UNIVERSIDADE FEDERAL DE SANTA CATARINA

Programa de Pós-Graduação em Engenharia de Produção

Área de concentração: Mídia e Conhecimento

PROGNÓSTICO DE DEMANDA DE POTÊNCIA ELÉTRICA PARA PLANEJAMENTO E OPERAÇÃO DE SISTEMAS ELÉTRICOS.

Cláudio Martin

Tese apresentada ao Curso de Pós-Graduação em Engenharia de Produção, Área de Concentração: Mídia e Conhecimento, da Universidade Federal de Santa Catarina - UFSC, como requisito para obtenção do título de Doutor em Engenharia de Produção.

Florianópolis

Dezembro 2005

Cláudio Martin

PROGNÓSTICO DE DEMANDA DE POTÊNCIA ELÉTRICA PARA PLANEJAMENTO E OPERAÇÃO DE SISTEMAS ELÉTRICOS.

Esta Tese foi julgada adequada para a obtenção do título de Doutor em Engenharia de Produção (Área de concentração: Mídia e Conhecimento) e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.

Florianópolis, 20 de dezembro de 2005

_____

Prof. Edson Pacheco Paladini, Dr.Eng.
Coordenador do Curso

BANCA EXAMINADORA

_____

Prof. Francisco Antonio Pereira Fialho, Dr. Eng.
Orientador

_____

Prof. Flávio Bortolozzi, Dr.
Examinador externo

_____

Prof. Luis Alberto Gómez, Dr.
Moderador

_____

Prof. Júlio César Nievola, Dr.
Examinador externo

_____

Prof. Edson José Rodrigues Justino, Dr.
Examinador externo

DEDICATORIAS

Para Amanda, que sempre se fez presente nos momentos difíceis.
Para Charles e Adriana, Fernando e Roberto que conviveram conosco as dificuldades.

AGRADECIMENTOS.

Aos amigos da UFSC, do CEFET-PR(agora UTFPR) e da PUC-PR que contribuíram para que este trabalho pudesse ser realizado.

SUMÁRIO

LISTA DE FIGURAS

RESUMO

MARTIN, Cláudio. Prognóstico de demanda de potência elétrica para planejamento e operação de sistemas elétricos. Florianópolis, 2005. 205p. Tese(doutorado em Engenharia de Produção) – curso de pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina.

O objeto da pesquisa é o estudo das técnicas utilizadas para o prognóstico de demandas de potência elétrica, apresentadas na forma de séries temporais, motivado pela necessidade de definir valores futuros com base em valores históricos, visando a utilização em estudos de planejamento de curto prazo de sistemas elétricos, utilizando a máquina de vetor de suporte como ferramenta de treinamento e prognóstico. A máquina de vetor de suporte desenvolvida nos anos 90, baseada na teoria estatística da aprendizagem se apresenta como uma máquina de aprendizagem universal, em que não é necessário determinar *a priori* o número de neurônios da camada de entrada, o qual é definido ao longo do processo. O projeto da máquina depende da extração de um subconjunto, dos dados de treinamento, que servirão como vetores suporte, representando uma característica estável do conjunto de dados de treinamento, evitando o excesso de ajuste da máquina de aprendizagem. A pesquisa efetuada apresenta as principais formulações existentes para regressão de séries temporais, focando sobre o desenvolvimento da máquina de vetor de suporte, apresentando um estudo de caso com série temporal de demanda de potência. Os resultados encontrados com a utilização da máquina de vetor de suporte são compatíveis com as necessidades do planejamento e operação de sistemas elétricos, apresentando erros de prognóstico menores, na ordem de 16,3% para os prognósticos dos perfis de demanda de meio de semana(terça a sexta-feira), e, de 37,3% para os prognósticos dos perfis de final de semana(sábado, domingo e segunda-feira), comparativamente com aqueles obtidos com a rede de função de base radial.

Palavras-chave: **Séries temporais, Prognóstico, Função núcleo, Máquina de vetor de suporte.**

ABSTRACT

MARTIN, Cláudio. Fortecasting of Electric Power Demand for Planning and Operation of Power Systems. Florianópolis, 2005. 205p. Doctorate thesis(Doctorate in Production Engineering) – Post-Graduation Program in Production Engineering, Federal University of Santa Catarina.

The research's aim is the study of techniques used for forecasting of electrical power demands presented in the form of time series, caused by the need of determining future values based on historical values, aiming at the used in studies of short-term planning of electrical systems, using the support-vector machine as tool for training and forecasting.

The support-vector machine developed in the 90's, based in the statistical learning theory, does not have characteristics present in the neural networks, that is, overfitting in the training stage caused by the excessive in-out examples of training set.

The support- vector machine is introduced as a machine of universal learning, in which it is not necessary determine *a priori* the number of neurons in the first layer, it is defined along the learning process. Changes of the kernel function present in the support -vector machine, can modify the type of the learning machine, and consequently the way of the function approaching. The machine's design depends on the extraction of a sub-set of the training data, that will serve as support vectors, presenting a stable characteristic of the training data set. The research introduces the main formularization of time series regression, focusing the development of the support-vector machine, presenting a study of case with time series of power demand. The results found with the utilization of the support-vector machine are compatible with the planning needs and the operation of power systems, presenting minor forecasting errors, about 16,3% for the forecasting of middle week demand profiles(Tuesday to Friday), and about 37,3% for forecasting of weekend demand profiles(Saturday to Monday), when compared to the results achieved with Radial Basis Function Networks.

Key-words: **Time Series, Forecasting, Kernel Function, Support Vector Machine.**

CAPÍTULO I – INTRODUÇÃO.

Métodos de prognóstico de valores, obtidos a partir de observações do ambiente, constituindo as chamadas séries temporais, tem sido pesquisados desde os primórdios, primeiramente de uma forma intuitiva, evoluindo para algoritmos matemáticos mais sofisticados, sempre com a finalidade de interpretar possíveis resultados futuros, antecipando eventos, permitindo a tomada de decisão com maior conhecimento do problema.

O prognóstico de séries temporais pode ser utilizado em diversas aplicações importantes, tais como dados em série (índice de bolsa de valores, câmbio, evolução de preços de mercado, etc), dados fisicamente observáveis (temperatura, índice pluviométrico, demanda de potência, etc) ou mesmo séries matemáticas (seqüência de Fibonacci, equações integrais e diferenciais)(Yang, 2003).

Uma série temporal com dados fisicamente observáveis é a da demanda de potência elétrica consumida por uma determinada cidade, região ou estado, cujo comportamento poderá subsidiar estudos do sistema elétrico visando questões econômicas, segurança e qualidade no fornecimento de energia aos consumidores.

As atividades, relacionadas com a operação e planejamento de um sistema elétrico, requerem o conhecimento prévio do consumo de energia elétrica de seus consumidores, que pode ser obtido a partir da série temporal representativa da demanda de potência elétrica. O prognóstico de carga futura pode ser categorizado como sendo de curtíssimo prazo(*very short-term load forecasting)*, de curto prazo*(short-term load forecasting)*, de prazo médio*(mid-term load forecasting)* e de longo prazo*(long-term load forecasting)*. O prognóstico de curtíssimo prazo refere-se a um tempo de alguns minutos até milhares de minutos, é requerido para os estudos de controle de carga –freqüência e funções de despacho econômico de carga do sistema de gerenciamento de energia. O prognóstico de curto prazo refere-se a um tempo de horas até uma semana e tem como objetivo alimentar as decisões relativas a programar a capacidade de geração do sistema, programar compras de combustíveis para geração, avaliar sistema de segurança do fornecimento de energia e planejamento de trocas de energia com outros sistemas. O prognóstico de médio prazo está situado em um período de tempo que pode ser de um mês até vários meses, e estrutura o planejamento energético do sistema dentro de um ano. O prognóstico de longo prazo estabelece um período de horizonte de estudo que

varia entre um ano e vários anos e é utilizado para o planejamento energético de longo prazo(Khotanzad, Rohani, Lu, Abaye, Davis, Maratukulam, 1997)(Charytoniuk, Chen, 2000).

A precisão do prognóstico de demanda de potência de curto prazo pode representar efeitos importantes no gerenciamento econômico de sistemas elétricos, que para o caso de valores prognosticados superiores aos requeridos, ter-se-á(Hobbs, Jitprapaikulsarn, Konda, Chankong, Loparo, Maratukulam, 1999):

- Unidades geradoras desnecessariamente reservadas, elevando os custos de combustível e de manutenção;
- Aquisição de energia mais cara não necessária, ou perda de uma oportunidade de venda de energia excedente;
- Produção de energia hidráulica que poderia ser gerada em um momento mais rentável;
- Cotação excessivamente elevada do preço real da energia, reduzindo vendas;
- Interrupções desnecessárias ou controle de carga, solicitadas, incomodando consumidores e reduzindo receitas.

Para o caso de se ter valores prognosticados inferiores aos requeridos, pode –se ter:

- Insuficiência de recursos para suprir restrições do comitê de segurança energética, tal como, margem de reserva girante necessária para cumprir o índice de confiabilidade do sistema;
- Necessidade de alocação de geração não econômica para suprir o crescimento de demanda não prevista, ou, alternativamente, solicitação de interrupção de cargas e controles não previstos caso o prognóstico tivesse sido correto;
- Reserva de compra de energia processado com preço menor do que aquele utilizado para atender a demanda superior de energia;
- Cotação do preço real da energia muito baixo, resultando na redução da incidência, da receita, nos custos da empresa.

Prognosticar demanda de energia elétrica é uma tarefa difícil, pois a série temporal representativa da demanda de potência apresenta vários níveis de sazonalidade: A

demanda de determinada hora é dependente não só da demanda da hora anterior, mas também da demanda da mesma hora do dia anterior e, da mesma forma, também, da demanda da mesma hora, no dia de mesma denominação da semana anterior. Encontrar valores de prognóstico com valores percentuais de erro na ordem de 10% é relativamente fácil, no entanto, o custo do erro sempre se apresenta muito elevado, justificando qualquer tentativa para reduzi-lo. Estimativas de custo do erro dão conta que o acréscimo de 1% no erro de prognóstico, resulta em um acréscimo de custo de operação do sistema elétrico da Grã Bretanha (em 1984) na ordem de £10 milhões por ano (Bunn, Farmer, 1985 apud Hippert, Pedreira, Souza, 2001 e Sfetsos, 2003).

As técnicas nas quais estão baseados os processos mais utilizados para prognosticar demandas de potência podem ser classificadas segundo duas abordagens: A abordagem que trata os padrões de carga medidos como séries temporais, e efetua o prognóstico de demanda futura a partir de valores ocorridos no passado. A outra abordagem considera que além dos valores ocorridos no passado, os valores futuros são bastante dependentes das condições ambientais (temperatura, umidade, etc). Esses modelos, que incluem variáveis ambientais, são limitados no seu uso considerando a imprecisão do prognóstico dessas variáveis e das dificuldades de modelagem da relação entre demanda de potência e essas variáveis (Park, Park, Lee, 1991 apud Kodogiannis, Anagnostakis,1999).

A abordagem apresentada abre a perspectiva de aplicação de novas modelagens, com a finalidade de melhorar os resultados de prognóstico e conseqüentemente reduzir o grau de incerteza no planejamento dos sistemas elétricos, permitindo uma substancial redução nos custos de operação.

A máquina de vetor de suporte tem, desde o seu aparecimento nos anos 90, atraído os pesquisadores da área de prognóstico, para o desenvolvimento de aplicações específicas, devido ao seu tratamento matemático, interpretação geométrica e uso prático.

A aplicação de máquinas de vetor de suporte se apresenta de uma forma sistemática, reproduzível e bem fundamentada na teoria estatística da aprendizagem.Treinar envolve a otimização de uma função de custo convexa: não apresenta pontos de mínimo local que venham a complicar o processo de aprendizagem(Bennet, Campbell, 2000).

1.1 Estabelecimento do problema.

Pesquisas efetuadas nos anos que precederam o período de reestruturação da indústria de energia elétrica em que ocorreram fusões e privatização do setor tem conduzido o estudo de prognóstico de demandas de potência a um patamar de performance satisfatório, na ordem de 1 a 2%(Bunn, Farmer, 1985 apud Bunn, 2000), com a finalidade de atender os requisitos operacionais do sistema elétrico.

Para as economias em desenvolvimento, as incertezas de operação e planejamento dos períodos futuros, conduzindo a padrões de segurança elevados, tem sido gerenciados a partir da construção de margens de reserva substancial dentro da modelagem do planejamento de capacidade instalada.

À abertura e chegada de mercados competitivos, foram associadas expectativas de participação de grandes consumidores e ganhos de eficiência para consumidores e empresas detentoras dos estoques de mercado*(shareholders)*. Um mercado ativo e competitivo de venda de energia no atacado poderá transformar o risco físico de capacidade inadequada para atender ao mercado em um risco financeiro de altos preços.

Enquanto o mercado de consumidores apresentar o mesmo perfil, as companhias podem suportar as incertezas de demandas através de técnicas financeiras do gerenciamento do risco ao invés de investimentos em plantas físicas. No entanto, este procedimento requer prognósticos com elevado grau de certeza, não somente para estimar demandas e preços mas também para determinar a extensão e períodos de mudanças voláteis.

Uma das principais implicações desse novo cenário competitivo é de que erros em prognóstico podem gerar redução substancial dos lucros, dividir mercados e reduzir o valor real das empresas. Assim sendo, o prognóstico futuro de demandas assume uma importância vital tanto para subsidiar as ações de planejamento e operação do sistema como também para estabelecimento de políticas de tarifação do mercado livre.

Técnicas associadas ao tema relacionado a prognóstico de demandas podem ser identificadas como(Bunn, 2000):

- Segmentação da variável a ser prognosticada em modelos separados;
- Combinação de prognósticos de vários modelos;
- Especificação e estimação de modelos generalizados usando redes neurais.

Uma máquina de aprendizagem apresenta uma habilidade de generalização limitada a um risco funcional dado pela soma de duas componentes, a do risco empírico e a do intervalo de confiança, conduzindo a duas formas construtivas possíveis de máquinas.
A primeira forma fixa o intervalo de confiança, a partir da construção de uma máquina apropriada, e minimiza o risco empírico(número de erros no conjunto de treinamento). É o caso da construção das redes neurais, que para valores elevados do intervalo de confiança, impostos ao projeto, poderá vir a ocorrer problemas de excesso de ajuste*(overfitting)*.
A segunda forma fixa o valor do risco empírico(por exemplo zero), e minimiza o intervalo de confiança(Vapnik, 2000). É o caso da máquina de vetor de suporte, que será utilizada para efetuar o estudo objeto do problema a ser estabelecido nesse item.
Observando o cenário descrito do setor de energia elétrica, aliado à necessidade de melhorar as ferramentas para determinar o prognóstico de demandas de energia e a crescente aplicação da máquina de vetor de suporte em aplicações específicas, surge naturalmente uma questão norteadora importante que pode ser formulada: Como desenvolver uma máquina de vetor de suporte para regressão de séries temporais específicas?

1.2 Objetivos.

O objetivo geral do estudo é desenvolver uma máquina de aprendizagem com vetores suporte, capaz de apresentar valores de prognóstico de demanda de potência compatíveis com as necessidades do planejamento e operação de sistemas elétricos.

Com base no objetivo geral elaboraram-se os seguintes objetivos específicos:

1 - Apresentar os principais métodos que podem ser aplicados para regressão de séries temporais;
2 - Analisar a teoria de máquinas de vetor de suporte aplicável para regressão de séries temporais;
3 -Identificar por meio de estudo de caso uma máquina de vetor de suporte capaz de prognosticar série temporal de demanda de potência elétrica de curto prazo;
4 -Apresentar resultados comparativos entre a máquina de vetor de suporte para regressão *(SVMR)* e uma rede de função de base radial *(RBF)* para o estudo de caso;

5 -Delinear uma possível relação entre as variáveis de entrada e os parâmetros da máquina de vetor de suporte para o caso em estudo.

1.3 Justificativa.

O tempo é uma grandeza de importância vital para os fenômenos da natureza, pois quase todos os problemas do mundo real podem ser atrelados a uma variável temporal. Cada tipo de dado apresenta uma dependência temporal, que pode apresentar–se de uma forma explícita com caracterização temporal ou implicitamente em que os dados são coletados de um processo que varia com o tempo(Rüping, 2001).

A análise estatística de séries temporais desenvolveu duas grandes classes de representação, aquelas denominadas de domínio do tempo e aquelas do domínio da freqüência. A análise no domínio do tempo é baseada na correlação entre os valores atuais e valores previamente observados, enquanto o domínio da freqüência decompõe a série temporal em componentes cíclicas de diferentes freqüências. A análise de séries temporais com a finalidade de aprendizagem, tem os seguintes objetivos: *descrição*(descreve a série temporal através de determinada estatística), *explanação*(entende o processo representado pela série temporal), *prognóstico*(prognostica valores futuros da série temporal) e *controle*(controla o processo atrás da série temporal para gerar valores futuros viáveis)(Rüping, Morik, 2003).

A máquina de vetor de suporte(SVM), modelagem a ser utilizada no presente trabalho, requer uma entrada de dados numéricos com comprimento fixo, não representando a variável tempo explicitamente(Morik, 2000).

A grande vantagem da máquina de vetor de suporte é de que as equações relevantes que descrevem a generalização do erro não dependem da dimensão do conjunto de dados, mas somente da margem de separação do hiperplano(superfície de decisão linear multidimensional), fazendo com que a modelagem utilizando SVM seja especialmente apropriada para conjunto de dados com alta dimensionalidade. Apesar de que esta argumentação não é estritamente válida - a margem de separação do hiperplano depende da geometria dos dados e também da dimensão - evidências empíricas dão conta de que esta propriedade se mantém na aplicação prática(Joachims, 1998 apud Rüping, Morik, 2003).

A utilização da máquina de vetor de suporte tem se estendido para os mais diversos campos de aplicação, tendo sido utilizada para efetuar o prognóstico de demanda de potência, constituindo-se em um primeiro trabalho de aplicação de SVM, com sucesso, para este tipo de série temporal(Chen, Chang, Lin, 2001).
O sucesso na aplicação da máquina de vetor de suporte, está relacionado com a escolha correta da representação dos dados. A excelente propriedade de generalização da SVM, especialmente sua boa performance em dados com alta dimensionalidade, torna fácil a tarefa de melhorar os resultados a partir da adição de características temporais adicionais e da construção de funções núcleo especializadas.

1.4 Delimitação do estudo.

Entre os anos 60 e 80, foi introduzida nova conceituação na estatística, que responde a uma importante questão apresentada a seguir:

O que necessitamos saber a priori a respeito de uma dependência funcional desconhecida, com a finalidade de estima - lá com base em observações?(Vapnik,2000).

Nos anos 20 e 30 a resposta a essa questão era –necessitamos saber quase tudo(Fischer, 1925 apud Haykin,2001). Pelo novo paradigma, com a finalidade de estimar a dependência entre os dados, é suficiente conhecer algumas propriedades gerais do conjunto de funções de onde provêem a dependência desconhecida. O conjunto de teorias desenvolvidas para responder à questão proposta deu origem à teoria estatística da aprendizagem, que corresponde a um dos temas gerais a serem abordados no presente trabalho.
A teoria da otimização responde pela filosofia matemática necessária para desenvolver soluções ao conjunto de equações representativas da modelagem adotada, considerando todas as possíveis restrições às quais estão sujeitos problemas de natureza complexa.
A junção da teoria estatística da aprendizagem, com a teoria da otimização, permitiu desenvolver máquinas capazes de responder com performance aos problemas de prognosticar seqüências numéricas, apresentadas na forma de séries temporais.
Com a necessidade de trabalhar com um conjunto de dados com alta dimensionalidade, foi desenvolvida uma máquina que utiliza um conjunto de vetores representativos da

série temporal em estudo, permitindo uma aproximação da função alvo sem incorrer no risco de haver um excesso de ajuste da curva(overfitting), denominados de vetores suporte.

A necessidade de trabalhar com sistemas não lineares, conduziu a pesquisa para transformar o espaço primordial em estudo, em um espaço dual de alta dimensionalidade, através de uma função de transformação, no qual é efetuada a otimização. A transformação dos vetores suporte para o espaço de alta dimensionalidade é efetuada com a utilização das funções núcleo, que de uma forma elegante constroem uma versão não-linear de um algoritmo linear.(Müller, Mika, Rätsch, Tsuda, Schölkopf, 2001).

Assim uma série temporal não-linear de dimensão elevada(por exemplo para prognosticar demandas de potência elétrica de curto prazo) é transformada em uma série temporal de vetores suporte, reduzindo a dimensão do vetor de entrada, obtendo uma solução linearizada através do mapeamento, com funções núcleo, em um espaço de alta dimensionalidade, definindo a máquina de vetor de suporte, conforme esquematizado na Figura 1.1.

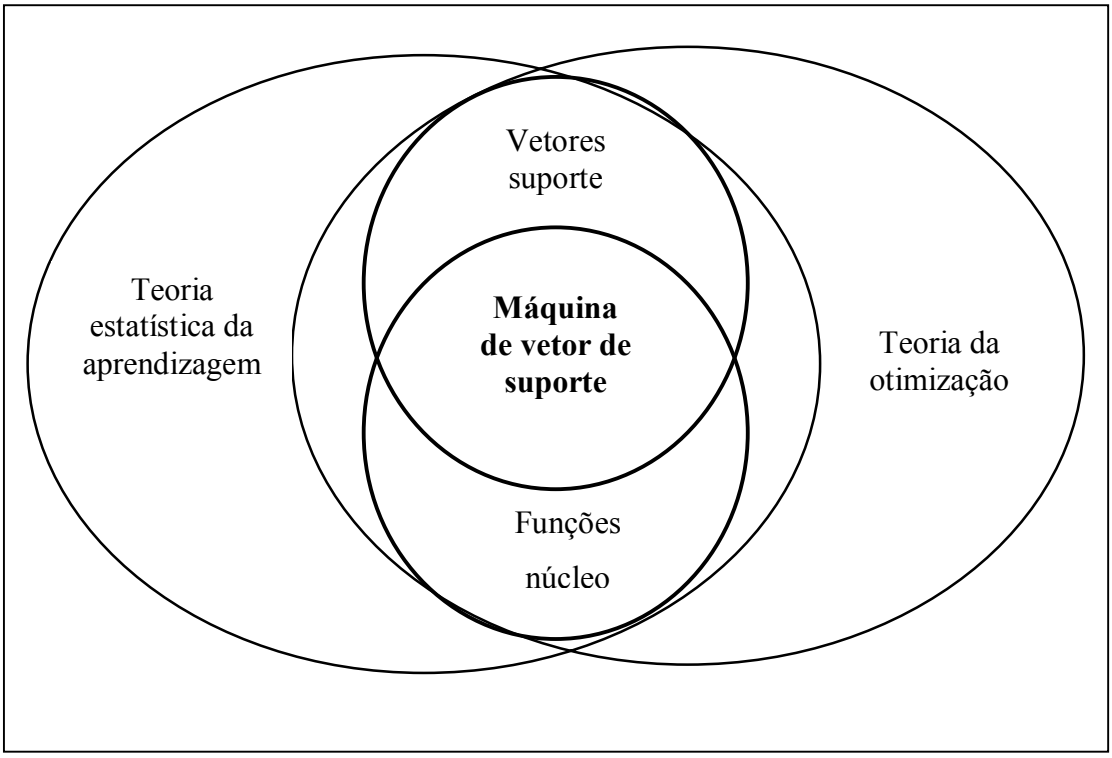


Figura 1.1 – Delimitação do estudo.

