que têm coordenadas (x=Perfuração Especial, y=NS). A coordenada para o(s) registro(s) é definida clicando-se sobre um dos pontos de coordenadas ("x" em vermelho) do gráfico.

4.2.2 Gráficos de Barras

O modo de visualização dos gráficos de barras apresenta três partes principais:

1. **Eixos:** Estes campos definem quais atributos correspondem ao eixo das abscissas (x) e das ordenadas (y) do gráfico de barras.

   Conforme pode ser visto na figura **??** destacado em azul,, ProfFinalSondador e TipoSonda respectivamente.

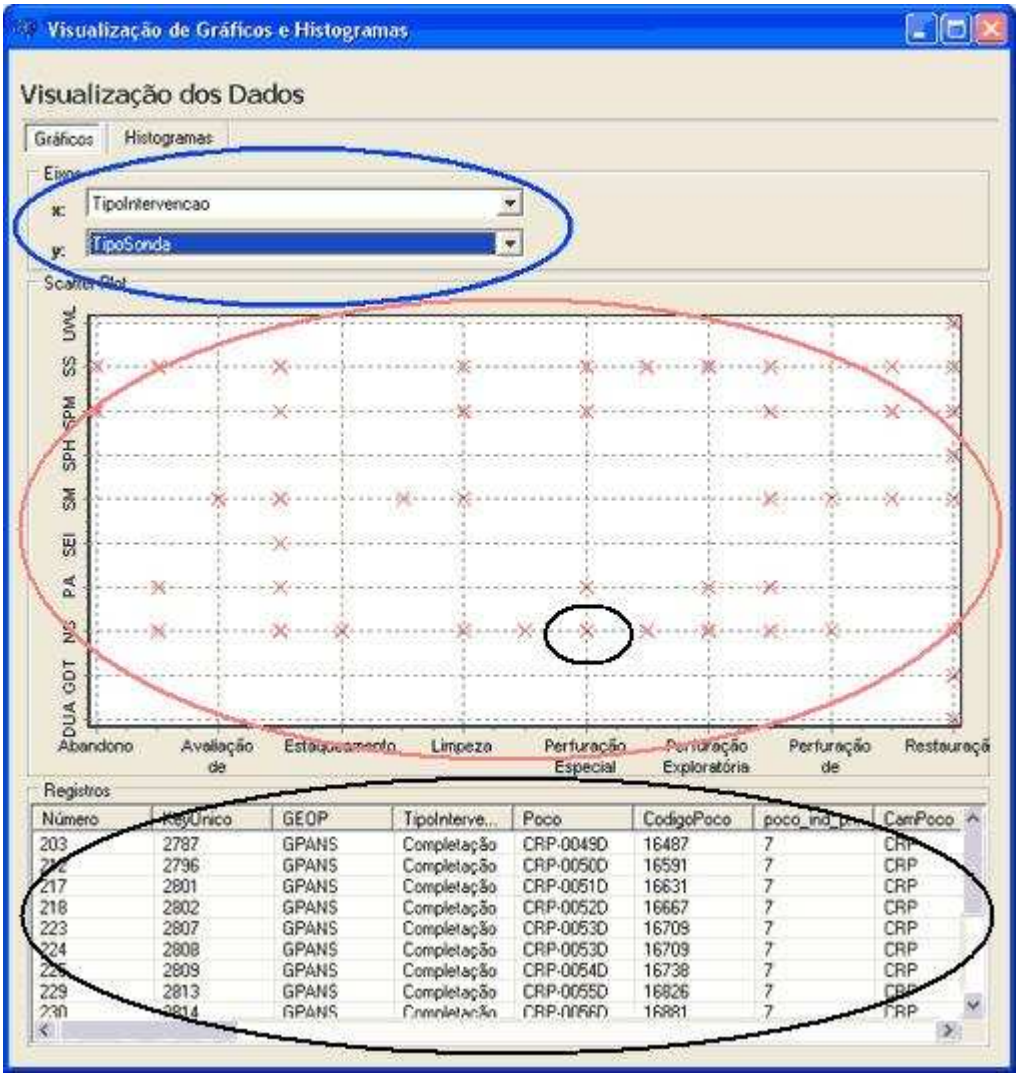

Figura 14: Gráficos de pontos.

2. **Gráficos de barras com legenda à direita:** Gráficos de barras verticais para o atributo da abscissa (x), cujas cores representam a freqüência em relação as categorias da ordenada (y). As **categorias** são definidas de acordo com o tipo do atributo, ou seja, para um atributo numérico serão o intervalo de classe dos seus valores e para um atributo nominal serão um dos seus possíveis valores.

   Por exemplo, conforme porde ser visto na figura 15, as barras verticais para a abscissa (x=ProfFinalSondador) são intervalos de classes dos seus valores e as categorias da ordenada (y=TipoSonda) são valores do atributo como NS, PA, SS.