1.5 Estrutura do trabalho.

Este trabalho está assim organizado:

O Capítulo I apresenta a importância de se obter o prognóstico de valores de demanda de potência, desde prognósticos de curtíssimo prazo até de longo prazo, sendo que cada um apresenta características específicas e se destinam a diferentes fases do planejamento do setor energético. A importância de se prognosticar demandas de potência sugere a utilização de procedimentos específicos de tratamento dos dados, recaindo sobre a máquina de vetor de suporte o escopo do presente trabalho. O objetivo apontado como relevante para o trabalho é a customização de uma máquina de vetor de suporte capaz de apresentar resultados compatíveis com a necessidade de planejamento elétrico, bem como, permitir o estabelecimento de algumas regras para obtenção das variáveis do modelo. A capacidade da análise estatística de séries temporais em dar respostas aos fenômenos de natureza social, aliada á capacidade de generalização da máquina de vetor de suporte e a sua boa performance para dados com elevada dimensionalidade, justificam a relevância da pesquisa efetuada.

O Capítulo II apresenta os diversos modelos para análise de séries temporais, dividindo a apresentação em modelagem linear e não-linear. Dentro da modelagem linear são apresentados os modelos ARMA, o ajuste exponencial e a modelagem de espaço de estado. A modelagem não-linear apresenta os modelos não lineares pré-definidos, não lineares com volatilidade variável e os modelos classificados como não lineares gerais. Nos modelos não lineares gerais estão classificados os de aprendizagem por reforço, aprendizagem supervisionada e não-supervisionada e a aprendizagem estatística. As redes neurais, classificadas como modelos não lineares gerais, são apresentadas na sua estrutura perceptron, redes de função de base radial e mapas auto-organizáveis. A máquina de vetor de suporte, classificada como modelo não linear geral e aprendizagem estatística supervisionada, é objeto de estudo dos capítulos subseqüentes. Para os casos de modelagem linear e não-linear são apresentados casos exemplos de ajuste de séries temporais de demanda de potência elétrica.

O Capítulo III apresenta os vetores suporte e as funções núcleo, iniciando com a teoria estatística da aprendizagem, definindo a minimização estrutural do risco composto pelo termo que representa o risco empírico e outro do intervalo de confiança. A forma de

tratamento dado a esses dois termos define o método de aprendizagem e a capacidade de generalização da máquina, em redes neurais e máquina de vetor de suporte. A teoria da otimização apresenta a formulação primordial e dual de solução de um problema geral, com a utilização dos multiplicadores de Lagrange, complementados com as condições KKT. A aplicação da otimização ao problema de regressão na máquina de vetor de suporte, com a introdução da função de perda insensível a ε, define os vetores suporte.

As funções núcleo efetuam o mapeamento do espaço de entrada de dados em um espaço de alta dimensionalidade, permitindo a transformação de um problema não-linear em uma otimização linear. São apresentadas as propriedades das funções núcleo e sua principal caracterização pelo teorema de Mercer(Mercer, 1909). A construção de funções núcleo é apresentada de forma a permitir obter núcleos, para estimar funções de valores reais, que geram funções geradoras flexíveis(splines), que geram expansões da série de Fourier e funções núcleo ANOVA. A construção de máquinas de vetor de suporte para regressão de séries temporais se constitui em obter através de algoritmos a solução do problema para um dado parâmetro de regularização C, uma função de perda ε e uma função núcleo k**.** São apresentados os algoritmos dos mínimos quadrados, de regressão de aresta*(ridge regression),* regressão com perda insensível a ε, quadrática e linear e regressão com vetores suporte e parâmetro ν.

O Capítulo IV apresenta a metodologia aplicada para o desenvolvimento do estudo e apresenta as características da série de demanda de potência elétrica utilizada.

O Capítulo V apresenta o estudo de caso de regressão de série temporal de demanda de potência para prognóstico de carga de curto prazo, comparando com resultados obtidos com uma rede de função de base radial. Os casos apresentados permitem visualizar o desempenho da máquina de vetor de suporte, frente às diversas variáveis que compõe o algoritmo de cálculo, tais como, o raio da região tubular ε, o parâmetro de regularização C, a função núcleo k e o número de iterações do processo de aproximação da solução.

O Capítulo VI apresenta as conclusões do trabalho e recomendações para trabalhos futuros.

CAPÍTULO II – MODELOS DE ANÁLISE DE SÉRIES TEMPORAIS.

Muitos métodos de análise de séries temporais foram desenvolvidos a partir da necessidade de prognosticar valores futuros com base em valores ocorridos no passado. Este capítulo apresenta os principais métodos utilizados para a análise, com modelagem linear e não-linear, e, alguns resultados obtidos no prognóstico de demandas de potência elétrica.

Séries temporais são, valores observados de forma seqüencial relacionados a um dado fenômeno, computados no tempo (Chatfield , 2003).

A maioria dos fenômenos pode ser representada através de séries temporais, sendo as mais freqüentes as séries temporais financeiras (índices da bolsa, cotação de moedas, etc.), séries temporais relacionadas às condições atmosféricas (temperatura, umidade do ar, velocidade do vento, índice pluviométrico e outros) e aquelas relacionadas às atividades de marketing, que tem a finalidade de orientar a aquisição de produtos pelo comércio e subsidiar a produção industrial.

No entanto, os resultados obtidos com a multiplicação dos métodos de representação, das ocorrências dos fenômenos temporais, e a conseqüente possibilidade de obter, a partir de valores ocorridos, prognósticos cada vez mais confiáveis de valores futuros, fez com que os mais variados fenômenos fossem representados, de forma a permitir, a utilização destas ferramentas.

Assim, também, o prognóstico de demanda de potência consumida em um determinado espaço geográfico delimitado, tornou-se uma das grandezas passíveis de serem representadas na forma de séries temporais e conseqüentemente permitir a utilização da matemática desenvolvida para tal fim.

As séries temporais apresentam características de valores de forma contínua no tempo, no entanto normalmente os valores são observados a intervalos de tempo igualmente espaçados pré-determinados, configurando uma série temporal discreta.

Normalmente os valores sucessivos de uma série temporal não são independentes, mas a análise deve levar em consideração a ordem temporal das observações, de modo a considerar a dependência dos valores no tempo.

Como o prognóstico futuro da série temporal geralmente não é determinado com exatidão a partir dos valores ocorridos no passado, a maioria das séries temporais, são

ditas estocásticas uma vez que somente parte dos valores passados contribui para o prognóstico, sendo necessário introduzir a idéia de distribuição probabilística.

2.1 Prognóstico

Determinar valores futuros de séries temporais observadas se constitui em um importante problema para várias áreas, dentre elas a econômica, planejamento de produção, determinação de preços de mercado e controle de estoques.

Existe uma variedade muito grande de procedimentos de prognóstico sendo que não há nenhum processo simples universalmente aplicável.

Os métodos de prognóstico podem ser classificados em três grupos (Chatfield, 2003):

Subjetivo é uma metodologia que utiliza dados qualitativos para obter resultados que permitam efetuar julgamentos intuitivos a partir do desenvolvimento de cenários capazes de indicarem valores para as varáveis desconhecidas do problema.

Alguns julgamentos são tomados de forma intuitiva para melhor definir a modelagem a ser utilizada no problema ou mesmo para proporcionar um ajuste aos resultados do prognóstico. Mesmo que muitos planejadores tenham incorporado uma capacidade de avaliar cenários de forma subjetiva, eles preferem realmente prognosticar situações futuras utilizando dados quantitativos, ao invés de confiar no seu próprio julgamento (Georgoff e Murdick, 1986).

Pesquisas efetuadas (Makidrakis e Hogarth, 1981), concluíram que mesmo a técnica quantitativa mais simples, desautoriza a utilização de avaliações intuitivas, geralmente desestruturadas, de “experts”, no julgamento para o ajuste de valores obtidos de prognósticos qualitativos, uma vez que sempre são acompanhados de uma perda de exatidão no processo. Este julgamento é desta forma, porque prognósticos intuitivos são suscetíveis a tendências, e os indivíduos são limitados nas habilidades de processar informações e manter relações consistentes entre variáveis (Sjoberg, 1982);

Univariável é um processo quantitativo onde o ajuste é baseado somente em valores presentes e passados da variável em estudo;

Multivariável é um processo quantitativo em que o prognóstico da variável em estudo depende do conhecimento do comportamento temporal de outras variáveis do problema.

Na prática para identificar o método que melhor se adapta ao problema em estudo, é comum considerar a aplicação de vários métodos de forma a gerar cenários que

permitam estabelecer a modelagem matemática que melhor representa a série de valores. Para o estudo de caso a ser desenvolvido, considerar-se-á a série temporal dependente somente dos próprios valores ocorridos no passado, considerando que a introdução de outras variáveis, a exemplo da temperatura, são limitadas no seu uso dada a imprecisão do prognóstico das mesmas e das dificuldades de modelagem(Park, Park, Lee, 1991 apud Kodogiannis, Anagnostakis, 1999)

2.2 Descrição de modelos

Os modelos utilizados para o prognóstico de séries temporais podem ser classificados em modelos lineares e não-lineares. Não existe uma definição clara do que venha a ser uma modelagem de série temporal estocástica linear e o que significa uma modelagem não-linear.

Assim um processo linear é classificado como geral, quando o valor atual da série temporal pode ser expresso como função linear de valores presente e passados de um processo puramente randômico.

Da mesma forma um método de prognóstico linear é aquele em que os valores prognosticados do futuro podem ser expressos como função linear de valores observados, incluindo o tempo presente (Chatfield, 2003).

Uma definição de sistema linear, originada do ponto de vista da estatística, em que a identificação de sistemas do tipo entrada/saída adquire uma importância maior, considera que para respostas do tipo $y_1(t)$ e $y_2(t)$, contínuas no tempo, correspondentes aos sinais de entrada $x_1(t)$ e $x_2(t)$, uma combinação linear dos valores de entrada deverá ser reproduzida na saída.

Assim a combinação linear $\lambda_1 x_1(t) + \lambda_2 x_2(t)$ deverá produzir a mesma combinação linear na saída, ou seja, $\lambda_1 y_1(t) + \lambda_2 y_2(t)$, onde λ_1 e λ_2 são quaisquer constantes.

Existem modelos que podem ser classificados de localmente lineares, no entanto são globalmente não-lineares. A definição exata de linearidade não é fácil de ser precisada, mas é possível mover-se gradualmente da linearidade para a não-linearidade.

A Figura 2.1 apresenta um diagrama de classificação dos modelos de análise de séries temporais.

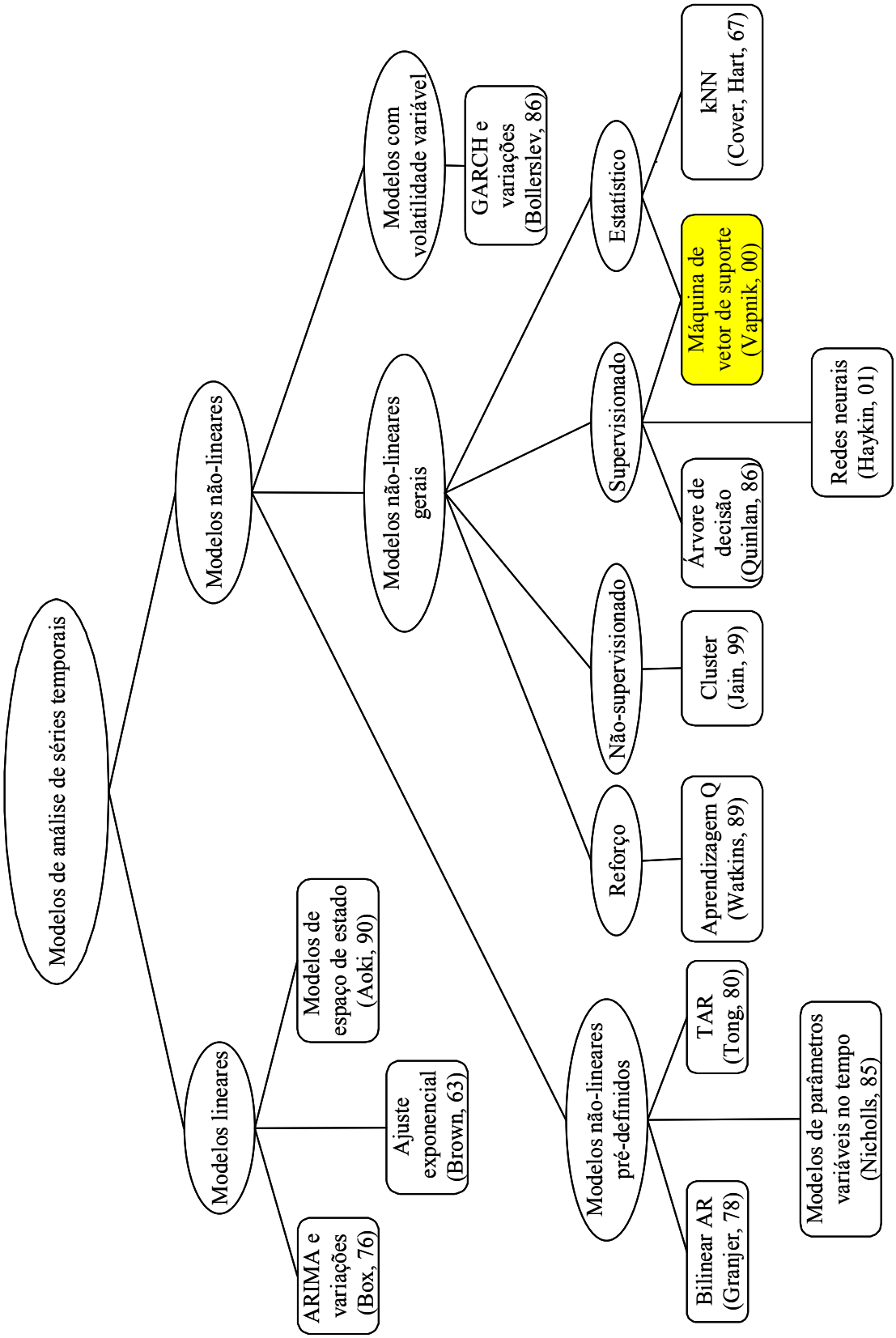


Figura 2.1 – Modelos de análise de séries temporais(Yang,2003).

2.3 Modelagem linear.

Os modelos lineares apresentam características de homogeneidade no que diz respeito á relação entre o sinal de entrada e saída, isto é, para um acréscimo no sinal de entrada corresponderá um acréscimo, de mesma proporção, ao sinal de saída.

Outra característica importante é a aditiva, que faz com que a entrada de dois ou mais sinais independentes, corresponderá na saída à soma dos dois ou mais sinais, provenientes das duas entradas, respectivamente. As duas características anteriores compõem o efeito de superposição da modelagem linear.

Uma outra característica igualmente importante é o da reprodução da saída para entradas em instantes diferentes no tempo, isto é, são invariantes no tempo.

Existem várias modelagens para sistemas lineares, dentre os quais destacamos ARMA e suas variantes, ajuste exponencial e modelagem de espaço de estado.

2.3.1 ARMA e suas variantes.

A relação geral, entre o valor atual Y_t de uma série temporal, com valores finitos observados do passado pode ser da forma (De Gooijer e Kumar, 1992) :

$$\boldsymbol{Y}_t = h(Y_{t-1}, \ldots\ldots, Y_{t-p}, \boldsymbol{\varepsilon}_{t-1}, \ldots\ldots, \boldsymbol{\varepsilon}_{t-q}) + \boldsymbol{\varepsilon}_t \qquad [2.1]$$

sendo $\boldsymbol{\varepsilon_t}$ um processo com ruído branco, isto é, uma seqüência de variáveis randômicas independentes identicamente distribuídas com média zero e variância σ^2, e h é uma função não-linear de valor real.

Expandindo segundo a série de Taylor em torno de determinado ponto fixo, podemos re-escrever na forma (Priestley, 1980 apud De Gooijer e Kumar, 1992):

$$Y_t + \sum_{i=1}^{p} \left[\phi_i(z_{t-1})\right] Y_{t-i} = \left[\mu(z_{t-1})\right] + \varepsilon_t + \sum_{j=1}^{q} \left[\theta_j(z_{t-1})\right] \varepsilon_{t-j} \qquad [2\text{-}2]$$

sendo que $z_t = \left(\varepsilon_{t-q+1},\ldots,\varepsilon_t, Y_{t-p+1},\ldots,Y_t\right)^{ts}$ é denominado vetor de estado e o expoente (*ts*) indica a transposição de sinais da função.

O modelo [2-2] apresentado acima é denominado de modelo dependente do estado, de ordem (p,q), e representa uma linearização local do modelo não linear geral, apresentado anteriormente em [2-1].

Os parâmetros desconhecidos do modelo, $\phi_i(x)\ (i=1,\ldots,p), \theta_j(x)\ (j=1,\ldots,q), \mu(x)$, são todos dependentes do estado do processo no instante de tempo $(t-1)$, e de σ^2.

Tomando-se no modelo apresentado em [2-2], e fazendo $\phi_i(x)\ (i=1,\ldots,p), \theta_j(x)\ (j=1,\ldots,q), \mu(x)$ todas constantes (isto é independentes de x), teremos o modelo denominado de ARMA(*p,q*), Modelo Auto-regressivo Média Móvel (Box e Jenkins, 1976 em Chatfield, 2003), que pode ser escrito:

$$Y_t + \sum_{i=1}^{p} \phi_i Y_{t-i} = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \qquad [2\text{-}3]$$

O modelo [2-3] pode ser separado em outros dois modelos, sendo que para *q=0* tem-se o AR(*p*) denominado Auto-regressivo de ordem *p*, e, para *p=0*, o modelo MA(*q*) denominado de Média Móvel de ordem *q*.

A modelagem ARMA(*p,q*) considera que a série temporal é estacionária, no entanto a maioria dos fenômenos que podem ser representados por séries temporais são não-estacionárias.

Com a finalidade de permitir utilizar a modelagem ARMA é necessário remover a característica de não-estacionaridade utilizando o método das diferenças, geralmente de primeira ordem, que gera uma nova série a partir da diferença de valores subseqüentes

da série original observada. Este processo é denominado ARIMA – Processo Geral Auto-regressivo Integrado de Média Móvel.

O processo ARIMA pode ser estendido para permitir a sua aplicação a séries que apresentam valores sazonais, utilizando o mesmo procedimento das diferenças, tornando uma série com sazonalidade e não-estacionária, em estacionária. Este processo é denominado SARIMA – Modelo Geral Sazonal ARIMA.

2.3.2 Ajuste exponencial (*Exponencial Smoothing*).

Técnica aplicável diretamente para o prognóstico de valores futuros da variável em estudo (Brown, 1963 apud Bedworth e Bailey, 1986). O procedimento apresenta todos os atributos da técnica de média móvel. A equação do processo de ajuste exponencial pode ser escrita:

$$\hat{Y}_t = \alpha Y_t + (1-\alpha)\hat{Y}_{t-1} \qquad [2\text{-}4]$$

sendo o termo $\hat{Y}_t$ o ajuste exponencial médio de todos os dados anteriores incluindo o do período *t*. A variável α é a constante de ajuste e assume valores no intervalo [0;1]. Expandindo a equação [2-4] para uma quantidade maior de termos observados da série temporal, e fazendo $\beta = (1-\alpha)$, teremos:

$$\hat{Y}_t = \alpha \sum_{i=0}^{t-1} \beta^i Y_{t-i} + \beta^t \hat{Y}_0 \qquad [2\text{-}5]$$

Na equação [2-5] observa-se que todos os termos medidos da série contribuem para o ajuste, e considerando que as variáveis α e β são fracionárias, os valores mais recentes observados, supostos serem os melhores valores, contribuem com uma maior parcela na formação do termo de ajuste, sendo que os demais contribuem com valores monotonamente decrescentes. Assim a contribuição do termo $\beta^t \hat{Y}_0$, para valores muito grandes de *t*, tenderá a zero.

O ajuste exponencial apresentado somente apresenta resultados satisfatórios para séries temporais que não sofrem efeitos de sazonalidade, e, que não apresentam variações sistemáticas de crescimento ou decrescimento.
Os procedimentos de prognóstico Holt e Holt-Winters de ajuste exponencial, permitem trabalhar com séries com características de variação sistemática e sazonalidade.

2.3.3 Modelagem de espaço de estados.

A modelagem de espaço de estados pode ser descrita pelo par de equações [2-6] e [2-7], denominadas de equação de transição e equação de observação, respectivamente:

$$s_t = B\ s_{t-1} + m_t \qquad [2\text{-}6]$$

$$x_t = A\ s_t + n_t \qquad [2\text{-}7]$$

Sendo que $\boldsymbol{x_t} \in \mathbb{R}^n$, t = 1,...,T são as observações no tempo t, e $\boldsymbol{s_t} \in \mathbb{R}^m$, t = 1,...,T os estados internos não observáveis do sistema dinâmico, denominados de espaço de estados.

Os vetores $\boldsymbol{m_t}$ e $\boldsymbol{n_t}$ são respectivamente, o vetor de desvios do processo e o ruído observado. Estas duas variáveis são assumidas como sendo não correlacionadas, e apresentarem valor médio nulo e variância conhecidas σ_n^2 e σ_m^2, sendo que a relação entre as variâncias $\left(\sigma_m^2/\sigma_n^2\right)$ é denominada relação sinal-ruído, e representa uma quantidade importante na determinação das características do modelo.