   Da mesma maneira, o efeito de "zoom" no gráfico (+/-) é obtido selecionando-se a área do gráfico desejada com o botão esquerdo do mouse pressionado. A legenda aparece ao lado direito do gráfico de barras.

3. **Informações da relação (x, y)**: Exibe informações específicas sobre a relação

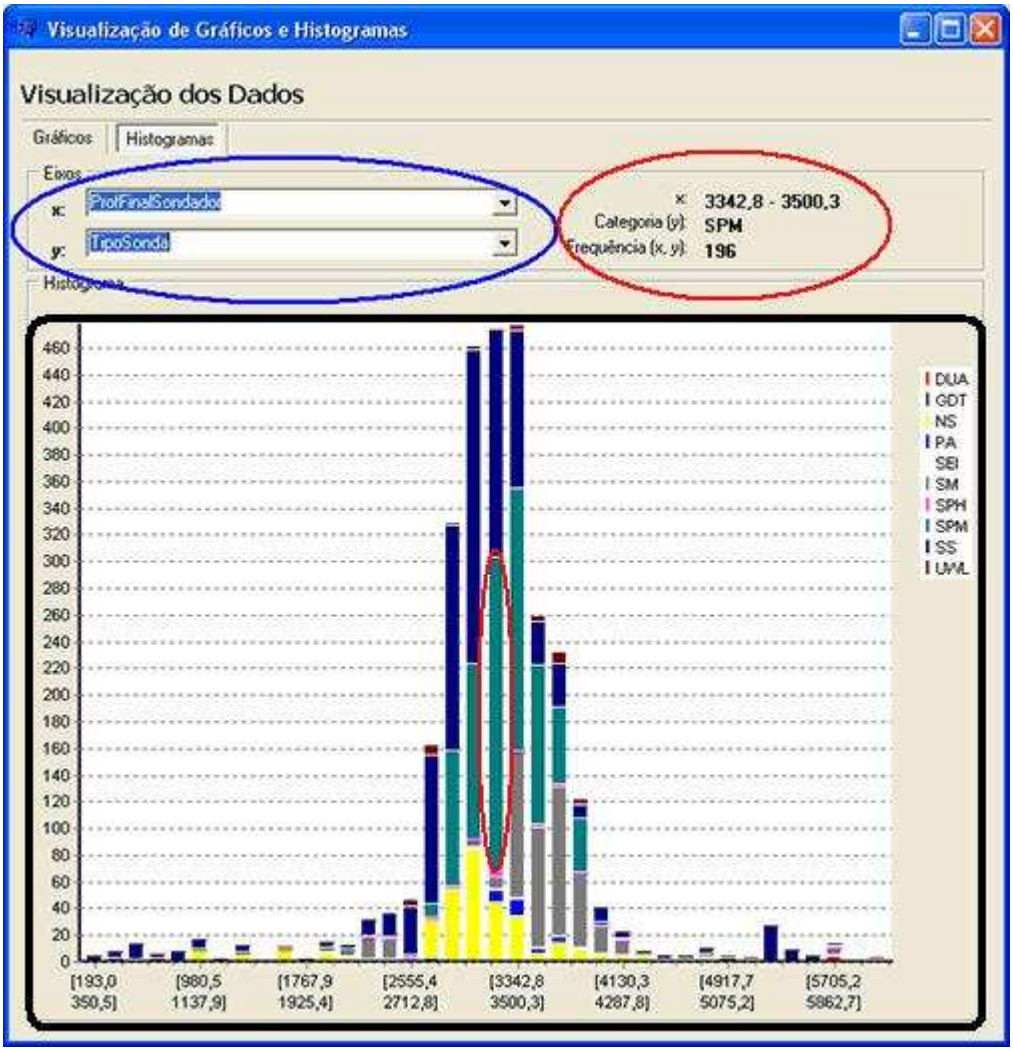

Figura 15: Gráficos de Barras

entre uma barra vertical (x) e uma categoria da ordenada (y). Cada categoria é representada por uma cor diferente na barra.

- **x:** valor da abscissa (x) que corresponde a barra vertical.
- **Categoria(y):** valor da categoria da ordenada (y) que está relacionada com a barra vertical (x).
- **Frequência (x, y):** valor da frequência de ocorrência nos dados da relação entre (x) e a categoria (y).

Porém, essas informações somente são exibidas quando o usuário clicar sobre uma das possíveis áreas da barra vertical, cada uma com sua cor específica. Por exemplo, na figura **??** destacado em vermelho, são apresentadas as seguintes informações:

- x: $[3342,8; 3500,3]$ representa o intervalo de classe para aquela barra vertical.