As matrizes **A** e **B** são respectivamente os parâmetros de linearização do mapeamento dos valores observados e dos valores prognosticados.

O filtro de Kalman (Aoki, 1990 apud Chatfield, 2003) é um dos mais conhecidos processos de modelagem de espaço de estados. Esta modelagem provê um método geral

para estimar o vetor de estados internos não observáveis s_t, que se constitui no objetivo do problema.

A essência do processo do filtro de Kalman estabelece um conjunto de equações denominadas de equações de prognóstico, as quais a partir de valores observados no período anterior estimam o valor no tempo presente, calculando em seguida o erro experimentado assim que o valor correto esteja disponível, efetuando as correções nas variáveis em estudo a partir de um conjunto de equações de correção de dados.

2.3.4 Aplicações da modelagem linear para prognóstico de demandas de potência.

A comparação entre cinco métodos (Moghram e Rahman, 1989) de avaliação da demanda de potência, para a mesma seqüência de dados da série histórica, na obtenção de prognóstico de curta duração (24 horas), não permitiu visualizar o método que apresentasse melhores resultados.

A série histórica foi dividida em valores de verão e de inverno e foram aplicados os métodos: Regressão Linear Múltipla, Série Temporal Estocástica (ARIMA), Ajuste Exponencial Geral (*Exponencial Smoothing*), Espaço de estados e filtro de Kalman e Conhecimento Baseado em Sistemas Especialistas, sendo que além dos valores históricos da série temporal, foram consideradas variáveis representativas das condições ambientais.

Dentre as aplicações efetuadas, pode-se observar que a modelagem com função de transferência(serie temporal estocástica), apresentou, para o caso de valores da série histórica do verão, erro de prognóstico médio absoluto da ordem de 0,5% e de 0,95% pára o valor de pico, mesmo tendo ocorrido erros da ordem de 2% ao longo das 24 horas de simulação do prognóstico de demanda de potência.

Este resultado é atribuído ao fato de que o processo estocástico não considera de forma relevante, mudanças grandes nas variáveis ambientais, mas é mais dependente de valores históricos da série.

Para o caso de valores da série histórica de inverno, o método que apresentou melhores resultados foi o Conhecimento Baseado em Sistema Especialistas, com um erro médio absoluto de 1,3% e de 2,18% para o valor de pico, tendo o método anterior apresentado erro médio absoluto de 2,7% e 6,75% para o valor de pico.

Podemos observar que não há coerência nos resultados de prognóstico apresentados, pois os métodos não reproduzem os mesmos resultados em situações diferentes, o que nos leva a crer que além das variáveis consideradas, outras correlações são igualmente importantes ou mesmo a utilização de modelagem não-linear.

Para levar em consideração a natureza estocástica da demanda horária de consumo de potência, foi desenvolvida modelagem (Park, Park e Lee, 1991) que considera a decomposição da série histórica da demanda em três componentes: a demanda nominal, a demanda residual e o tipo de demanda.

A demanda nominal é obtida observando-se que a série temporal de demanda dos dias úteis da semana, segue uma mesma curva de variação para qualquer dia da semana, permitindo que se componha uma representação temporal de 24 valores da série representativa do dia útil.

A demanda nominal é definida como sendo a demanda consumida sob condições normais de fatores econômicos, ciclos de trabalho e fatores metereológicos, quando não ocorrem eventos especiais de desligamento de cargas. Esta componente não tem agregada ao seu valor influências sazonais, crescimentos passados e variações das condições ambientais.

Para a curva de demanda representativa do final de semana, a série temporal é composta pela diferença entre a demanda nominal e a demanda atual daquele dia. Esta nova variável obtida da diferença mencionada, é a terceira componente, define o tipo de demanda.

A segunda componente da decomposição, a demanda residual, corresponde à modelagem de erro, que não inclui termos anteriores da série temporal, sendo identificada como sendo um processo randômico estacionário de média zero.

A extração dos valores componentes da demanda nominal foi efetuada utilizando a modelagem de variáveis de estado utilizando o modelo AR(p), sendo a estimação de estado e o prognóstico efetuado com a utilização do algoritmo do filtro de Kalman.

Uma vez efetuado o prognóstico da demanda nominal através do uso do filtro de Kalman, a modelagem do tipo de demanda pode ser efetuada com procedimentos mais simples, tendo sido utilizado o ajuste exponencial (*exponencial smoothing*), dada a sua simplicidade na aplicação e razoável acerto nos resultados.

A componente referente à demanda residual, originariamente é definida como sendo a diferença entre a demanda no tempo atual e a demanda nominal, para o caso de dias úteis, sendo obtida a partir de um modelo AR(q) e assumindo que é um processo estacionário de média zero, o parâmetro do modelo pode ser estimado pelo método recursivo dos mínimos quadrados.

O método foi aplicado para dois anos (1983 e 1987), tendo-se obtido resultados com erros relativos absolutos médios totais, para um período de 24 horas, na ordem de 1,50% e 1,40% respectivamente, sendo que os valores, de erro médio, observados para o final de semana (domingo), foram 2,41% e 2,23% respectivamente.

Metodologia aplicada na *Tókio Electric Power Company – TEPCO* utilizando formulação de regressão linear, não se mostrou adequada para todas as estações do ano, tendo sido desenvolvido nova metodologia baseada em métodos de translação e reflexão dos dados (Haida e Muto, 1994), melhorando consideravelmente o resultado do prognóstico.

A metodologia proposta utiliza dados de valores de pico de demanda, ocorridos no passado e variáveis ambientais, temperatura e umidade do ar.

O método assume que há uma relação linear entre variáveis ditas explanatórias (temperatura e umidade) e o valor de pico dentro de um período pequeno de tempo, tendo sido observado que esta afirmação é mais verdadeira nos meses com temperaturas extremas, que é o caso do verão e inverno.

Para as estações do ano em que ocorrem variações de temperatura de transição entre inverno - verão e vice-versa, que são a primavera e o outono, a relação entre o valor de pico da demanda e estas variáveis guardam uma relação de não-linearidade, tendo que ser tratada de forma diversa.

Com a finalidade de reduzir o erro de prognóstico nas estações de transição (primavera, outono), o método propõe inicialmente determinar uma função polinomial da temperatura $f_j(X_j)$, a partir dos dados do ano anterior, que utilizando o método dos mínimos quadrados determine o valor diário de pico da demanda de potência.

Esta função $f_j(X_j)$ é utilizada para determinar a modelagem de prognóstico do dia, situado no ano em estudo, utilizando os últimos dados dos dias que precedem a referida data.
A translação de dados tem a finalidade de corrigir a distância entre as temperaturas de ocorrência do menor valor do pico de demanda, que necessariamente não ocorrem na mesma temperatura, verificadas nos dois anos subseqüentes da base de dados tomados para o estudo.

A reflexão dos dados determina o menor valor da função de transformação $f_j(X_j)$, situado entre os dados de temperatura coletados entre o verão e o inverno, a partir da derivação da função polinomial.

O ponto encontrado divide a reta de regressão linear (temperatura x pico de demanda de potência) em dois conjuntos de valores, demandas para temperaturas mais quentes e demandas para temperaturas mais frias, que são situações características das estações de transição, onde normalmente não ocorrem temperaturas extremas.

A partir desta constatação é efetuada a correção do valor de pico de demanda de potência e da função de transformação da temperatura, e desta forma apresentar resultados melhores nos períodos situados nas estações de transição.

Prognósticos efetuados, para os anos de 1989 a 1992, resultaram para valores médios absolutos do erro, utilizando regressão linear simples, função de transformação sem translação e reflexão e utilizando a função de transformação com translação e reflexão, 1,93%, 1,70% e 1,68% para a primavera e 2,16%, 1,55% e 1,43% para o outono, respectivamente.

No passado, os métodos de prognóstico levavam em consideração a experiência adquirida pelos operadores dos sistemas elétricos. Esta experiência pode ser de grande valia para a estimação de valores iniciais de métodos paramétricos. Uma metodologia desenvolvida utilizando este conceito, foi, a apresentação de uma modelagem ARIMA – *Autoregressive Integrated Moving Average*, modificada para atender os requisitos de utilização de valores estimados pelos operadores como dados de entrada (Amjady, 2001).

Do ponto de vista matemático, desde que o valor estimado (pelos operadores) da variável de saída possa apresentar uma correlação com o valor real da saída, a utilização desse termo poderá aumentar a capacidade de prognóstico do procedimento modificado ARIMA, comparativamente com o procedimento ARIMA normal.

Resultados obtidos confirmam a melhora do processo com a utilização do procedimento ARIMA modificado, tendo apresentado erros absolutos médios e de pico para a demanda horária e de pico variando entre 1,45% a 1,99% e 4,15% a 5,05% respectivamente, considerados vários dias da semana incluindo feriados públicos e sexta-feira (final de semana no mundo islâmico), e aplicados em dias com estações quentes e frias.

Os mesmos dados aplicados a uma metodologia ARIMA normal, apresentaram erros absolutos médio e de pico, da ordem de 50% a 115% e 38% a 98% superiores, respectivamente, para os mesmos pontos calculados anteriormente com o método ARIMA modificado.

Considerando os valores prognosticados pela experiência dos operadores, chega-se a resultados com erros absolutos médio e de pico, da ordem de 86% a 153% e 69% a 125% superiores, respectivamente, relativamente ao procedimento ARIMA modificado.

Com a crescente desregulamentação do setor elétrico, a energia elétrica tornou-se um produto comerciável através de bolsa de valores, ou através de empresas especializadas.

Torna-se então necessário determinar as demandas de energia excedentes que possam ser oferecidas ao mercado consumidor, bem como também prognosticar valores futuros do preço desta energia, informação esta importante para o mercado produtor e consumidor, para subsidiar o planejamento das estratégias de oferta, de forma a maximizar seus benefícios e instalações, respectivamente.

Pesquisa desenvolvida nesse sentido (Nogales, Contreras, Conejo e Espínola, 2002), baseou-se em dois procedimentos: método de regressão dinâmica e o método da função de transferência.

Ambos os métodos correlacionam o preço atual com valores anteriores do preço e das demandas de potência, mantendo uma certa similaridade com os métodos ARMA.

Resultados obtidos de aplicações a dois mercados de energia (Espanha e Califórnia EUA), apresentaram erro médio absoluto diário variando entre 3,8% a 8,7% e 3,6% a 8,3%, quando aplicado para uma semana do mês de Agosto na Espanha, e 1,9% a 4,0% e 2,2% a 3,7%, quando aplicado a uma semana de Abril na Califórnia, utilizando os métodos de regressão dinâmica e função de transferência, respectivamente.

2.4 Modelagem não linear.

Apesar de que muitos modelos estudados dizem respeito ao comportamento linear das variáveis em estudo, não há razão para crer que os processos na natureza apresentem na sua maioria um comportamento linear (Chatfield, 2003). As razões, pelas quais os estudos prendem-se mais à modelagem linear, provavelmente estão relacionadas à facilidade computacional e a formulação matemática mais simples, o que não ocorre para a modelagem de fenômenos não lineares, observados na natureza.

Conforme a Figura 2.1, considerar-se-á três tipos de modelagens não lineares:

-Modelos não-lineares pré-definidos;

-Modelos não-lineares com volatilidade variável;

-Modelos não-lineares gerais.

2.4.1 Modelos não lineares pré-definidos.

Muitos modelos foram originados da modelagem linear AR(p), tendo-se proposto modelos com mudança a partir da introdução de parâmetros variáveis no tempo, utilizando funções determinísticas ou estocásticas ou mesmo determinadas de alguma forma a partir de dados ocorridos no passado.

A modelagem AR(*p*), com coeficientes dependentes do tempo, é obtida fazendo os parâmetros de autoregressão, funções determinísticas do tempo. Os processos originados desta modelagem são a modelagem de parâmetros variáveis no tempo (Nicholls e Pagan, 1985 apud Chatfield, 2003) e a modelagem autoregressiva de coeficientes randômicos (Nicholls e Quinn, 1982 apud De Gooijer e Kumar, 1992).

Outra modelagem, denominada bilinear (Granger e Andersen, 1978 apud De Gooijer e Kumar, 1992), pode ser obtida considerando na equação [2-2] $\mu(x)$ *e* $\phi_i(x)(1=1,...,p)$ constantes e $\theta_j(z_{t-1}) = \theta_j + \sum_{v=1}^{Q} c_{jv} Y_{t-v} (j=1,...,q)$:

$$Y_t + \sum_{i=1}^{p} \phi_i Y_{t-i} = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{j=1}^{q} \sum_{v=1}^{Q} c_{jv} Y_{t-v} \varepsilon_{t-j} \qquad [2\text{-}8]$$

A modelagem bilinear completa pode ser obtida efetuando-se algumas transformações na equação [2-8] e introduzindo $p = q = 0$ e $\mu=0$:

$$Y_t = \sum_{u=1}^{P}\sum_{v=1}^{Q} c_{uv} Y_{t-v} \varepsilon_{t-u} \qquad [2\text{-}9]$$

Uma outra modelagem considera a aplicação de finitas possibilidades para a modelagem AR, a partir da mudança de parâmetros em cada instante no tempo de passagem de um membro para outro da série, definido por uma constante pré-determinada.

A essência do modelo é a linearização de trechos do espaço de estado do modelo não-linear, através da introdução de pontos de mudança –*threshold* – a partir do qual os parâmetros do equacionamento são alterados para atender novas condições da série em estudo.

O modelo denominado SETAR- *Self-Exciting Threshold Autoregressive Model* – apresentou generalização suficiente para capturar certas características não possíveis de serem obtidas com modelos lineares (Tong e Lim, 1980 apud De Gooijer e Kumar, 1992). Fazendo na equação [2.2], $\mu(x) = \phi_0^{(j)}, \phi_j(x) = 0$ *e* $\phi_i(z_{t-1}) = \phi_i^{(j)}$ sendo $Y_{t\text{-}d} \in \mathrm{R}^{(j)} (i = 1,..,p; j = 1,..,\ell)$ onde *d* é um inteiro positivo e $\mathbb{R}^{(j)}$ é um conjunto de números reais que indica os pontos de mudança ("*threshold*"), pode-se escrever o modelo SETAR(ℓ*; p,...,p*):

$$Y_t + \phi_0^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} Y_{t-i} = \varepsilon_t^{(j)} \qquad [2\text{-}10]$$

2.4.2 Modelos não lineares com volatilidade variável.

Os modelos não lineares descritos anteriormente (AR(*p*), bilinear e SETAR) podem ser descritos como tendo não-linearidade estruturada (Chatfield, 2003).

Os modelos ditos de volatilidade variável correspondem aos modelos com variações na sua variância, sendo o mais conhecido o modelo ARCH – *Autoregressive Conditional Heteroscedastic Model.*

O modelo ARCH(*q*) de ordem *q* é dado pela relação(Engle, 1982 apud De Ggooijer e Kumar, 1992):

$$Y_t = \varepsilon_t \sigma_t = \varepsilon_t \left[\alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \right]^{1/2} \qquad [2\text{-}11]$$

A não linearidade deriva da variância condicionada (termo entre colchetes da equação [2-11]) de $\{\varepsilon_t\}$, porque nos demais modelos não lineares, a condicionalidade do meio varia com o tempo.

A generalização do modelo, GARCH – *Generalized Autoregressive Conditional Heteroscedastic Model* (Bollerslev, 1986 apud De Ggooijer e Kumar, 1992), se dá para o caso em que a variância, além de ser função de valores anteriores de $\{\varepsilon_t\}$ como na equação [2-11], se apresenta como:

$$Y_t = \varepsilon_t \sigma_t = \varepsilon_t \left[\alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \right]^{1/2} \qquad [2\text{-}12]$$

que é o modelo GARCH de ordem *(p,q),* onde $\alpha_0 \geq 0$ e α_i , $\beta_j \geq 0$ para todo *i,j*. Identificar apropriadamente um modelo GARCH não é uma tarefa fácil, sendo que muitos analistas assumem GARCH(1,1) como modelo padronizado.

2.4.3 Modelos não lineares gerais.

Um modelo classificado nesta categoria é a modelagem de aprendizagem por reforço, que baseado na Fig. 2.2 apresenta a seguinte funcionalidade: Um agente é conectado ao ambiente, capaz de percebê-lo e tomar determinadas ações (Kaelbling, Littman e Moore, 1996). Em cada passo da iteração o agente *"B"* recebe como entrada, "*i*" que

corresponde a alguma indicação do estado atual do ambiente "*s*"; o agente então gera uma ação "*a*" na saída. A ação modifica o estado do meio e o valor dessa transição de estado é fornecida ao agente através de um sinal de reforço "*r*".

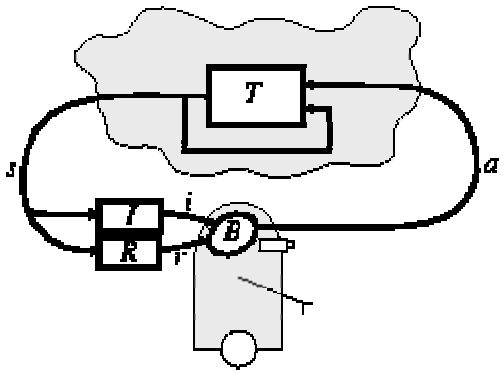


Figura 2.2 – Modelagem padrão de aprendizagem por reforço

(Kaelbling, Littman e Moore, 1996).

O comportamento do agente "*B*" é gerar ações que tendem a incrementar os valores do sinal de reforço, provenientes da interação com o ambiente. Dessa forma o agente pode aprender, ao longo do tempo, através da sistemática de tentativa e erro.

O problema consiste em determinar um procedimento ótimo que leve a uma solução, baseado na experimentação de várias seqüências possíveis de ações, observando os resultados ocorridos. Tal procedimento existe, e é, encontrado, em um processo estocástico denominado aprendizagem *Q* (Watkins, 1989 apud Haykin, 2001).

O fator Q, para cada estado "*i*" e ação "*a*", é definido como o *custo imediato mais a soma dos custos descontados de todos os estados sucessores que seguem a política,* definida como sendo um *mapeamento de estados para ações* (Haykin, 2001).

O equacionamento do valor de custo ótimo $\mathbf{Q^*(i,a)}$, que segue o critério de otimização de Bellman (Bellman, 1957), para um estado "*i*" e uma ação "*a*" é apresentado na equação [2-13] :

$$Q^*(i,a) = \sum_{j=1}^{N} p_{ij}(a)\left[g(i,a,j) + \gamma \min_{b \in s} Q^*(j,b) \right] \text{ para todo } (i,a) \qquad [2\text{-}13]$$

onde a função "*p*" representa a probabilidade da transição do estado "*i*" para o estado "*j*" dado que ocorreu a ação "*a*", a função "*g*" é o custo imediato da transição para passar do estado "*i*" para o estado "*j*", o fator de desconto varia de $0 \leq \gamma < 1$ e os custos $Q^*(j,b)$ representam os custos ótimos de todas as possíveis ações que podem suceder a ação "*a*".

Outra categoria de modelagem não-linear é constituída pelos métodos não supervisionados de classificação, que efetuam a separação de dados em agrupamentos (*clusters*), seguindo algum critério de proximidade dos padrões a serem analisados.

Dentre os métodos mais difundidos e mais simples de ser utilizado (Jain, Murty e Flynn, 1999) encontra-se o *k*-means que utiliza o critério do quadrado do erro para proximidade dos padrões, onde para o agrupamento $\boldsymbol{\Psi}$ do conjunto de padrões $\boldsymbol{\Omega}$ (contendo $\boldsymbol{K}$ grupos) é equacionado como na equação [2-14], onde o erro $e^2(\Omega,\Psi)$ é calculado em função do padrão x_i^j de entrada pertencente ao agrupamento $\boldsymbol{j}$, sendo c_j a coordenada do centróide do mesmo grupo.

$$e^2(\Omega,\Psi) = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|x_i^j - c_j\right\|^2 \qquad [2\text{-}14]$$

O algoritmo inicia com uma distribuição aleatória dos padrões de entrada, e efetua a localização do mesmo baseado na proximidade com o centróide até a convergência do critério (não há mais posicionamento de padrões para outros agrupamentos ou o erro cessa de decrescer).

O algoritmo *k*-means apresenta uma sensibilidade elevada à seleção do valor inicial do processo, e pode convergir para a um mínimo local se a partição inicial não for apropriada. Existem processos que permitem a identificação de uma melhor escolha do centróide inicial que serão apresentados no decorrer do trabalho.

As redes neurais artificiais, também se apresentam como sendo uma ferramenta poderosa para a identificação de padrões. A rede de Kohonen (Kohonen, 1990), apresentada juntamente com as redes neurais na seqüência, é o método equivalente ao algoritmo *k*-means e tem sido aplicada para grandes conjuntos de dados.

Uma outra forma possível de classificação leva em consideração a natureza estatística da classificação dos padrões segundo uma coleção de exemplos.

O processo de classificação denominado *k*-NN *(k – Nearest Neighbor)* (Cover e Hart, 1967), assume que a coleção de exemplos ($\mathbf{x_i}$,$\boldsymbol{\theta_i}$), onde $\mathbf{x_i}$ são as observações e $\boldsymbol{\theta_i}$ são as categorias a que pertence o padrão observado, são valores independentes e identicamente distribuídos, segundo uma distribuição (x,θ).

Dado um conjunto de ***n*** pares (x_1,θ_1),....., (x_n,θ_n), em que os valores x_i's assumem o valor **d,** em um espaço métrico **X,** e os θ_i's assumem as categorias {1,2,...M}, um novo par (x, θ), onde só é observável a medida x, na qual quer se determinar a categoria a que pertence, utilizando o conjunto de pontos classificados corretamente, teremos $x_i^t \in \{x_1, x_2, \ldots, x_n\}$ como ponto mais próximo de **x,** se min $d(x_i, x) = d(x_n^t, x)$, fazendo com que a observação tenha a mesma distribuição probabilística na sua categoria de classificação.

A árvore de decisão é uma metodologia classificada como supervisionada, e se baseia, a partir de uma questão inicial, na procura da pergunta que possa melhor discriminar a direção da resposta procurada.

O processo pode ser binário, conforme Figura 2.3, onde cada nó abre duas novas possibilidades, geralmente como resposta sim ou não (Berry e Linoff, 1997). Este processo é repetido até que chegue aos nós mais afastados da origem, onde não é mais possível a formulação de pergunta para a divisão dos dados, onde geralmente se encontram os resultados procurados.

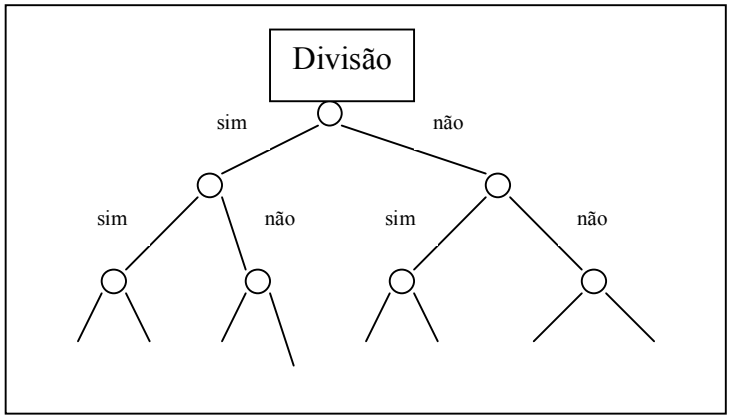


Figura 2.3 – Representação gráfica de uma árvore de decisão binária.

A árvore de decisões pode assumir divisão ternária, assim a pergunta formulada no nó conduzirá a uma resposta mais concreta sobre a base de dados em estudo.

Um dos algoritmos mais utilizados é denominado CART – *Classification and Regression Trees,* (Breiman, Friedman, Olshen e Stone, 1984 apud Haykin, 2001), que constrói uma árvore binária, iniciando o processo com resultados pré-classificados utilizados como conjunto de treinamento.

A finalidade do processo prévio é determinar qual a divisão inicial que melhor poderá conduzir ao resultado esperado, sendo que o potencial da divisão é avaliado por um *índice de diversidade* do conjunto de resultados.

A melhor divisão inicial é aquela que apresenta os menores índices de diversidade dos conjuntos de resultados, conseqüentemente maximiza a função:

Diversidade(antes da divisão) – {diversidade(resposta sim) + diversidade(resposta não)}

Outro algoritmo utilizado no processo de árvore de decisão é o C4.5 (Quinlan, 1986 apud Yang, 2003), cujo processo é similar ao CART quanto à divisão dos dados em conjuntos, para verificar o critério de inicialização.

O critério utilizado para verificar o acerto inicial está baseado na relação de ganho, que é uma informação que leva em consideração diferentes valores de resultados de testes. Assim se C for o número de classes e $p(D,j)$ a proporção de casos em D que pertencem à classe j, a incerteza residual a respeito da classe na qual D pertence pode ser expressa por(Quinlan, 1996):

$$Info(D) = -\sum_{j=1}^{C} p(D,j) \ \log_2 [p(D,j)] \qquad [2\text{-}15]$$

e o correspondente ganho de informação para o teste *T* com *k* resultados, na [2-16].

$$Gain(D,T) = Info(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \; Info(D_i) \qquad [2\text{-}16]$$

O critério utilizado para verificar o acerto no valor inicial está baseado na relação de ganho, dado pela equação [2-17], onde *Split(D,T)* corresponde à informação sobre a divisão do conjunto *D***,** e *T* são os testes com *k* resultados e D_i são subconjuntos de *D***.**

$$Split(D,T) = -\sum_{i=1}^{k} |D_i/D| \log_2 \left| \frac{D_i}{D} \right| \qquad [2\text{-}17]$$

Uma metodologia que pode ser classificada como sendo estatística e ao mesmo tempo ser classificada como de aprendizagem supervisionada é a máquina de vetor de suporte (Vapnik, 2000), que será apresentada no capítulo III.

2.4.4 Modelos não lineares gerais: Redes neurais.

As redes neurais artificiais provêm da observação do comportamento biológico dos neurônios no mundo real, onde nem todos executam exatamente as mesmas funções, ou exatamente da mesma forma. Apesar, de não apresentarem comportamentos similares, existem propriedades que podem ser atribuídas a todos os neurônios (Kartalopoulos, 1996):

1. *Existem muitas conexões em paralelo entre muitos neurônios.*
2. *Muitas conexões paralelas providenciam mecanismos de realimentação a outros neurônios e a si mesmo.*
3. *Alguns neurônios excitam outros neurônios enquanto inibem a operação de outros.*
4. *Algumas partes da rede estão pré-definidas, enquanto outras partes estão em evolução ou em desenvolvimento.*
5. *A saída não é necessariamente de natureza binária, isto é, 1- 0 ou sim/não.*
6. *A operação das redes neurais ocorre de forma assíncrona.*

7. *As redes neurais apresentam, um mecanismo robusto e lento de sincronismo, mais lento que as pulsações cardíacas que suportam as suas funções vitais.*
8. *As redes neurais executam um programa que é distribuído de forma completa e não de forma seqüencial.*
9. *As redes neurais não se apresentam com um processador central, mas sim o processamento é distribuído.*

A Figura 2.4 ilustra o modelo básico de um neurônio e a correspondente rede neural artificial, sendo que os valores x_{ik} correspondem às entradas e q_k à saída.
O sinal y_k é resultante da condição de disparo do elemento Σ, e corresponde ao resultado de $\Sigma\, w_{ik}\, x_{ik}$, que por sua vez aciona a função de ativação não linear, que pode ser uma função limiar, sigmoidal ou outras, que define a saída da rede, q_k.

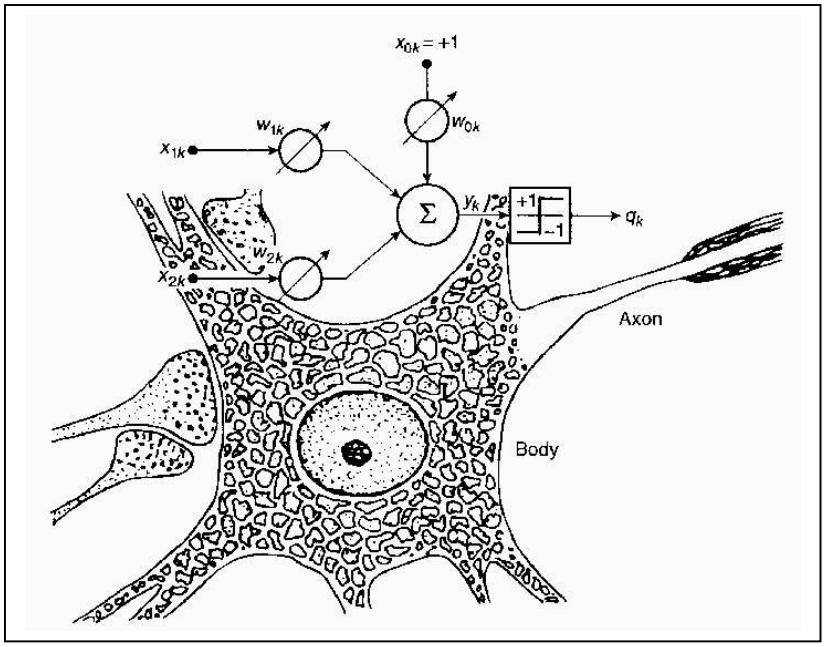


Figura 2.4 – Modelo básico de neurônio e a rede neural correspondente (Kartalopoulos, 1996).

Podemos definir uma rede neural vista como uma máquina adaptativa (Aleksander e Morton, 1990 apud Haykin, 2001):

Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

1. *O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.*
2. *Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.*

Existem características básicas que devem ser especificadas para cada rede neural artificial para permitir o seu funcionamento. Podemos identificar cinco características que definem a arquitetura (Pfeifer, 2000):

(1) *A característica do nó (neurônio)* – define como o nó efetua as operações de soma das entradas, como efetua a transformação através da função de ativação e como são tratados esses dados e como são transformados em dados de saída e transmitidos pelo axônio.

(2) *A conectividade* – especifica quais nós estão interligados e em que direção.

(3) *A regra de propagação* – especifica como um sinal ativado que trafega pelo axônio é transmitido aos neurônios, aos quais estão conectados.
As regras de propagação definem a topologia das redes que podem ser classifica-las em redes *alimentadas adiante*, em que não ocorre a existência de ciclos de realimentação, e as redes recorrentes ou *feedback*, em que existem ciclos de realimentação, conforme mostrado na Figura 2.5.

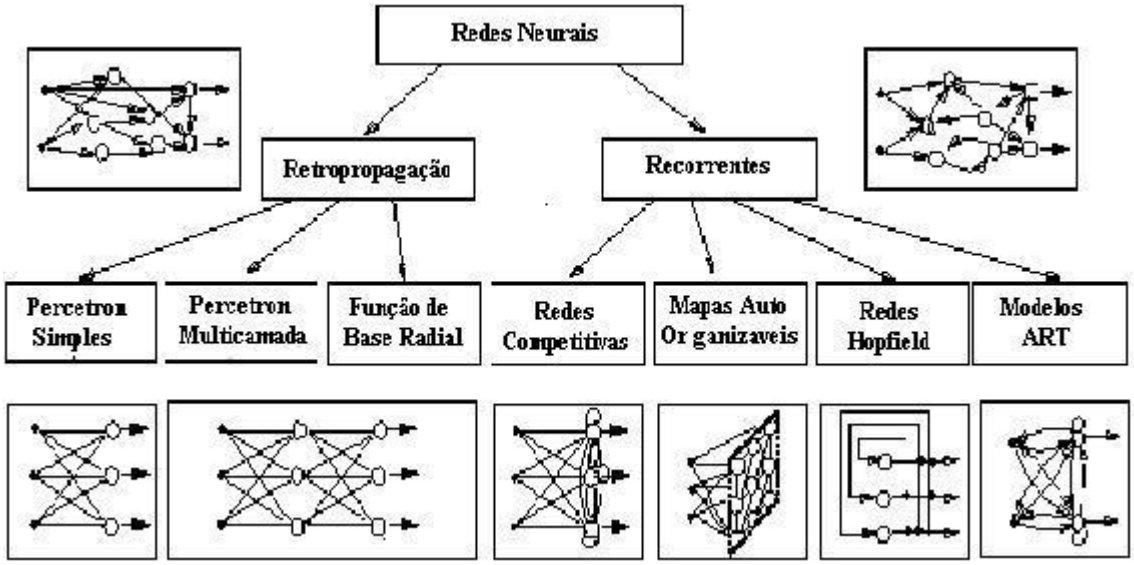


Figura 2.5 – Divisão das redes neurais segundo a direção do fluxo de sinais
(Jain, Mao e Mohiuddin, 1996)

(4) *As regras de aprendizagem* – especificam como a intensidade das conexões entre neurônios, sofrem mudanças com o tempo.
Basicamente podem ser definidos três tipos de regras de aprendizagem: supervisionada, com reforço e não supervisionada.

(5) *Como introduzir a rede em um sistema físico* – especifica o sistema que faz a aquisição de dados físicos para a rede e as conexões necessárias para trabalho em tempo real.

-Perceptrons

O perceptron na sua forma original (Rosenblatt, 1958 apud Haykin, 2001) foi desenvolvido com base no modelo de McCulloch-Pitts (McCulloch e Pitts, 1943 apud Kartalopoulos, 1996), que é um modelo que apresenta um combinador linear (somatória dos produtos dos sinais de entrada com os pesos sinápticos), seguido por um limitador abrupto que realiza a função sinal, conforme mostrado na Figura 2.6.
O perceptron é um sistema de classificação de padrões que utiliza aprendizagem supervisionada para definição dos pesos sinápticos, necessitando de valores de entrada e saída conhecidos, para permitir o processamento do treinamento.

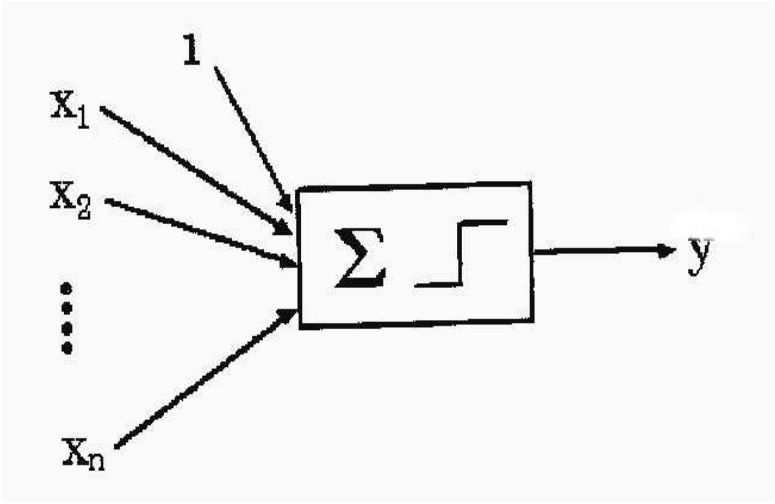


Figura 2.6 – Perceptron classificador (Fonte: Wolfram Research).

As redes de múltiplas camadas alimentadas adiante constituem uma classe importante de redes neurais. São as redes mais utilizadas em inúmeras aplicações e são conhecidas por perceptrons de múltiplas camadas (MLP, *multilayer perceptron*), conforme mostrado na Figura 2.7, representando uma generalização do perceptron de camada única.

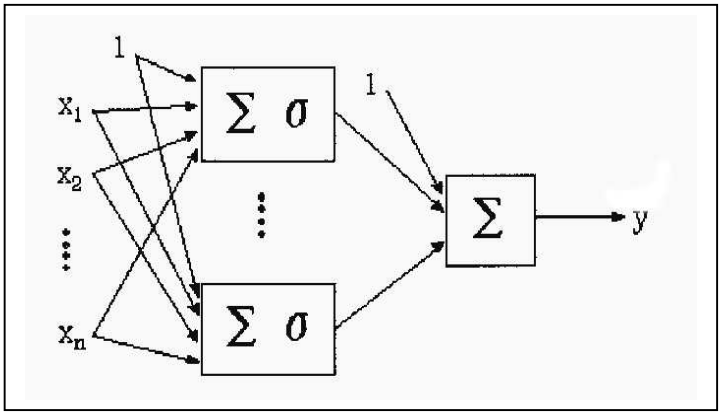


Figura 2.7 – Perceptron com uma camada oculta e uma saída
(Fonte: Wolfram Research).

O algoritmo de retropropagação(*backpropagation*)(Werbos, 1974; Parker, 1982; Rumelhart, Hinton e Williams, 1986a, 1986b apud Simpson 1992) , consiste basicamente em dois passos : passo para frente , a *propagação* , onde o sinal de entrada

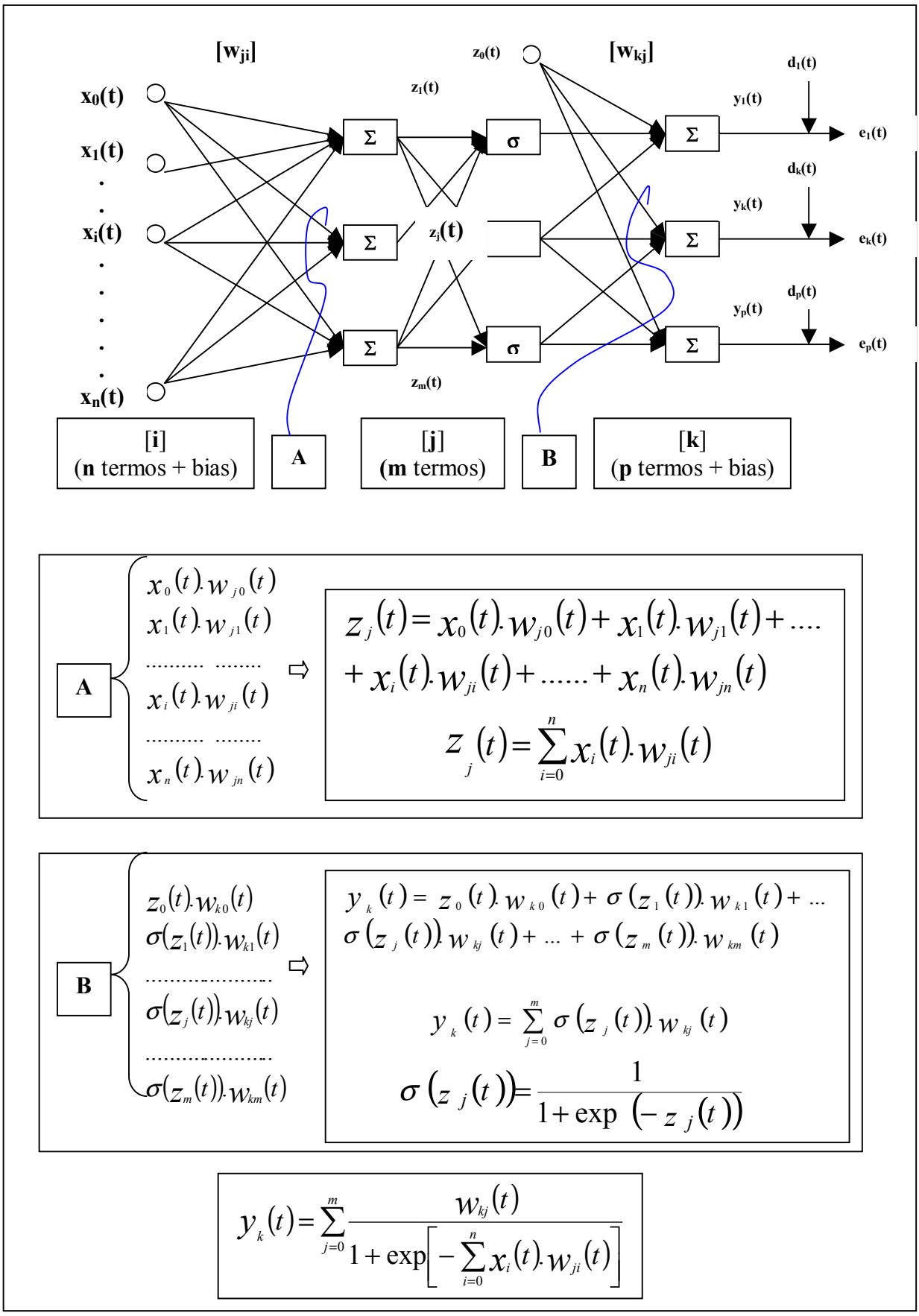
se propaga através da rede e produz uma resposta real na saída. A resposta real é comparada com a resposta desejada, produzindo um sinal de erro, o qual se propaga na direção da entrada, passo para trás, a *retropropagação*, corrigindo os pesos sinápticos.

Na Figura 2.8 está representado o equacionamento do passo de propagação do sinal de entrada, do algoritmo de retropropagação, sendo que o vetor de entrada $\mathbf{x}=[x_0(t),x1(t),x_2(t)....,x_i(t)...., x_n(t)]$ se propaga através da rede, passando na camada oculta *j* produzindo um sinal de saída $\sigma(z(t))$ ao passar pela função de ativação σ, apresentando uma resposta real *y* na camada de saída *k*, onde comparada com a resposta desejada *d*, resulta um sinal de erro ***e*.**

Na Fig 2.9 está apresentado o equacionamento da fase de retropropagação. A correção do erro *e* é efetuada através da função energia do erro *E*, que representa uma medida do desempenho de aprendizagem.
A correção Δw a ser aplicada aos pesos é equacionada a partir da regra Delta, que é obtida pela primeira derivada da função de custo *E* relativamente aos pesos e bias, que é função do gradiente local δ, o qual busca uma direção para mudança do peso que reduza o valor de *E*.

O algoritmo de retropropagação apresenta várias propriedades que o tornam atrativo. Dentre as diversas propriedades, a detecção no de espaço de entrada, realizada pela transformação não-linear que ocorre na camada de neurônios ocultos, de características salientes dos dados de treinamento(Haykin, 2001).

Uma propriedade que não está explicitamente programada no modelo, mas, no entanto, é resultado do processamento maciçamente em paralelo, é a tolerância a erros e a ruídos, que na sua ocorrência provocam somente uma degradação gradual na performance da rede, permitindo a re-aprendizagem a partir da apresentação de novos valores de entrada, reconduzindo a rede ao mesmo nível de desempenho(Pfeifer, 2000).



$$x_0(t).w_{j0}(t)$$
$$x_1(t).w_{j1}(t)$$
$$\ldots\ldots\ldots\ldots$$
$$x_i(t).w_{ji}(t)$$
$$\ldots\ldots\ldots\ldots$$
$$x_n(t).w_{jn}(t)$$

A ⇨

$$z_j(t) = x_0(t).w_{j0}(t) + x_1(t).w_{j1}(t) + \ldots. + x_i(t).w_{ji}(t) + \ldots\ldots + x_n(t).w_{jn}(t)$$

$$z_j(t) = \sum_{i=0}^{n} x_i(t).w_{ji}(t)$$

$$z_0(t).w_{k0}(t)$$
$$\sigma(z_1(t)).w_{k1}(t)$$
$$\ldots\ldots\ldots\ldots$$
$$\sigma(z_j(t)).w_{kj}(t)$$
$$\ldots\ldots\ldots\ldots$$
$$\sigma(z_m(t)).w_{km}(t)$$

B ⇨

$$y_k(t) = z_0(t).w_{k0}(t) + \sigma(z_1(t)).w_{k1}(t) + \ldots \sigma(z_j(t)).w_{kj}(t) + \ldots + \sigma(z_m(t)).w_{km}(t)$$

$$y_k(t) = \sum_{j=0}^{m} \sigma(z_j(t)).w_{kj}(t)$$

$$\sigma(z_j(t)) = \frac{1}{1+\exp(-z_j(t))}$$

$$y_k(t) = \sum_{j=0}^{m} \frac{w_{kj}(t)}{1+\exp\left[-\sum_{i=0}^{n} x_i(t).w_{ji}(t)\right]}$$

Figura 2.8 - Fase de Propagação(Haykin, 2001).

$$Erro\ \ na\ \ saída\ \ do\ \ nó\ \ k \Rightarrow e_k(t) = d_k(t) - y_k(t)$$

$$Função\ \ energia\ \ do\ \ erro \Rightarrow E(t) = \frac{1}{2}\sum_{k=1}^{p} e_k^2(t) = \frac{1}{2}\sum_{k=1}^{p}\left[d_k(t) - y_k(t)\right]^2$$

$$Derivada\ \ da\ \ função\ \ energia = \frac{\partial E(t)}{\partial w_{ji}(t)} = \frac{\partial E(t)}{\partial e_k(t)} \cdot \frac{\partial e_k(t)}{\partial y_k(t)} \cdot \frac{\partial y_k(t)}{\partial z_j(t)} \cdot \frac{\partial z_j(t)}{\partial w_{ji}(t)}$$

$$\frac{\partial E(t)}{\partial e_k(t)} = \frac{\partial}{\partial e_k(t)}\left[\frac{1}{2}\sum_{k=1}^{p} e_k^2(t)\right] = e_k(t) \qquad \frac{\partial e_k(t)}{\partial y_k(t)} = \frac{\partial}{\partial y_k(t)}\left[d_k(t) - y_k(t)\right] = -1$$

$$\frac{\partial y_k(t)}{\partial z_j(t)} = \frac{\partial}{\partial z_j(t)}\left[\varphi_j(t)\right] = \varphi'_j(t) \qquad \frac{\partial z_j(t)}{\partial w_{ji}(t)} = \frac{\partial}{\partial w_{ji}(t)}\left[\sum_{i=0}^{n} x_i(t) \cdot w_{ji}(t)\right] = x_i(t)$$

$$\frac{\partial E(t)}{\partial w_{ji}(t)} = -e_k(t) \cdot \varphi'_j(t) \cdot x_i(t)$$

$$Regra\ \ Delta \Rightarrow \Delta w_{ji}(t) = -\eta \frac{\partial E(t)}{\partial w_{ji}(t)} = \eta \cdot e_k(t) \cdot \varphi'_j(t) \cdot x_i(t) = \eta \cdot \delta_j(t) \cdot x_i(t)$$

$$\begin{pmatrix} Correção \\ do\ \ peso \\ \Delta w_{ji}(t) \end{pmatrix} = \begin{pmatrix} Parâmetro\ \ de \\ aprendizagem \\ \eta \end{pmatrix} \cdot \begin{pmatrix} Gradiente \\ local \\ \delta_j(t) \end{pmatrix} \cdot \begin{pmatrix} Sinal\ \ de\ \ entrada \\ do\ \ neurônio\ \ j \\ x_i(t) \end{pmatrix}$$

Figura 2.9 - Fase de retropropagação(Haykin, 2001).