- Categoria (y): SPM é o valor da ordenada (y=TipoSonda) para aquela área selecionada da barra do gráfico de barras.
- Frequência: 196 representa a frequência de ocorrência da relação entre $[3342,8; 3500,3]$ e SPM nos dados.

5 *Metodologia de trabalho*

Foi realizado um estudo detalhado e aprofundado de algumas técnicas de análise exploratória de dados existentes na literatura, assim como, o estudo de caso do software WEKA, que incorpora diversos aspectos e conceitos de análise exploratória de dados. Com tudo isso, buscou-se um melhor embasamento teórico para a realização da proposta em questão.

5.1 Requisitos funcionais

- Importar os dados de uma base de dados corporativa, permitindo a visualização e edição destes dados.
- Possibilitar a nomeação e alteração dos campos da base de dados.
- As alterações nos dados não serão propagadas para a estrutura original da base de dados, ou seja, serão locais.
- Classificar os campos da base de dados de acordo com a natureza destes dados:
    1. Qualitativo Ordinal
    2. Qualitativo Nominal
    3. Quantitativo Discreto
    4. Quantitativo Contínuo
- Permitir a preparação dos dados para a etapa de mineração através de técnicas de seleção, pré-processamento e transformação destes dados.
- Permitir o tratamento de valores não preenchidos (missings).
- Poder obter uma distribuição mais simétrica dos dados, através de processos de transformação destes dados.

5.2 Especificação do sistema

5.2.1 Implementação

A implementação consiste no desenvolvimento de um módulo de software integrado, incorporando aspectos de mineração de dados, de acordo com os requisitos apresentados. A implementação do aplicativo será feita na linguagem de programação C++.

Parte da modelagem de classes UML proposta para a estrutura básica do sistema é apresentada na figura 16:

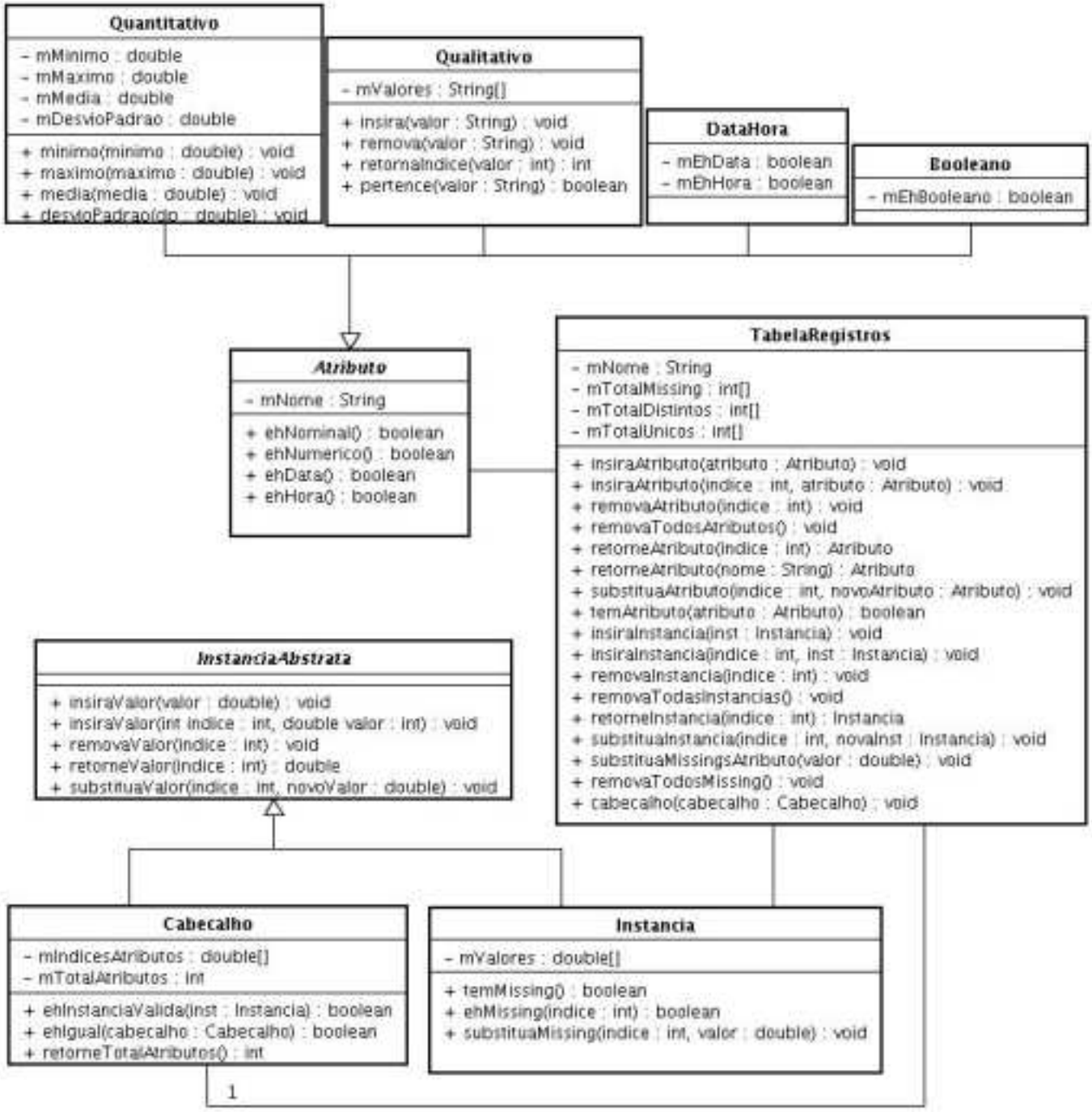

Figura 16: Diagrama de Classes da estrutura básica do sistema

6 Conclusões

A principal tarefa do módulo neural do E&P Risk IV é fazer uma estimativa do tempo total para realizar uma nova operação de perfuração ou completação de um poço de petróleo baseada em um conjunto de variáveis de dados. O processo de descoberta dessa informação é amplo e passa pelas seguintes etapas: importação e tipificação, limpeza e seleção, análise de agrupamentos e modelo de previsão.

A grande contribuição deste trabalho para o processo foi a implementação de soluções para a realização das duas primeiras etapas, basicamente, a preparação dos dados para o algoritmo minerador. A preparação dos dados é um passo importante e crítico para problemas de análise de dados modelados com redes neurais, pois a qualidade destes dados têm um grande impacto sobre os resultados finais, no caso, a estimativa do tempo total. Além disso, outro ponto forte é o sistema de imputação para o tratamento dos dados não preenchidos.

Desta maneira, pode-se afirmar que a ferramenta cumpriu todos os requisitos especificados através de diversas funcionalidades que facilitam a realização da tarefa de preparação. Uma das principais funcionalidades é o tratamento dos valores não preenchidos ("missings"), visto que, a restrição para a realização do treinamento da rede neural modelada para a análise de agrupamentos é que não existam valores "missings" nos dados de entrada.

Além disso, a ferramenta pode desempenhar o papel de um "filtro" para os dados. Através das ferramentas de edição e visualização pode-se identificar ruídos, valores corrompidos, inconsistências, redundâncias, ou seja, fazer uma limpeza nos dados.

Aplicou-se também de forma positiva um processo de tipificação automático de variáveis que serve para auxiliar o usuário não especialista do domínio a identificar os respectivos tipos. Este foi um dos maiores desafios do sistema devido a complexidade do problema.

Como trabalho futuro, pode-se incorporar novas técnicas avançadas para o processo de seleção de variáveis, por exemplo, algoritmos genéticos, assim como, novas técnicas para

o processo de tipificação de variáveis. Além disso, definir novos critérios para identificar correlações e associações entre campos da base de dados.

Um aspecto positivo do E&P Risk IV é a sua flexibilidade para a importação dos dados, ou seja, permite a obtenção dos dados de diferentes formatos de fontes. Desta maneira, outra idéia de trabalho futuro seria a extensão da ferramenta de importação dos dados para possibilitar a integração de várias fontes de dados em um único repositório centralizado.

Referências

BARBETTA, P. A. *A Estatística Aplicada às Ciências Sociais*. [S.l.]: Editora da UFSC, 1998.

BARBETTA, P. A. *Estatística para Cursos de Engenharia e Informática*. [S.l.]: Editora Atlas, 2004.

BUSSAB, W. de O. *Estatística Básica*. [S.l.]: Saraiva, 2003.

LOPES, C. H. *Descoberta de Conhecimento e Mineração de Dados*. PUC-Rio de Janeiro, 1999.

ROISENBERG, M. Risk assessment of drilling and completion operations in petroleum wells using a monte carlo and neural network approach. In: PROCEEDINGS OF THE 2005 WINTER SIMULATION CONFERENCE. [S.l.], 2005.

ROSE, P. R. *Risk Analysis and Management of Petroleum Exploration Ventures*. USA: The American Association of Petroleum Geologists, 2001. (AAPG Methods in Exploration, 12).

SANTOS, J. G. dos. *SETIP - Sistema Especialista para Tipificar Dados de uma Pesquisa: Variáveis Qualitativas e Quantitativas*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2001.

SILVA, W. T. da. *Anais do XXIV Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2004.

THE UNIVERSITY OF WAIKATO. *Weka Software*. [S.l.]. Disponível em: <http://www.cs.waikato.ac.nz/ ml/weka/>. Acesso em: 06 jul. 2005.

YU, L. An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2006.