O processo de treinamento de uma rede pode ser visto como um ajuste de curva. Uma rede neural que recebe poucos exemplos de entrada, durante o processo de aprendizagem, apresentará *pouco ajuste* dos pontos discretos quando da entrada de valores novos na rede, conforme mostrado na Figura 2.10. Da mesma forma se for apresentado para a rede um número excessivo de valores ou quando a rede é treinada em excesso, tende a memorizá-los, apresentando um *excesso de ajuste*, conforme mostrado na Figura 2.10, caracterizando a perda da propriedade de generalização(Lawrence, Giles e Tsoi, 1996).

A propriedade de generalização é uma propriedade inteligente do sistema capaz de dar resposta ao mundo real, apresentando respostas similares para entradas similares.

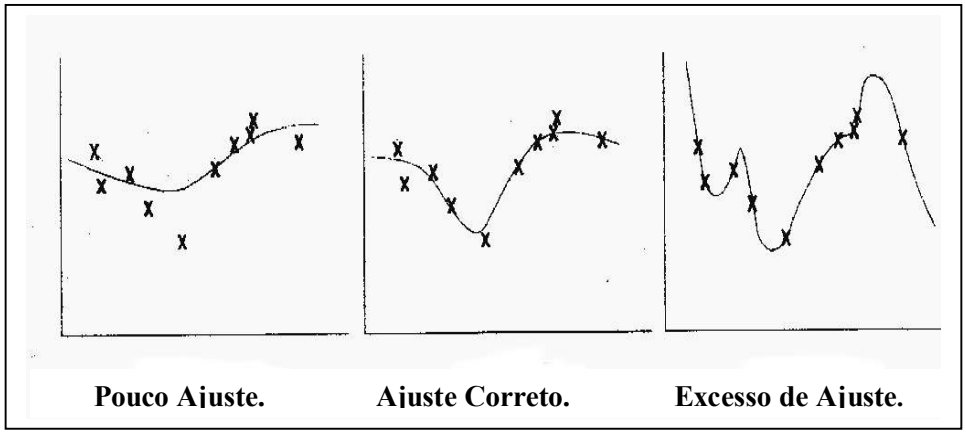


Figura 2.10 – Ajuste de pontos discretos (Lawrence, Giles e Tsoi, 1996).

Em uma pesquisa de ajuste de funções com perceptrons de duas camadas(Lawrence, Giles e Tsoi, 1997), com variações no número de neurônios da camada oculta, não foi verificado um elevado grau de *ajuste em excesso,* tendo sido observado, que as redes com um número maior de neurônios na camada oculta, podem as vezes, melhorar o desempenho no treinamento reduzindo o erro de generalização.

O acréscimo no grau de liberdade melhora a convergência do sistema para um mínimo global, isto é, o acréscimo de parâmetros diminui a chance do sistema convergir para mínimos locais ou platôs(Kröse e Smagt, 1993 apud Lawrence, Giles e Tsoi, 1997).

-Redes de funções de base radial.

Da teoria de aproximação de funções em um espaço de alta dimensionalidade deriva a idéia de redes de função de base radial(Powell, 1985 apud Haykin, 2001), que provem da utilização, na sua camada oculta, de funções que formam uma base arbitrária para os padrões, denominadas de *funções de base radial*(Bullinaria, 2003).

A rede de funções de base radial, conforme Figura 2.11, na sua forma natural é constituída por três camadas com funções completamente diferentes: 1)A camada de entrada constituída por nós de fonte, que conectam a rede ao seu ambiente; 2)A segunda camada, que é a camada oculta, geralmente de alta dimensionalidade, constituída por neurônios que aplicam uma transformação não linear do espaço de entrada para o espaço oculto; 3)A camada de saída, geralmente com um neurônio, fornece uma resposta linear, à função de ativação (função de base radial) aplicada aos sinais de entrada.

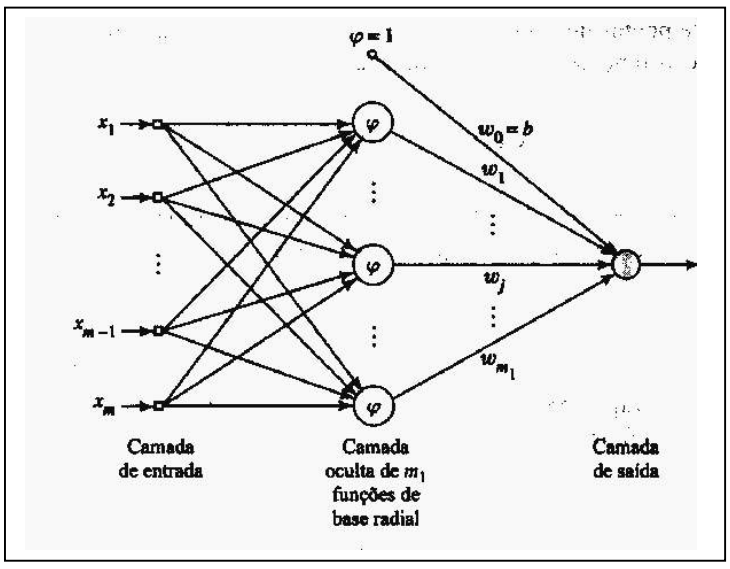


Figura 2.11 – Rede de função de base radial(Haykin, 2001).

A solução de problemas complexos de classificação de padrões utilizando uma rede de funções de base radial, necessita efetuar uma transformação não linear para um espaço de alta dimensionalidade, o que está de acordo com o *teorema de Cover* da

separabilidade de padrões (Cover, 1965 apud Haykin, 2001), que pode ser formulado como:

Um problema de classificação de padrões dispostos não linearmente em um espaço de alta dimensionalidade tem maior probabilidade de ser linearmente separável do que em um espaço de baixa dimensionalidade.

O problema a ser resolvido é determinar a geometria da superfície(hiper-superfície), dada pela função $F(x_e)=y_e \quad (e=1,2,....N)$, que satisfaça a condição de passar por todos os pontos $\{(x_e, y_e)\}_{e=1}^{N}$ do conjunto de dados conhecidos, endereçando o problema à teoria de interpolação multivariada.

A técnica de funções de base radial sugere a função de interpolação da seguinte forma(Powell, 1988 apud Haykin, 2001), (Nikolaev, 2003):

$$F(x)=\sum_{i=1}^{N} w_i \varphi(\|x-x_i\|) \qquad [2\text{-}18]$$

onde $\varphi(\|x-x_i\|)$ é o conjunto de funções não-lineares de base radial, x_i são os centros dessas funções interpolantes e $\|x-x_i\|$ é uma norma, geralmente euclidiana.

Podemos representar a equação [2-18] através de um conjunto de equações lineares simultâneas, definindo os vetores **[*F(x)*]=[*d*]** (vetor de respostas desejadas), **[*W*]** o vetor de peso linear e as funções de base radiais na matriz **[*Φ*],** como em [2-19]:

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & .. & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & .. & \varphi_{2N} \\ .. & .. & .. & .. \\ \varphi_{N1} & \varphi_{N2} & .. & \varphi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ .. \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ .. \\ d_N \end{bmatrix} \qquad [2\text{-}19]$$

Explicitando [*w*], a equação [2-19] pode ser re-escrita de forma compacta como apresentado em [2-20], considerando que uma grande classe de funções de base radial φ

conduz a uma matriz **[Φ]** não-singular (pontos x_i distintos), portanto, que admite inversa.

$$[w] = [\Phi]^{-1}[d] \qquad [2\text{-}20]$$

Tendo-se o valor dos pesos *w*, tem-se a função *F(x)* que representa uma superfície contínua diferenciavel, que passa exatamente em cada um dos pontos.

As funções de base radial mais utilizadas, e que conduzem a matrizes **[Φ]** não singulares, estão apresentadas em [2-21], [2-22] e [2-23].

Multiquádricas: $\varphi(r) = \left(r^2 + c^2\right)^{1/2} \quad para \quad c > 0 \quad e \quad r \in R$ [2-21]

Multiquádricas inversas: $\varphi(r) = \dfrac{1}{\left(r^2 + c^2\right)^{1/2}} \quad para \quad c > 0 \quad e \quad r \in R$ [2-22]

Funções gaussianas: $\varphi(r) = \exp\left(-\dfrac{r^2}{2\sigma^2}\right) \quad para \quad \sigma > 0 \quad e \quad r \in R$ [2-23]

A obtenção de superfícies com base em um conjunto de pontos de dados, que podem ser esparsos, poderá constituir-se em um problema de reconstrução de hiper-superfícies mal-formulado, provenientes: 1) da não existência de sinal de saída distinta para cada entrada da rede; 2) da não existência de tanta informação na amostra de treinamento capaz de reconstruir o mapeamento de entrada-saída, e; 3)da existência de ruídos ou imprecisão dos dados de treinamento adicionando incertezas na reconstrução do mapeamento de entrada-saída(Haykin, 2001).

A teoria da regularização(Tikhonov, 1963 apud Haykin, 2001), apresenta um método para resolver problemas mal-formulados, denominado de *regularização*, que consiste basicamente em *estabilizar* a solução por meio de algum funcional, levando em consideração informações prévias de que os ajustes ocorrem de forma *suave*, ou seja, entradas similares correspondem a saídas similares.

A teoria da regularização introduz duas componentes na equação de regularização [2-24], quais sejam: O termo de erro padrão $E_s(F)$, que mede o erro(distância) padrão entre a resposta desejada e a resposta real, e o termo de regularização $E_r(F)$, que penaliza mapeamentos que não se ajustam(Bullinaria, 2003).

$$E(F) = E_s(F) + E_r(F) = \frac{1}{2}\sum_p \sum_k \left[y_k(x_p) - d_k^p\right]^2 + \frac{1}{2}\lambda \sum_k \int \left|P\, y_k(x)\right|^2 dx \qquad [2\text{-}24]$$

sendo λ o parâmetro de regularização que define a relativa importância do ajuste comparado com o erro, e P pode assumir várias formas, sendo a idéia geral mapear funções $y_k(x)$ que apresentam grande curvatura e tem portanto grandes valores de $\left|P\, y_k(x)\right|^2$ e contribuem com grande penalização no valor total da função erro.

A solução do problema de regularização consiste em encontrar uma função $F_\lambda(x)$ que minimiza o funcional de Tikhonov $E(F)$, definido pela equação [2-24].

A equação [2-25], após algumas transformações matemáticas, apresenta a solução do problema de regularização, onde w_i representa os coeficientes de expansão da rede de regularização e $G(x,x_i)$ é a função de Green centrada em x_i, que constitui uma base em um subespaço N-dimensional(Poggio e Girosi, 1990 apud Haykin, 2001).

$$F_\lambda(x) = \sum_{i=1}^{N} w_i G(x, x_i) \qquad [2\text{-}25]$$

Expandindo a equação [2-25] e desenvolvendo a função de Green como uma *função gaussiana multivariada*, teremos a forma apresentada em [2-26]

$$F_\lambda(x) = \frac{1}{\lambda}\sum_{i=1}^{N}\left[d_i - F(x_i)\right] \cdot \exp\left(-\frac{1}{2\sigma_i^2}\left\|x - x_i\right\|^2\right) \qquad [2\text{-}26]$$

A equação [2-25], permite implementar uma rede de regularização, como apresentada na Figura 2.12, que consiste de três camadas. A primeira corresponde aos nós de entrada, com dimensão igual ao número de variáveis independentes do problema, a

segunda é constituída de neurônios ocultos, com uma unidade para cada ponto dos dados, com unidades não-lineares ativadas com funções de Green, e a terceira camada com uma única unidade linear, totalmente conectada à camada oculta.

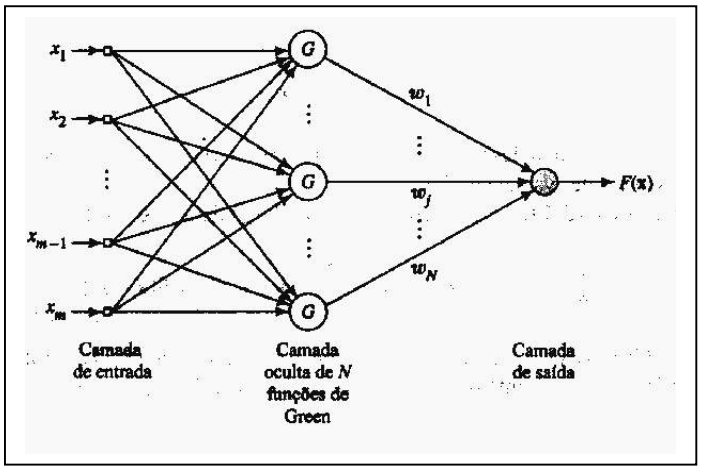


Figura 2.12 – Rede de regularização(Haykin, 2001).

As estratégias de aprendizagem estão relacionadas com o tipo de escolha que será efetuada para a determinação dos centros das funções de base radial da rede e dos parâmetros livres da rede. Os métodos apresentados estão formulados com base na teoria da interpolação.

1) Uma abordagem assume *centros fixos selecionados ao acaso,* escolhidos aleatoriamente, que para problemas em que os dados de entrada estão distribuídos de uma forma representativa, poderá representar uma boa escolha(Lowe, 1989 apud Haykin, 2001).

2) O método de *seleção auto-organizada de centros* determina os centros das funções de base radial a partir de um processo de agrupamento, que pode ser obtido com o algoritmo *k-means.*

O método adotado pelo algoritmo *k-means* pode conduzir resultados para mínimos locais, pois é dependente da escolha inicial dos centros das funções, o que poderá resultar em uma rede desnecessariamente grande. Esta limitação foi superada com o desenvolvimento de um algoritmo *k-means* aperfeiçoado(Chen, 1995; Chinrungrueng e Séquin, 1994 apud Haykin, 2001) que capacita o algoritmo a convergir para uma configuração ótima ou próxima da ótima , independente da localização inicial dos centros.

3) Na *seleção supervisionada de centros*, a rede assume sua forma mais generalizada em que os centros das funções de base radial e todos os outros parâmetros livres sofrem um processo de aprendizagem supervisionada.

Experimento de desempenho de redes, comparando perceptrons de múltiplas camadas e redes de função de base radial, concluíu(Wettschereck e Dietterich, 1992 apud Haykin, 2001):

> *-As redes RBF, com aprendizagem não-supervisionada das localizações dos centros, e aprendizagem supervisionada dos pesos da camada de saída, não generalizam tão bem como os perceptrons de múltiplas camadas treinados com o algoritmo de retropropagação.*
>
> *-As redes RBF, com aprendizagem supervisionada das localizações dos centros, bem como, dos pesos da camada de saída, são capazes de superar substancialmente o desempenho de generalização dos perceptrons de múltiplas camadas.*

-Mapas auto-organizáveis.

As arquiteturas de redes e o processamento de sinais utilizados para representar o sistema nervoso podem ser divididos em três categorias, cada uma baseada em diferente filosofia(Kohonen, 1990).

Redes de retropropagação*(Feedefoward networks)*(Rumelhart, Hinton e Williams, 1986b), transformam vetores de entrada em vetores de saída, sendo que os parâmetros internos de caminhamento dos sinais são, geralmente, ajustados através de supervisão externa.

Nas redes realimentadas*(feedback network)*(Hopfield, 1982 em Kohonen, 1990) a informação de entrada define o estado inicial da atividade do sistema de realimentação, e após o estado de transição o estado assintótico final é identificado como a saída do processo de computação.

Na terceira categoria, células vizinhas em uma rede neural competem em suas atividades através de interação mútua lateral e desenvolvem adaptações entre detectores específicos de diferentes padrões de sinal. Esta categoria é denominada auto-organizável*(self-organizing)*.

Baseada na forma de organização cerebral dos diferentes sinais sensoriais, foram desenvolvidos dois *modelos de mapeamento de características* diferentes. O primeiro modelo foi proposto sobre bases biológicas, para explicar o problema do mapeamento retinotópico da retina para o córtex visual(Willshaw e von der Malsburg, 1976 apud Haykin, 2001). O segundo é um modelo que se assemelha a uma rede neural artificial cujas células recebem sinais entrada de vários padrões ou classes, através de um processo de aprendizagem não-supervisionado(Kohonen, 1990). O modelo proposto pertence à categoria de algoritmos de codificação vetorial, produzindo um mapeamento topológico que localiza otimamente um número fixo de vetores em um espaço de entrada de dimensionalidade mais elevada, conforme mostrado na Figura 2.13.

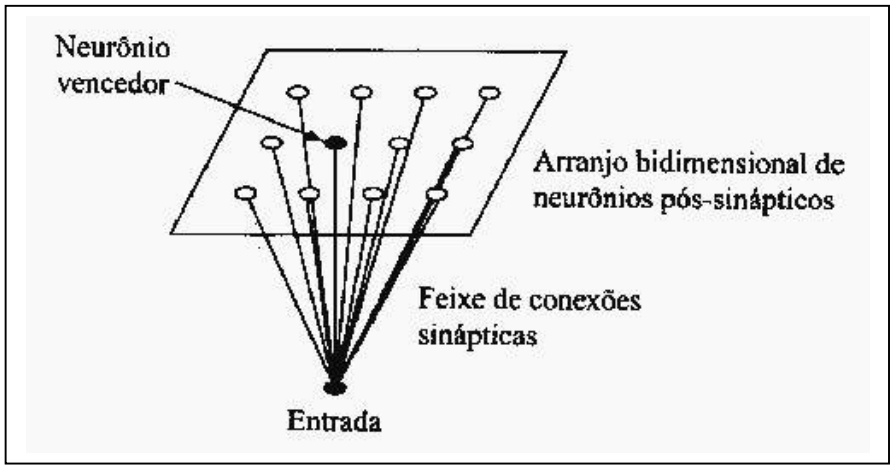


Figura 2.13 – Modelo de mapa auto-organizável de Kohonen(Haykin,2001)

O mapa auto-organizável(SOM-*self-organizing map*) tem por objetivo principal transformar um sinal de entrada de dimensão arbitrária em um mapa uni- ou bidimensional e realizar esta transformação de uma maneira topologicamente ordenada.

A formação do mapa auto-organizável de características é, determinado a partir de três processos essenciais(Haykin, 2001):

- O *processo competitivo* corresponde à identificação do neurônio vencedor que melhor se adapta com o vetor de entrada. Assim se tivermos um padrão de entrada dado por $x = [x_1, x_2, \ldots, x_m]'$ e o vetor de pesos sinápticos do neurônio *j* dado por $w_j = [w_{j1}, w_{j2}, \ldots, w_{jm}]'$, o neurônio vencedor da competição será identificado a partir da minimização da distância euclidiana, obtida a partir da equação [2-27], sendo ℓ o número de neurônios da grade.

$$i(x) = \arg \min_j \left\| x - w_j \right\| \quad j = 1,2,\ldots,\ell \qquad [2\text{-}27]$$

- O *processo cooperativo* que localiza o centro de uma vizinhança topológica de neurônios cooperativos, é determinado a partir de uma função unimodal $h_{j,i(x)}$, denominada de vizinhança topológica, da distância lateral $d_{j,i}$ entre o neurônio vencedor *i* e o neurônio ativado *j*.

A vizinhança topológica $h_{j,i(x)}$ deve satisfazer a condição de simetria em relação ao ponto máximo, definido como $d_{j,i}=0$, e deve decrescer monotonamente com $d_{j,i}\to\infty$, recaindo a escolha típica na função *gaussiano*. Para o caso de uma grade bidimensional a distância lateral $d_{j,i}$ é determinada pela distância entre os vetores, r_j que posiciona o neurônio excitado *j*, e r_i que posiciona o neurônio vencedor *i*, conforme mostrado na equação [2-28]. Uma característica do algoritmo é diminuir a vizinhança topológica com o tempo, o que é obtido pela introdução de uma função de decaimento exponencial em σ(n), onde *n* é o número de iterações e τ_1 é uma constante de tempo.

$$h_{j,i(x)}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)} \right) = \exp\left(-\frac{\left\| r_j - r_i \right\|^2}{2\left(\sigma_0 \exp\left(-\frac{n}{\tau_1} \right) \right)^2} \right) \qquad [2\text{-}28]$$

- O *processo adaptativo*, tem a finalidade de introduzir modificações no vetor de pesos sinápticos w_j do neurônio *j*, relativamente ao vetor de entrada *x*, para tornar a

grade auto-organizável. A equação [2-29] apresenta a forma de atualização do vetor de pesos sinápticos, onde $\eta(n)=\eta_0 \exp(-n/\tau_2)$, é o parâmetro da taxa de aprendizagem, que decresce monotonamente com o tempo, da mesma forma que a função σ(n).

$$w_j(n+1) = w_j(n) + \eta(n)\, h_{j,i(x)}(n)\left(x - w_j(n)\right) \qquad [2\text{-}29]$$

2.4.5 Aplicações da modelagem não-linear para prognóstico de demandas de potência.

Diversos métodos têm sido propostos(Hippert, Pedreira, Souza, 2001), para o prognóstico da demanda de potência elétrica, com a finalidade de reduzir o erro intrínsico à natureza do fenômeno.
Uma técnica utilizada é baseada na similaridade(Senjyu, Higa, Yue-Jin, Uezato, 1997 apud Senjyu, Takara, Uezato e Funabashi, 2002), isto é, utiliza dados de dias que apresentam as mesmas condições ambientais, do dia em que se deseja obter o prognóstico. Estes métodos têm a vantagem, de poder apresentar resultados satisfatórios, não somente das não linearidades da curva de demanda de potência, mas também de valores associados a feriados e fins de semana.

A maioria dos trabalhos apresenta resultados de prognóstico das próximas 24 horas ou a demanda de pico do próximo dia, utilizando prognóstico da temperatura como informação para a determinação do prognóstico de demanda de potência.

Tal metodologia poderá apresentar resultados, com elevada margem de erro, quando em presença de variações bruscas nas condições ambientais. No entanto, a utilização da temperatura do dia em que se deseja obter os dados de demanda de potência para a hora seguinte, conduz a resultados coerentes e satisfatórios.

Método apresentado(Senjyu, Takara, Uezato e Funabashi, 2002) utilizando correção de dados de dias similares para a determinação do prognóstico da hora seguinte, calcula a norma Euclidiana com fatores de peso para avaliar a similaridade entre o dia a ser prognosticado e os dias anteriores a serem pesquisados.

A pesquisa de valores similares cobre um período de 30 dias anteriores ao dia prognosticado, e, 60 dias no entorno do dia similar ocorrido nos dois anos anteriores. Esses valores, correspondentes aos dias similares, são utilizados em um algoritmo de retropropagação para efetuar a aprendizagem de uma rede neural, tendo como variáveis de entrada os desvios da temperatura e da demanda de potência entre os valores atuais e os valores correspondentes aos dias similares e um fator de correção de prognóstico.

A variável de saída da rede é o fator de correção de prognóstico, o qual permite obter o prognóstico de demanda de potência e o desvio do valor ocorrido, cujos valores são utilizados para alimentar *on-line* o processo de aprendizagem. A correção *on-line* permite obter resultados de mudanças bruscas de temperatura, de forma a efetuar as mudanças horárias que venham a ocorrer nas demandas de potência, correspondentes.
O método aplicado a dados da Okinawa Electric Power Company resultou em um erro percentual médio absoluto (MAE) de 1,18% para um período de um ano e de 0,9% quando aplicado para prognosticar demanda de potência em uma semana com poucas variações da temperatura, e, 1,23% e 1,11% em situações de variações bruscas na temperatura.
Vários trabalhos apresentados utilizam estruturas tipo (MLP-*multilayer perceptron)* perceptron de múltiplas camadas com variantes para atender condições específicas do problema.
Pesquisa levada a efeito(Khotanzad, Afkhami-Rohani, Lu, Abaye, Davis e Maratukulam, 1997) para determinação do prognóstico de carga, utiliza múltiplas redes neurais, com módulos semanal, diário e horário compostos de valores de entrada de demandas de potência, temperaturas e umidade do ar, e saídas com prognósticos de demandas de potência para as próximas 24 h.
A arquitetura de primeira geração(Khotanzad, Hwang, Abaye, Maratukulam,1995) é constituída por 38 redes perceptrons de múltiplas camadas, agrupadas em três módulos. Esses três módulos foram modelados para atender variações semanais, diárias e horárias da demanda de potência. Cada módulo gera o prognóstico de carga independente dos outros dois módulos. Esses três prognósticos são combinados adaptativamente, de modo a obter o prognóstico final da demanda de potência.

Essa arquitetura foi implementada em mais de 20 empresas de energia elétrica e a sua performance foi satisfatória e superior a outros algoritmos existentes na época (1992-1995), apesar de ter apresentado redução na precisão devido a condições climáticas extremas.

Esta imprecisão levou ao desenvolvimento da arquitetura de segunda geração(Khotanzad, Afkhami-Rohani, Lu, Abaye, Davis e Maratukulam, 1997), constituída de 24 redes menores tipo perceptron de múltiplas camadas, sendo cada uma para cada hora do dia. Nesta arquitetura foram considerados efeitos observados no comportamento da demanda, tais como, divisão da demanda de potência diária em categorias, segundo o período horário de sua ocorrência. Assim foi efetuada a divisão horária em: Madrugada e cedo pela manhã(1h às 9h); Manhã até início da tarde e noite(10h às 14h e 19h às 22h); Pico da tarde(15h às 18h); Tarde da noite(23h e 24h), sendo que cada categoria é afetada pelas mesmas variáveis, apesar de existir diferenças entre os fatores que influenciam os períodos horários das diferentes categorias.

As duas gerações de arquiteturas, quando aplicadas para prognosticar a demanda futura de potência, de 10 empresas de energia elétrica e seis meses de dados, apresentaram valores do erro absoluto médio(MAE), conforme tabela da Fig.2.14 , onde pode se observar que o valor do erro cresce a medida que o dia prognosticado se distancia do atual, e a arquitetura de segunda geração suplanta a de primeira geração.

		Dias futuros						
Prognóstico	Arquitetura	1	2	3	4	5	6	7
Todas as horas	Geração II	2,19	2,53	2,98	3,13	3,38	3,59	3,67
%	Geração I	2,52	2,87	3,21	3,53	3,73	3,88	4,00
Carga de pico	Geração II	2,48	2,87	3,03	3,23	3,39	3,52	3,66
%	Geração I	2,98	3,25	3,58	3,89	3,91	4,02	4,12
Diária total	Geração II	1,70	2,39	2,47	2,64	2,68	2,72	2,82
%	Geração I	2,26	2,88	3,02	3,05	3,10	3,16	3,19

Figura 2.14 – Erro de prognóstico utilizando algoritmo de Geração I e II.(Khotanzad, Afkhami-Rohani, Lu, Abaye, Davis e Maratukulam, 1997).

Um dos problemas observados(Khotanzad, Afkhami-Rohani, Maratukulam, 1998), na primeira e segunda gerações de arquiteturas, é a resposta muito lenta a variações rápidas

na carga provocadas por frentes de temperaturas quentes ou frias. Apesar dos mecanismos adaptativos associados às arquiteturas serem capazes de promover alterações na carga, é possível que estas alterações ocorram em um período longo, antes que esses modelos iniciem as alterações nos prognósticos de demanda de potência. Outra questão diz respeito ao perfil da demanda em feriados e dias especiais, que nas duas arquiteturas apresentam problemas, pois foram tratados como finais de semana, não apresentando erros inferiores àqueles encontrados para os dias de semana normais. Uma terceira questão diz respeito à necessidade de efetuar re-treinamento das redes após alguns anos, para melhorar a qualidade dos dados de entrada, mas o número elevado de redes a serem treinadas para as arquiteturas de primeira e segunda, gerações, dificulta sobremaneira, esta tarefa.A terceira geração de arquitetura consiste de três módulos, duas redes neurais e um adaptador para combinar as saídas, sendo que cada um deles recebe o mesmo conjunto de entrada e produz prognóstico de carga para o mesmo dia, mas utiliza diferentes estratégias para obtê-las.

A entrada em cada rede é composta de 24 valores de carga horária do dia anterior, 24 valores de temperatura horária do dia anterior, 24 valores de prognóstico de temperatura do dia seguinte e 7 valores correspondendo à indicação do tipo de dia da semana.

A primeira rede é treinada para prognosticar a carga regular(base) do dia seguinte(24 valores horários), denominada BLF - *Based Load Forecaster*. A segunda rede de prognóstico, avalia a variação nas demandas horárias, entre o dia anterior e o atual, denominada CLF – *Change Load Forecaster*.

Os valores BLF tendem, a responder mais lentamente às variações rápidas da carga, ao passo que, os valores CLF respondem mais rapidamente às variações bruscas da carga pelo fato de utilizar valores de demanda do dia anterior e prognosticar as variações experimentadas por esta demanda, para o dia seguinte. O prognóstico horário final é obtido pela combinação linear dos valores horários BLF e CLF, com os coeficientes da combinação linear obtidos através de um algoritmo recursivo de mínimos quadrados.

A comparação entre os valores MAE de demandas de 10 empresas de energia elétrica está apresentada na tabela de Fig. 2.15, onde se observa uma melhora sensível nos valores obtidos na terceira arquitetura relativamente à segunda.

		Dias futuros						
Prognóstico	Arquitetura	1	2	3	4	5	6	7
Todas as horas %	Geração III	2,05	2,53	2,72	2,82	2,89	2,94	2,99
	Geração II	2,26	2,83	3,08	3,26	3,41	3,55	3,92
Carga de pico %	Geração III	2,12	2,57	2,71	2,80	2,89	2,97	3,03
	Geração II	2,34	2,82	3,06	3,28	3,45	3,60	3,75

Figura 2.15 – Erro de prognóstico utilizando algoritmo de Geração II e III.(Khotanzad, Afkhami-Rohani, Maratukulam, 1998).

Para a determinação do prognóstico de curtíssimo prazo(Charytoniuk, Chen, 2000), desde um minuto até alguns minutos no futuro, a metodologia a ser utilizada requer outros requisitos, diferentes daqueles utilizados para a determinação de prognósticos de curto prazo.

Ao invés de modelar a relação entre a carga, horário, condições ambientais e outros fatores que afetam a demanda de potência, neste caso, é utilizada a extrapolação para o futuro próximo, de padrões de valores recentemente observados.

O valor a ser prognosticado no futuro de curtíssimo prazo, é constituído pelo valor horário imediatamente anterior adicionado de uma parcela, obtida a partir do prognóstico de um coeficiente de incremento relativo, determinado com base nos *n* valores incrementais relativos, anteriores, da carga, utilizando uma rede neural.

Na utilização de valores incrementais, da ordem de 20 a 60 minutos, aplicada para prognosticar uma demanda de potência de 24h, resultaram valores MAE entre 0,4% e 1,1%.

Um valor importante a ser determinado na série representativa da demanda de potência é o valor de pico diário, com a finalidade de prover o planejamento operacional, do sistema, em estudo.

A comparação entre métodos de aprendizagem permite conhecer o desempenho relativo com a finalidade de obter aquele que apresenta o melhor resultado. Trabalho conduzido neste sentido(Saini, Soni, 2002), apresenta resultados(MAE) para condições ambientais de inverno, verão, tempo chuvoso e tempo seco para quatro tipos de algoritmos de aprendizagem, conforme apresentado na tabela da Fig.2.16, onde observa-se que o terceiro e quarto algoritmos apresentam os melhores resultados.

Algoritmo	Dias no futuro								
	Dia atual	1	2	3	4	5	6	7	Média
Steepest descent	3,08	2,17	2,64	2,74	2,98	2,41	2,89	3,29	2,78
Levenberg-Marquardt	2,69	2,50	2,96	2,84	2,89	3,04	3,47	2,54	2,87
Broyden-Fletcher -Goldfarb-Shanno	2,28	2,47	2,08	2,21	2,50	2,32	2,40	2,77	2,38
One-step secant	2,08	2,39	2,02	2,23	2,30	2,58	2,78	2,92	2,41

Figura 2.16 – Erro de prognóstico, utilizando quatro tipos de algoritmos de aprendizagem com método BP(*back-propagation)*(Saini, Soni, 2002).

Com a finalidade de observar o comportamento interno de uma rede neural, normalmente denominada de "caixa preta", devido à dificuldade de observação, foi desenvolvida uma estrutura neural analizável(Iizaka, Matsui, Fukuyama, 2002), que permite através do método de treinamento extrair conhecimentos próprios e expor as razões dos resultados de prognóstico, com correlação independente entre variáveis de entrada e demanda de pico.

Conforme mostrado na Fig. 2.17, a estrutura da rede neural é composta por dois tipos de neurônios da camada oculta, um módulo denominado de *conexões esparsas,* que apresenta correlação independente entre um conjunto de variáveis de entrada e de saída, e um módulo denominado de *totalmente conectado,* que apresenta interações entre as unidades de entrada com a finalidade de extrair correlações dos dados de treinamento e expor as razões do prognóstico.

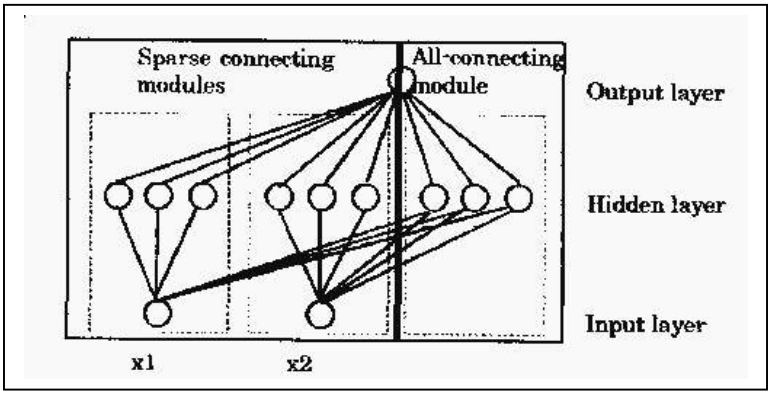


Figura 2.17 – Estrutura de rede neural analizavel (Iizaka, Matsui, Fukuyama, 2002).

No método proposto os agrupamentos de variáveis utilizados para constituir os módulos esparsos foram: os valores de pico de demanda dos dias anteriores, as temperaturas máximas e mínimas ocorridas e o dia de semana, sendo que os mesmos dados são utilizados para os neurônios totalmente conectados.

O treinamento é efetuado para cada variável separadamente possibilitando a análise da influência daquela variável no valor de saída.

Resultados, para a determinação de valor de pico do dia seguinte, utilizando redes neurais e treinamento convencional com o algoritmo de retropropagação, rede analizavel com treinamento simultâneo de todos os conjuntos de variáveis e rede analizavel com treinamento independente de cada conjunto de variável, estão apresentados na tabela da Fig. 2.18.

Método	MAE %	
	Primavera	Verão
Convencional retropropagação	2,66	1,64
Analizável convencional	2,39	1,60
Analizável independente	2,53	1,57

Figura 2.18 – Erro de prognóstico de pico de demanda, utilizando três métodos diferentes para a estrutura neural analizavel(Iizaka, Matsui, Fukuyama, 2002).

Para responder a questões sobre a dependência da modelagem com redes neurais com determinados dados específicos de demanda de potência, ou mesmo com respeito à robustez do modelo, para responder a todas as variações experimentadas pelos dados disponíveis e também quanto a efetividade do prognóstico de demanda para as próximas 24h, pesquisa efetuada(Lu, Wu, Vemuri, 1993) considerou dados de duas empresas de energia elétrica e efetuou estudos comparativos para responder às questões levantadas.

As redes neurais das duas empresas utilizavam como dados de entrada valores passados de temperaturas, correlação entre hora e demanda de potência, dia da semana e horário,

e como saída única o prognóstico de demanda de potência para a próxima hora. Esse valor da primeira hora era utilizado para obter valor de prognóstico para a segunda hora e assim sucessivamente.

Uma terceira rede foi proposta com entradas de temperatura média do dia anterior, dias da semana e demandas horárias(24h) ocorridas nos dias anteriores, tendo como saída o prognóstico das demandas horárias do dia seguinte(24 valores). O resultado obtido na aplicação das duas redes específicas, aos dados de cada empresa, respectivamente, e, a aplicação dos dados de ambas na terceira rede, está resumida na tabela de Fig.2.19, onde se pode observar que a modelagem específica para cada empresa conduz a resultados com menor erro.

Período de prognóstico	MAE %			
	Redes específicas		Rede geral	
	Empresa A	Empresa B	Empresa A	Empresa B
24h	1,97	2,77	2,40	6,11
48h	1,90	2,65	2,31	4,09
72h	2,03	2,71	2,20	4,75
96h	1,96	2,81	2,17	3,18
120h	1,98	2,90	2,04	4,57
144h	1,93	3,00	2,33	3,77
168h	1,96	3,03	2,87	3.56
Valor médio	1,96	2,84	2,34	4,30

Figura 2.19 – Erro de prognóstico, utilizando redes específicas e rede geral para dados de duas empresas(Lu, Wu, Vemuri, 1993).

Baseado nos resultados apresentados nos testes de prognóstico, a pesquisa concluiu que: *Não há critério firme para selecionar estrutura de rede capaz de processar um conjunto de demandas horárias e dados de temperatura. Os modelos não são únicos e sistemas com diferentes características de carga requerem diferentes estruturas. Redes neurais são sensíveis a valores errados, conforme demonstrado em testes efetuados com a introdução de valores errôneos no conjunto de dados.*

Redes neurais convencionais quando utilizadas, para modelar prognósticos, de demandas de potência, não podem oferecer bons resultados, uma vez que devem considerar o conjunto de dados de todos os dias do ano, em presença de múltiplos perfis devido a fatores sazonais, econômicos e culturais.

Valores do erro percentual médio absoluto(MAE), relativamente baixos, quando comparados com a maioria dos casos, foram encontrados em pesquisa(Marin, Garcia-Lagos, Joya, Sandoval, 2002) desenvolvida com a finalidade de obter um modelo global para prognóstico de demandas de potência .
O modelo, conforme mostrado na Fig. 2.20, é composto de uma fase inicial em que os dados históricos são classificados com o algoritmo de mapas organizáveis de Kohonen (SOM), condicionada a fatores metereológicos, sociais e econômicos específicos de cada região de concessão. A segunda fase é composta de redes para cada classe, treinadas com os dados obtidos na primeira fase. A terceira fase determina o tipo de estudo a ser executado, a curva de demanda horária do dia seguinte ou a demanda da hora seguinte, cada um necessitando de dados diferentes.

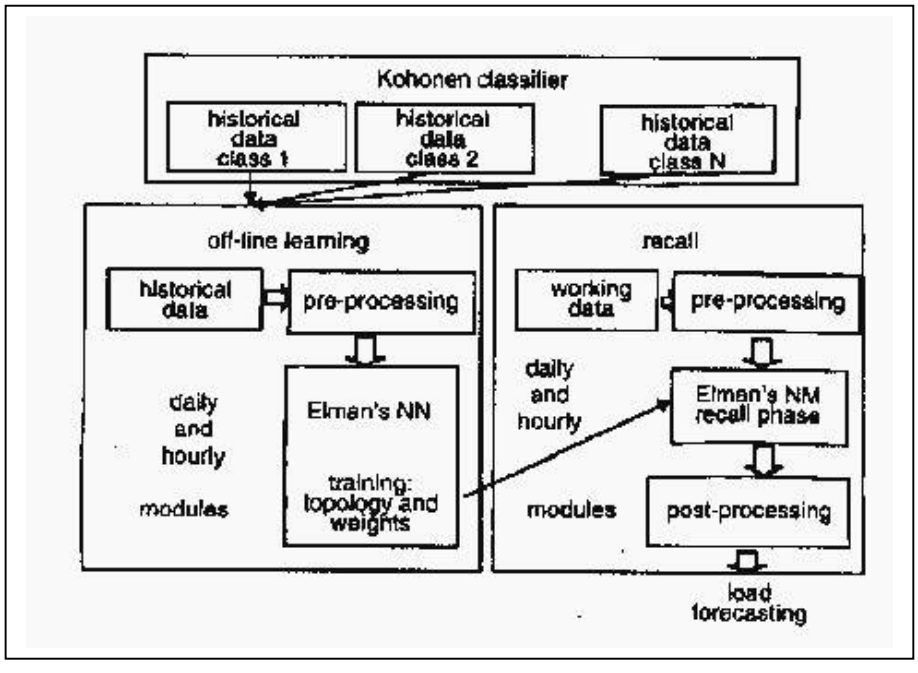


Figura 2.20 –Modelo global para prognóstico de demanda da potência. (Marin, Garcia-Lagos, Joya, Sandoval, 2002)

Testes efetuados com dados de demandas de potência horária cobrindo os anos de 1989 a 1999 resultaram, nos valores apresentados na tabela da Fig. 2.21, de acordo com a categoria classificada pelo algoritmo de mapas organizáveis.

Classificação	MAPE %
Dias de semana	1,45
Sábados	1,46
Domingos	1,62
Segunda-feira	1,65
Festas religiosas	1,87

Figura 2.21 – Erro de prognóstico utilizando o modelo global(Marin, Garcia-Lagos, Joya, Sandoval, 2002).

Procedimento conduzido na mesma linha(Lamedica, Prudenzi, Sforna, Caciotta, Cencelli, 1996), com a finalidade de obter procedimentos para dias com demandas anômalas, apresenta uma metodologia dividida em três estágios. O primeiro estágio providencia um critério de identificação das características dos dias, no que diz respeito à demanda de potência, estabelecendo *clusters* com perfis semelhantes de carga. O segundo estágio consiste na atualização do processo de informação obtido no primeiro estágio. Esse estágio é efetuado pelos operadores do sistema. O terceiro estágio efetua o prognóstico utilizando redes neurais tipo perceptron de múltiplas camadas.

A modelagem adotada, não utiliza variáveis ambientais nos dados de entrada, considerando que, em condições de períodos anômalos, predominam as influências dos aspectos sociais no perfil da carga.

Testes efetuados para períodos anômalos apresentaram, na época, erros percentuais médios absolutos da ordem de 5% a 20%, que para as necessidades atuais são insatisfatórios, no entanto, já haviam detectado a necessidade de estabelecer procedimentos específicos para os períodos em que a demanda de potência se apresenta de forma não padronizada.

As pesquisas têm sido desenvolvidas, com o objetivo de melhorar os resultados de prognóstico, bem como, obter métodos que possam melhor responder ás variações de carga, experimentadas pelos sistemas.

Pesquisa atual(Ling, Leung, Lam, Lee, Tam, 2003) introduz a conceituação de algoritmo genético nas redes neurais para a determinação de prognóstico de séries temporais.

Conforme Fig. 2.22, os neurônios ocultos(camada central) apresentam duas funções de ativação: A função de ativação estática(SAF) e a função de ativação dinâmica(DAF) que governa a relação entrada – saída do neurônio. Para a função SAF, os parâmetros são fixos e sua saída depende dos dados de entrada do neurônio. Para a função DAF, os parâmetros dependem da saída de outros neurônios e da saída da sua função SAF.

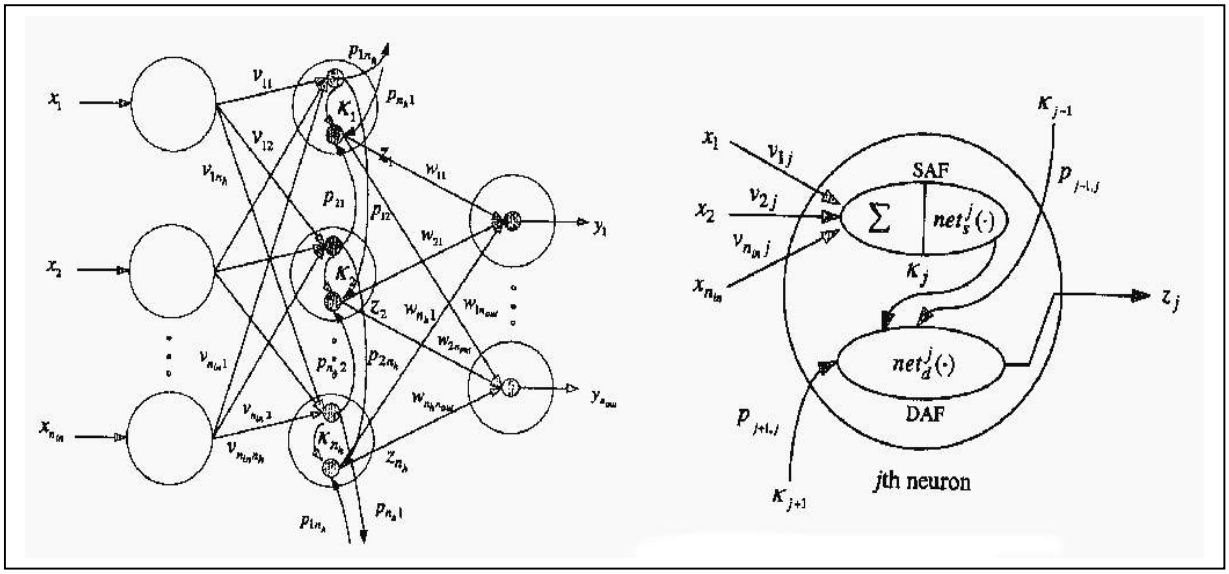


Figura 2.22 – Rede neural e modelo de neurônio com duas funções de ativação(Ling, Leung, Lam, Lee, Tam, 2003).

Os testes realizados para a determinação do prognóstico de carga futura apresentaram valores de erro percentual médio absoluto 1,91% e 1,93%, para quarta-feira e domingo, respectivamente, correspondendo a uma melhora na performance da ordem de 30% com relação aos resultados da aplicação dos mesmos dados a uma rede convencional.

Pesquisa comparativa entre métodos que utilizam algoritmo de retropropagação baseada em métodos do gradiente conjugado(Saini, Soni, 2002), apresentaram os resultados da tabela da Fig. 2.23, onde o valor MAE representa a média dos erros médios absolutos, obtidos a partir dos valores dos erros computados para cad dia de prognóstico futuro. Os algoritmos de gradiente conjugado, utilizados foram: Steepest Descent Back-Propagation(SDBP), Fletcher-Reeves(FR), Polak-Ribiere(PR), Powell-Beale(PB), Scaled Conjugate Gradient(SCG).

Algoritmo	MAE %
SDBP	2,78
FR	2,43
PR	2,32
PB	2,31
SCG	2,40

Figura 2.23 – Erro de prognóstico de vários algoritmos de gradiente conjugado (Saini, Soni, 2002).

Com a finalidade de efetuar uma separação dos dados de acordo com a similaridade, muitos métodos utilizam o agrupamento *(clusters)* de dados, aplicando algoritmos específicos, tais como mapas auto-organizáveis, K-means e outros.

Pesquisa efetuada utilizando um algoritmo denominado *Deterministic Annealing Clustering(DA)*(Mori, Yuihara, 2001)*, p*rocura uma solução ótima para cada "temperatura" da energia livre do sistema em estudo, através de método de pesquisa de proximidade dos vizinhos, de uma forma determinística, que independe das condições iniciais e apresenta bons resultados em termos de pesquisa global.

A tabela da Fig. 2.24 apresenta os resultados comparativos entre os métodos Multi Layer Perceptron(MLP), MLP com agrupamento K-means(MLPK) e MLP com agrupamento *Deterministic Annealing(MLPDA)*.

Método	MAE - %			
	Aprendizagem		Prognóstico	
	Médio	Máximo	Médio	Máximo
MLP	1,357	6,729	1,648	4,909
MLPK	1,292	6,708	1,411	4,112
MLPDA	1,297	7,253	1,394	3,660

Figura 2.24 – Erro de prognóstico de vários métodos(Mori, Yuihara, 2001).

O desenvolvimento de variantes na determinação do prognóstico de cargas tem dominado a pesquisa, com a introdução de novas formas de obtenção, através do uso de redes neurais, dos valores futuros das demandas de potência de uma hora, um dia e uma semana.

A utilização da teoria de análise de microondas tem recebido, grande atenção, na aplicação de análise de distúrbios transitórios e sinais, existentes nos sistemas de potência.

A transformação em microondas decompõe, o sinal de entrada em escalas de sinais com diferentes resoluções, sendo que cada escala indica as diferentes freqüências contidas no sinal original. Pela observação do escalamento na transformação dos sinais, a ocorrência e classificação do tipo, de distúrbio, podem ser identificadas.

Uma modelagem utilizando a teoria de microondas foi desenvolvida(Huang, Yang, 2001) para modelar a alta não linearidade e comportamento dinâmico do sistema de cargas para melhorar a performance das redes neurais tradicionais.

A rede neural está construída com três camadas, sendo que a primeira camada decompõe o sinal de entrada em diversas escalas de sinal, cuja função de ativação é composta por uma condição de admissibilidade obtida através da série de Fourier, a segunda camada é constituída por valores W, obtidos por um processo de competição entre dois conjuntos de indivíduos, o conjunto representativo, de "parentes", da população envolvida no processo, sendo que a cada indivíduo está agregado um valor de desqualificação e, o conjunto de mesmo tamanho formado por "filhos" de cada indivíduo pertencente ao primeiro conjunto, e a terceira camada, correspondente à saída da rede, é composta pela soma dos produtos entre os valores do caminhamento entrada-

saída, da função de Fourier obtida na primeira camada e os fatores de peso que interligam a segunda camada e a de saída.

Na aplicação de dados para quatro tipos de carga(verão, inverno, transformador e alimentador) e quatro tipos de dias, da empresa Taipower, para prognóstico horário, diário e semanal foram obtidos os valores apresentados na Fig.2.25, onde se observa que os resultados de erro MAE máximo e médio obtidos utilizando a teoria de microondas são sensivelmente melhores que aqueles obtidos com a aplicação de algoritmos de redes neurais tradicionais.

	Valor máximo MAE%			Valor médio MAE %		
	Horário	Diário	Semanal	Horário	Diário	Semanal
Redes neurais tradicionais	6,25	6,96	11,37	1,43	2,32	1,97
Teoria de microondas	4,26	5,90	6,64	1,13	1,72	1,54

Figura 2.25 – Erro de prognóstico, utilizando redes tradicionais e teoria de microondas(Huang, Yang, 2001)

Na literatura existem vários métodos de obtenção do prognóstico de séries temporais que podem ser basicamente divididas em duas categorias principais: Métodos paramétricos e métodos baseados na inteligência artificial.

Um método alternativo pesquisado é a aplicação de método de regressão não-paramétrico que utiliza, para prognóstico, dados obtidos diretamente da série histórica.

A modelagem de demanda de potência deve refletir sua conformação geral, tal como, a existência de padrões diários, a dependência com a temperatura e outras variáveis relevantes e sua natureza randômica. Assim, a propriedade da carga pode ser descrita em termos de uma função de densidade de probabilidade multivariada (PDF) da carga, tempo, temperatura e outros possíveis fatores(Charytoniuk, Chen, Van Olinda, 1998).

A função PDF=f(P,$\boldsymbol{x}$), sendo P um valor médio da carga horária e $\boldsymbol{x}=(x_{1,\ldots,} x_r)^t$ é o vetor de variáveis que afetam o valor da demanda de potência, tais como, hora do dia, e fatores ambientais, pode ser obtida dos dados históricos e outras variáveis que afetam o problema, diretamente, pela estimação da densidade não-paramétrica dos exemplos.

Os erros resultantes da aplicação em dados de três semanas no verão, utilizando regressão não paramétrica com variação do modelo de carga com a temperatura, regressão não paramétrica com carga estática e redes neurais tradicionais, estão apresentados na tabela da Fig. 2.26.

Tipo de prognóstico	MAE %
Não paramétrico dinâmico	2,78
Não paramétrico estático	3,01
Redes neurais	2,64

Figura 2.26 – Erro de prognóstico utilizando métodos não paramétricos e redes neurais(Charytoniuk, Chen, Van Olinda, 1998).

Uma questão importante e de grande relevância na determinação de prognóstico de demanda de potência, diz respeito à seleção de variáveis que devem participar como valores de entrada de modo a obter resultados satisfatórios na saída da rede.

Pesquisa desenvolvida nesse sentido(Drezga, Rahman, 1998), ressalta a importância da escolha correta de variáveis, pois considera que as redes neurais apreendem relações entre variáveis de entrada e saída desde que os valores de entrada-saída formem pares. Se as variáveis de entrada não apresentam relevância no processo, não é possível esperar que a rede neural possa obter uma representação correta da dependência entre as variáveis. Assim, se forem utilizadas variáveis que apresentam baixa correlação ou até mesmo insignificante, para valores de entrada, a pesquisa computacional, na fase de treinamento da rede, poderá resultar em acréscimo no tempo de processamento, muitas vezes com resultados inferiores, pois, a busca se faz preferencialmente de características mutuamente independentes.

Pesquisa efetuada(Piras, Germond, Buchenel, Imhof e Jaccard, 1996)identificaram 32 tipos diferentes de variáveis de entrada na determinação do prognóstico de demanda de potência, conseqüentemente muitas combinações podem ser construídas.

A técnica utilizada na pesquisa de variáveis que apresentem correlação elevada para a determinação de prognóstico de demandas de potência, utiliza as observações da série temporal em estudo para reconstruir a dinâmica de complexidade do sistema(*phase-*

space embedding). A técnica transforma estados desconhecidos do sistema, dentro de um sistema derivado das medidas observadas.

A reconstrução (*phase-space*) é efetuada através de método que utiliza valores em atraso da série temporal. O método não se apresenta tão sensitivo a ruídos, e, portanto, mais adequado para dados obtidos de medições.

Aplicação do método, aos dados de duas empresas de energia resultaram, na determinação de prognóstico de demanda de potência para as próximas 24h, em um valor MAE em torno de 2% e um valor máximo do erro entre 8% e 10% para todos os meses considerados no estudo.

Aplicação do método(Drezga, Rahman, 1999) foi efetuado em dados de duas empresas determinando valores de prognóstico horários e de pico de demanda.

A identificação de variáveis foi efetuada utilizando a técnica *phase space embedding*(Drezga, Rahman, 1998), para identificar um conjunto compacto de variáveis locais.

O conjunto de treinamento para cada horizonte a ser prognosticado (dias de semana com cinco dias consecutivos e fins de semana com sábado e domingo), foi determinado utilizanda a técnica kNN – *k nearest neighbors.*

Atenção especial foi tomada na determinação do número de neurônios da camada oculta, que foi obtido utilizando-se uma rede auxiliar em que o valor alvo a ser atingido na fase de treinamento da rede, é substituído pelo seu vizinho próximo, repetindo-se o treinamento para várias configurações da camada oculta, obtendo-se a sua arquitetura pelo menor resultado no erro de treinamento.

Com a finalidade de aumentar a capacidade de generalização da rede neural, e considerando que estudos teóricos e experimentos práticos confirmam que, a combinação de valores estimados podem aumentar significantemente a precisão do prognóstico(Perrone, 1994 em Drezga, Rahman, 1999), foram utilizadas duas redes neurais disjuntas, com entradas randômicas, treinamento com parada mais cedo evitando o excesso de ajuste e obtenção do valor final a partir da média dos valores individuais de saída de cada rede.

Alguns resultados obtidos na aplicação aos dados das empresas estão apresentados na tabela da Fig. 2.27 onde o erro do prognóstico horário está compatível com outros

métodos utilizados e o erro do prognóstico de pico de carga apresenta resultados mais baixos que outros métodos.

Empresa	Prognóstico horário de carga			Prognóstico de pico de carga	
	24h	120/48h	1h	24h	120/48h
	Dias de semana				
A	2,05	2,44	0,98	2,43	1,88
B	2,04	2,15	1,13	2,12	2,67
	Fim de semana				
A	2,47	2,59	1,05	2,25	2,11
B	2,25	2,43	1,28	2,11	1,56

Figura 2.27 – Erro de prognóstico utilizando método de identificação de variáveis de entrada (Drezga, Rahman, 1999).

2.5 Síntese.

Esse capítulo apresentou os modelos de análise de séries temporais, do ponto de vista da modelagem linear e não linear, apresentando pesquisas efetuadas, com a utilização das mais variadas modelagens, bem como a utilização de soluções híbridas com mais de um modelo, na procura de melhores resultados.

A importância de se efetuar a análise dos modelos para a solução de redes temporais, reside no fato de que existe, uma quantidade muito grande de fenômenos que podem ser representados por uma sucessão de valores no tempo, e que de alguma forma foram responsáveis pelo desenvolvimento dos métodos existentes, uma vez que a curiosidade sempre conduziu o homem a tentar prognosticar valores futuros baseados em acontecimentos ocorridos em um passado recente.

A pesquisa a ser desenvolvida nos próximos capítulos diz respeito à analise quantitativa, com a previsão de determinação de valores futuros de uma série temporal, motivando a apresentação de modelos de análise com séries temporais discretas.

Na modelagem com funções lineares, são apresentados os métodos ARMA e suas variantes, o ajuste exponencial por regressão, que possivelmente foi um das primeiras tentativas de ajuste de valores discretos e o modelo de espaço de estado onde é apresentado o filtro de Kalman.

Os modelos classificados como não lineares apresentam a modelagem ARMA com a introdução da variável tempo nas variáveis do modelo perseguindo a não linearidade do problema, a modelagem com volatilidade variável que utiliza a conceituação de variância variável, os modelos denominados gerais onde é apresentada a aprendizagem por reforço, o modelo k-means que efetua o agrupamento de valores similares em torno de um centro, o modelo kNN-Nearest Neighbor que classifica os grupos segundo a proximidade da vizinhança e os modelos de árvore de decisão, que procura a direção daquela decisão que melhor responder à questão colocada.

Pesquisas recentes efetuadas na determinação de prognóstico de séries temporais de demanda de potência, com a apresentação de trabalhos onde cada fase da procura do valor futuro é efetuada com um determinado método específico, configurando o que se denomina de sistemas híbridos, incluindo em muitos casos algoritmos genéticos, lógica Fuzzy e outros métodos, mostram a atualidade do tema.

Dentre os modelos apresentados na Fig. 2.1, não foi apresentada a modelagem com máquina de vetor de suporte, que pode ser classificada como um modelo supervisionado com aprendizagem estatística que aprende relações não lineares com uma máquina linear, a partir do mapeamento do espaço de entrada em um espaço com características de alta dimensionalidade.

Esse modelo será objeto de estudo dos próximos capítulos, uma vez que o trabalho a ser apresentado utilizará como base para o desenvolvimento a máquina de vetor de suporte.

CAPÍTULO III – VETORES SUPORTE E FUNÇÕES NÚCLEO.

Assim como o perceptron de múltiplas camadas, treinado com o algoritmo de retropropagação, a máquina de vetor de suporte pode ser utilizada para classificação de padrões e regressão linear, apresentando algumas características vantajosas em relação às redes neurais. Neste capítulo apresentar-se-á a teoria da máquina de vetor de suporte e a construção de algoritmos de aprendizagem com utilização de funções núcleo.

A máquina de vetor de suporte é um sistema de aprendizagem treinado com um algoritmo baseado na teoria estatística de aprendizagem, que conceitualmente implementa a seguinte idéia: vetores do espaço de entrada são mapeados não-linearmente em um espaço com características de alta dimensionalidade, através de um mapeamento escolhido *a priori*. Neste espaço de características, é construída uma superfície de decisão linear, que se constitui de um hiperplano de separação ótima e que apresenta propriedades especiais que garantem alta habilidade de generalização da máquina de aprendizagem(Cristianini, Shawe-Taylor, 2002),(Cortes, Vapnik, 1995),(Vapnik, 2000).

A Fig. 3.1 mostra um diagrama esquemático do processo de construção do espaço de alta dimensionalidade e do hiperplano ótimo que efetua a separação entre exemplos positivos e negativos de forma que a margem de separação seja máxima. A máquina de vetor de suporte mapeia o espaço de entrada num espaço com características de alta dimensionalide e contrói neste, um hiperplano ótimo de separação de exemplos.

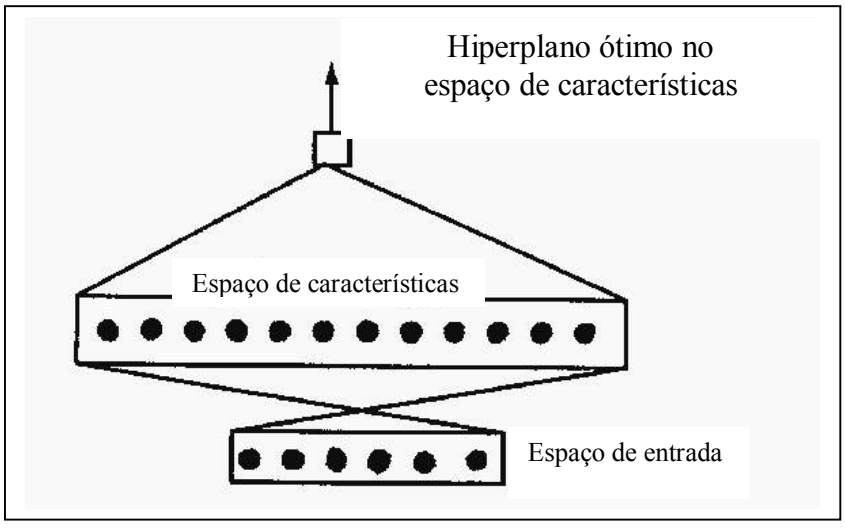


Figura 3.1 – Diagrama esquemático da máquina de vetor de suporte(Vapnik, 2000).

Os processos de aprendizagem, dentre os quais situa-se a máquina de vetor de suporte, são de natureza estatística, considerando que na determinação de valores obtidos pelo algoritmo de aprendizagem, os desvios em relação aos valores alvos são expressos em termos estatísticos(Haykin,2001).

3.1 Teoria estatística da aprendizagem.

O modelo geral de aprendizagem a partir de exemplos é composto de três componentes, como mostrado na Fig 3.2:

- Um gerador (**G**) de vetores randômicos $\boldsymbol{x} \in R^n$, processado independentemente com valor fixo e desconhecido da função de distribuição probabilística ***F(x)***.
- Um supervisor (**S**) que retorna um valor de saída **y** para cada vetor de entrada **x**, de acordo com uma função de distribuição condicional ***F(y /x)***, também fixa, mas desconhecida.
- Uma máquina de aprendizagem (**LM**) capaz de implementar um conjunto de funções $\boldsymbol{f(x,\alpha), \alpha \in \Lambda}$, onde $\boldsymbol{\Lambda}$ é um conjunto de parâmetros.

Durante o processo de aprendizagem, a máquina observa o par *(x,y)*(conjunto de treinamento). Após o treinamento, a máquina deve retornar, para cada valor de ***x***, um valor $\hat{y}$ O objetivo é retornar um valor $\hat{y}$ que esteja situado próximo à resposta ***y*** do supervisor.

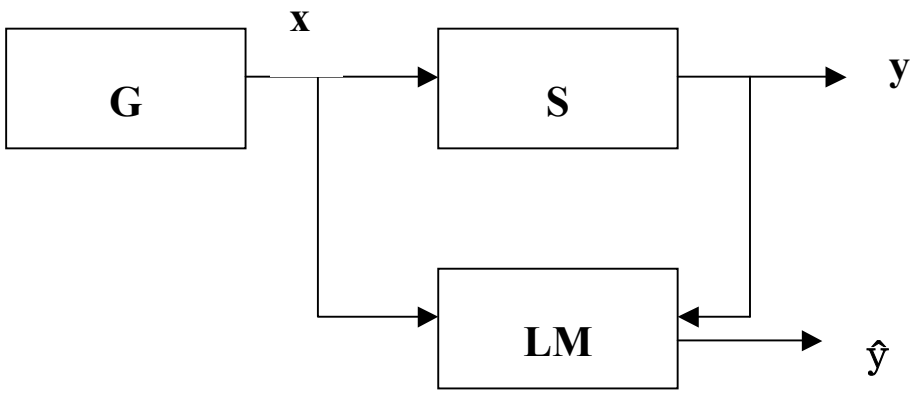


Figura 3.2 – Modelagem de aprendizagem a partir de exemplos(Vapnik, 2000).

O objetivo da modelagem é achar um modelo do espaço de hipóteses, conforme mostrado na Fig 3.3, o qual está inserido no espaço constituído pelos valores alvo, apresentando um determinado grau de erro, através de uma função subentendida oculta.
O erro de generalização do processo de aprendizagem estatística é constituído por duas componentes:

- O erro de aproximação resultante do fato de o espaço de hipóteses se apresentar menor do que o contorno do espaço alvo, fazendo com que a função oculta representativa da resposta $f_0(x)$, esteja situada externamente ao contorno do espaço de hipóteses. Este termo representa a incapacidade da máquina definida pela função $f(x,\alpha), \alpha \in \Lambda$ de aproximar com precisão a função de regressão $f(x) = E[\hat{y}/X = x]$ onde E é o operador estatístico do valor esperado(esperança matemática).
- O erro de estimação, devido ao procedimento de aprendizagem, resulta da seleção técnica não otimizada do modelo para o espaço de hipóteses. Este termo representa a não adequação da informação contida na amostra de treinamento, representado pelo conjunto $\Im = \{(x_i, y_i)\}_{i=1}^{N}$, acerca da função de regressão $f(\boldsymbol{x})$.

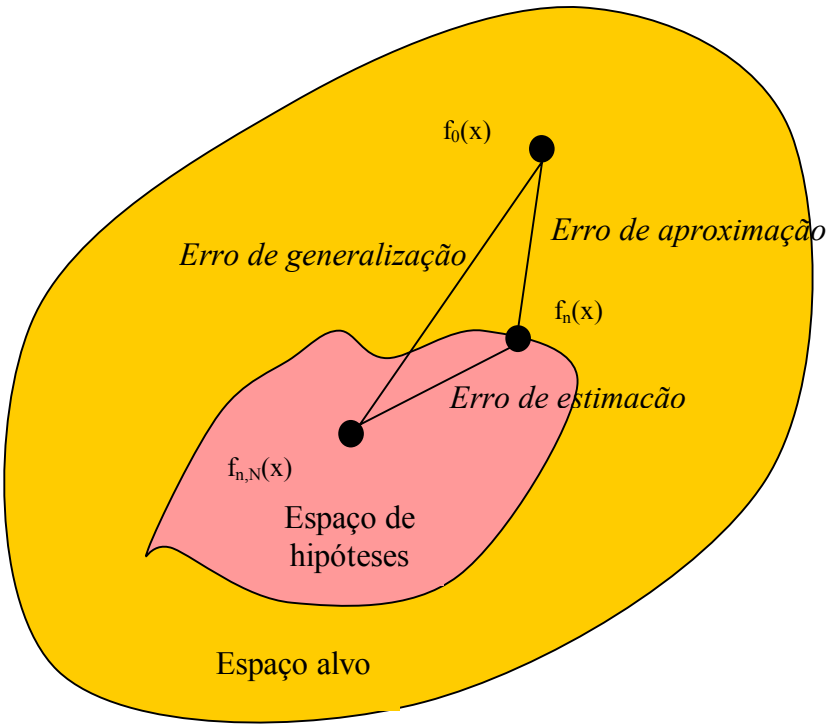


Figura 3.3 – Modelagem do erro(Gunn, 1998)

3.1.1 Minimização do risco funcional.

Com a finalidade de obter a melhor avaliação da aproximação para a resposta do supervisor **S** (Figura 3.2), pode ser medida a perda ou discrepância *c(y,f(x,w))* entre a resposta *y* do supervisor para uma dada entrada *x* e a resposta *f(x,w)* providenciada pela máquina de aprendizagem. O valor esperado da perda pode ser dado pelo *funcional de risco*(Vapnik,2000)*:*

$$R(f)=\int c(y,f(x,w))dP(x,y) \qquad [3\text{-}1]$$

O objetivo é determinar uma função $f(x,w_0)$ que seja capaz de minimizar o risco funcional $R(f)$ tendo se em vista que a função distribuição probabilística $P(x,y)$ é desconhecida e a única informação avaliável é o conjunto de treinamento, da forma:

$$(x_1,y_1),(x_2,y_2),.....,(x_\ell,y_\ell) \qquad [3\text{-}2]$$

3.1.2 Minimização do risco empírico.

Para superar a dificuldade matemática introduzida pelo desconhecimento da função distribuição probabilística $P(x,y)$, foi desenvolvido o princípio indutivo da minimização do risco empírico, definido pela equação [3-3](Vapnik,2000):

$$R_{emp}(f)=\frac{1}{\ell}\sum_{i=1}^{\ell}c(y_i,f(x_i,w)) \qquad [3\text{-}3]$$

O risco empírico $R_{emp}(\boldsymbol{f})$ difere do risco funcional $R(f)$ em dois aspectos desejáveis(Haykin,2001):

1 - O risco empírico não depende, de forma explícita, da função de distribuição desconhecida $P(\boldsymbol{x},y)$.

2 - Teoricamente ele pode ser minimizado em relação ao vetor de peso *w*.

O *princípio da minimização do risco empírico* contempla uma interpretação física. Ao iniciar o treinamento de uma máquina de aprendizagem, todas as possíveis funções aproximativas são viáveis. Com o avanço do processo de treinamento aumenta a probabilidade daquelas funções que melhor podem representar o conjunto de dados de entrada. Quando o tamanho da amostra ℓ cresce de forma a tornar o espaço de entrada densamente povoado, o ponto mínimo do funcional de risco empírico $R_{emp}(\boldsymbol{f})$, converge para o ponto mínimo do funcional de risco $R(f)$.

Para amostras pequenas do espaço de entrada, grandes desvios podem ocorrer e um excesso de ajuste(*overfitting*) poderá ocorrer, conforme mostrado na Fig 3.4. Assim um pequeno erro de generalização, dado pela equação [3-3], não poderá ser obtido pela simples minimização do erro de treinamento.

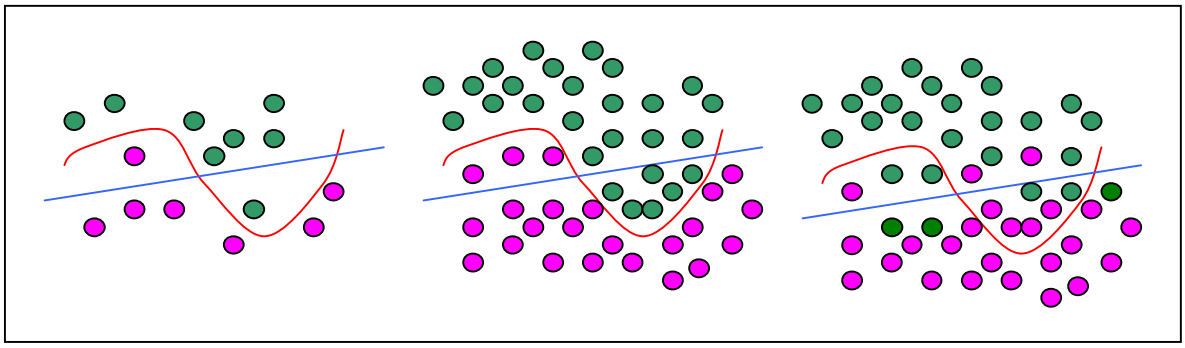


Figura 3.4 –Ilustração do dilema do ajuste em excesso(*overfitting*): Dado somente um pequeno número de exemplos(esquerda), as hipóteses em linha vermelha e azul são verdadeiras, sendo a vermelha mais complexa mas apresentando pequeno erro de treinamento. Somente com um número elevado de exemplos é possível observar qual a decisão que reflete a verdadeira distribuição de forma mais correta. Se a hipótese vermelha for a correta, a hipótese com linha azul apresenta falta de ajuste(*underfitting)*(meio); se a linha azul for a hipótese correta a hipótese vermelha apresenta excesso de ajuste(*overfitting*)(direita).(Muller, Mika, Rätsch,Tsuda, Schölkopf, 2001)

3.1.3 Dimensão VC (Vapnik-Chervonenkis).

A teoria estatística da aprendizagem ou teoria VC(Vapnik-Chervonenkis), estabelece como imperativo restringir a classe de funções de onde *f(x,w)* é obtido, para aquelas que possuam a capacidade de representar o conjunto de dados de treinamento. A teoria VC providencia limites na determinação do risco a partir do controle da complexidade da classe de funções. A minimização desses limites que depende do risco empírico e da capacidade da classe de funções, define o *princípio da minimização estrutural do risco* (Chen, Lin, Schölkopf, 2004).

A dimensão VC é um parâmetro que mede a capacidade da família de funções de classificação realizadas pela máquina de aprendizagem.

A dimensão VC de um conjunto de funções *f(x,w)* é o máximo número *h* de vetores $\mathbf{x_1,x_2,.....,x_h}$ que podem ser separados em duas classes, de todas as possíveis formas 2^h, utilizando funções do conjunto, isto é, o máximo número de vetores que pode ser fragmentado pelo conjunto de funções. Se para um dado número ℓ de pontos existir um conjunto ℓ de vetores que podem ser fragmentados pelo conjunto de funções, então a dimensão VC é igual a infinito(Vapnik,2000).

Para conceituar melhor fragmentação (separação de dados) e dimVC, considere-se que os dados de entrada estão situados em R^2, e o conjunto de funções *{ f(x,w)}* consiste de linhas. Assim para uma dada linha, os pontos situados em um lado assumem a classe +1, e os situados no outro lado assumem a classe –1.

Conforme mostrado na Fig 3.5, onde a orientação da linha indica o lado em que os pontos assumem a classe +1, podemos observar que é possível obter para três pontos não alinhados, a separação de no máximo três pontos, pelo conjunto de funções, não sendo possível obter o mesmo resultado para quatro pontos em R^2, conforme mostrado na Fig 3.6. Assim a dimVC de um conjunto de linhas orientadas em R^2 é três.

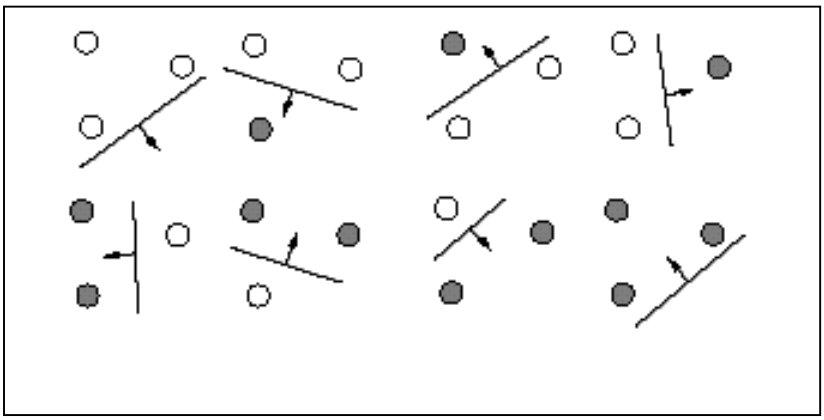


Figura 3.5 – Ilustração de três pontos não alinhados, pertencentes a duas classes diferentes, em um espaço de dimensão $\mathbf{R}^2$, separados por linhas orientadas(Burges,1998),(Markowetz, 2003)

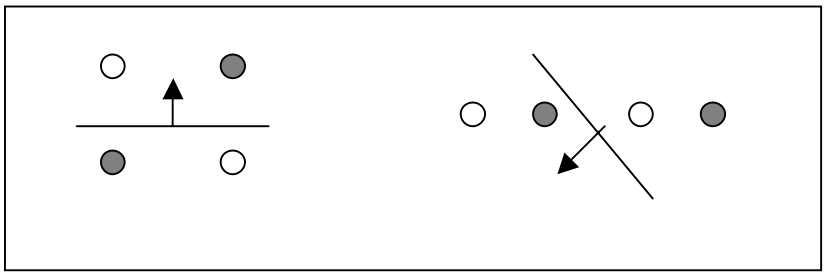


Figura 3.6 – Ilustração de quatro pontos em um espaço de dimensão $\mathbf{R}^2$, onde não é possível efetuar a separação com linhas orientadas de pontos pertencentes a classes diferentes(Markowetz,2003).

Considerando hiperplanos no espaço R^n, pode-se provar o seguinte teorema(Burges 1998):

Considere-se um conjunto de m pontos em $\boldsymbol{R}^n$. *Escolha-se qualquer um dos pontos como origem. Então, os m pontos podem ser separados por planos orientados, se e somente se, a posição vetorial dos demais pontos for linearmente independente.*

A dimVC de um conjunto de hiperplanos orientados em R^n é *n+1*, desde que sempre podemos escolher *n+1* pontos(caso da Fig 3.5), escolhendo um dos pontos como origem, de forma que a posição vetorial dos *n* pontos remanescentes seja linearmente independente, mas nunca será possível encontrar *n+2* de tais pontos(caso da Fig 3.6), uma vez que os demais *n+1* vetores de R^n não são linearmente independentes.

3.1.4 Minimização estrutural do risco.

A teoria de controle da habilidade de generalização de máquinas de aprendizagem estuda a construção de um princípio indutivo para a minimização do risco funcional utilizando dados de treinamento com pequeno número de exemplos de treinamento.
O tamanho da amostra de treinamento ℓ é considerada pequena se a relação ℓ/h(relação entre os padrões de treinamento ℓ e a dimensão Vapnik-Chervonenkis h da máquina de aprendizagem) é pequena, assim como $\ell/h < 20$(Vapnik,2000).

Para a construção de métodos com número pequeno de exemplos em que, o conjunto de funções *f(x,w)* contém número infinito de elementos e apresenta uma dimensão VC finita *h,* o risco esperado para uma probabilidade $1-\eta$ e para $\ell > h$, é dado por(Osuna, Freund, Girosi,1997):

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h\left(\ln\frac{2\ell}{h}+1\right) - \ln\left(\frac{\eta}{4}\right)}{\ell}} \qquad [3\text{-}4]$$

O objetivo é minimizar o erro de generalização *R(f), definido como a freqüência de erros cometidos pela máquina quando é testada com exemplos não vistos anteriormente(Haykin,2001),* que pode ser alcançado através da obtenção de um pequeno erro de treinamento $R_{emp}(f)$ mantendo a classe de funções menor possível. Dois extremos podem ser observados na equação [3-4]: (i) uma classe de funções muito pequena (como F_1 com dimVC=h_1, na Fig 3.7) apresenta um valor desprezível para o termo com raiz quadrada, mas mantem um erro de generalização elevado, enquanto (ii) um número elevado de classes de funções(assim como F_k na Fig 3.7 com dimVC=h_k), apresenta um erro empírico desprezível, mas um valor elevado para o termo com raiz quadrada.

A melhor classe de funções situa-se entre esses extremos, conforme mostrado na Fig 3.8, onde é possível obter uma função que represente os dados de entrada de uma forma muito boa apresentando baixo risco na sua obtenção.

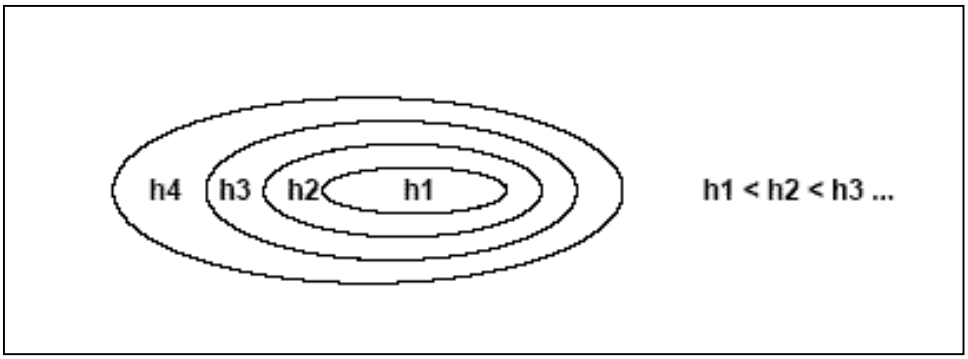


Figura 3.7 – A estrutura do conjunto de funções é determinada pelos subconjuntos de funções aninhadas $F_1 \subset F_2 \subset \ldots\ldots\ldots \subset F_k \ldots\ldots$ com dimensão VC finita com $h_1 \leq h_2 \ldots\ldots \leq h_k$ ... (Burges,1998)

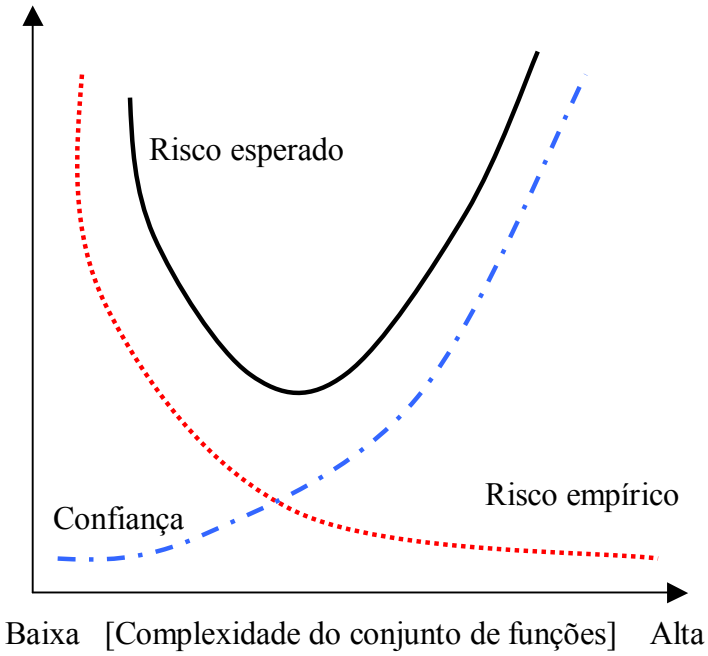


Figura 3.8 – Ilustração esquemática da equação do erro de generalização. A linha pontuada representa o erro de treinamento (risco empírico), a linha traço-ponto o limite superior do termo complexo(risco de confiança). Para o conjunto de funções de grande complexidade o erro empírico decresce enquanto o limite superior do risco de confiança é crescente. Para certa complexidade da classe de funções pode-se obter valores menores para o risco esperado (linha cheia). (Müller, Mika, Rätsch, Tsuda, Schölkopf, 2001).

3.1.5 Métodos de aprendizagem e capacidade de generalização.

De acordo com a teoria de controle da habilidade de generalização de processos de aprendizagem, para garantir um nível elevado de generalização constrói-se uma estrutura $F_1 \subset F_2 \subset ..... \subset F$, obtida do conjunto de funções perda, e, tomando um elemento apropriado F_k da estrutura e uma função pertencente a este elemento que minimize os limites correspondentes, como, por exemplo, dado pela equação [3-4], aqui reescrita de forma sintética,

$$R(f) \leq R_{emp}(f) + \Phi\left(\frac{\ell}{h}\right) \qquad [3\text{-}5]$$

onde o primeiro termo $R_{emp}(f)$ é o risco empírico e o segundo termo $\Phi(l/h)$ o intervalo de confiança.

Podem ser adotados dois modos construtivos para minimizar o lado direito da desigualdade [3-5].

O primeiro modo, na fase de projeto da máquina, consiste em estabelecer um intervalo de confiança $\Phi(\ell/h)$ fixo para a máquina de aprendizagem, para um conjunto de treinamento ℓ que apresenta uma dimensão VC dada por h.

No processo de treinamento, a máquina minimiza o primeiro termo da desigualdade [3-5] (o número de erros do conjunto de treinamento). Se para um certo conjunto de dados de treinamento o projeto da máquina se tornar muito complexo, o intervalo de confiança tornar-se-á muito grande. Neste caso mesmo que o erro empírico seja minimizado para zero, o número de erros no conjunto de teste permanecerá elevado. Este fenômeno configura o excesso de ajuste*(overfitting)*.

Para evitar o excesso de ajuste (para obter um valor pequeno para o intervalo de confiança) pode-se construir uma máquina com dimensão VC pequena, o que dificulta a

aproximação da função aos dados de treinamento (para obter um valor pequeno para o primeiro termo da desigualdade [3-5]).
Para obter um erro de aproximação pequeno e simultaneamente um intervalo de confiança pequeno, ter-se-á que achar uma arquitetura de máquina baseada em informações prévias a respeito do problema a ser resolvido.
Assim, a solução do problema consiste primeiro em determinar a arquitetura de uma máquina capaz de apresentar um compromisso entre o excesso de ajuste*(overfitting)* e uma aproximação ruim dos dados de treinamento, e, secundariamente, achar para esta máquina, a função que minimiza o número de erros de treinamento. Essa forma de resolver o problema pode ser descrita:

Mantenha o intervalo de confiança $\Phi(\ell/h)$ fixo(através da construção de uma máquina apropriada) e minimize o risco empírico $R_{emp}(f)$.

A segunda forma de resolver o problema pela minimização da desigualdade [3.5] pode ser descrita:

Mantenha o valor do risco empírico $R_{emp}(f)$ fixo(por exemplo igual a zero) e minimize o intervalo de confiança $\Phi(\ell/h)$.
Essas duas formas de implementar uma máquina de aprendizagem deram origem a duas formas de resolver o problema:

1)redes neurais (que implementam a primeira forma), e,

2)máquinas de vetor de suporte (que implementam a segunda forma).

A implementação de algoritmos de máquina de vetor de suporte é efetuada através de metodologia que se utiliza, da teoria de otimização como base de desenvolvimento teórico, com a utilização, principalmente, de multiplicadores de Lagrange para a obtenção dos vetores suporte representativos do problema em estudo.

3.2 Teoria da otimização.

Como visto anteriormente, a tarefa de aprendizagem está formulada como um problema de otimização. A função hipótese é procurada de forma a minimizar um certo funcional, a função de risco. No caso de máquinas de aprendizagem linear, a tarefa é determinar um vetor de parâmetros que minimiza(ou maximiza) certa função de custo, geralmente sujeita a certas restrições. Dependendo da forma específica da função de custo e da natureza das restrições às quais está sujeita, pode-se distinguir um certo número de classes de otimização de problemas para as quais existem estratégias para soluções eficientes. Para o caso em que a função de custo é uma função quadrática convexa com restrições lineares, o método de otimização adequado para solução é denominado *programa quadrático convexo,* o qual mostrou-se adequado para o treinamento de máquinas de vetor de suporte (Cristianini, Taylor, 2002).

3.2.1 O problema primordial.

A forma geral do problema a ser considerado é determinar o máximo ou mínimo de uma função, sujeita a algumas restrições. A forma geral em que se apresenta o problema de otimização está apresentada a seguir:

Definição 3.1: *(Problema da otimização primordial)* Dadas as funções f, g_i e h_i, definidas no domínio $\Omega \subseteq R^n$,

$$\begin{array}{lll} \text{minimizar} & f(\mathbf{w}) & \mathbf{w} \in \Omega \\ \text{sujeito às restrições} \left\{ \begin{array}{ll} g_i(\mathbf{w}) \leq 0, & i = 1,2,.....,k, \\ h_i(\mathbf{w}) = 0 & i = 1,2,.....,m, \end{array} \right. & & \end{array} \qquad [3\text{-}6]$$

sendo que $f(\mathbf{w})$ é denominada de *função objetiva,* e as demais relações, de restrição de desigualdade $g_i(\mathbf{w})$ e restrição de igualdade $h_i(\mathbf{w})$, respectivamente(Luenberger, 1973). O valor ótimo da função objetivo é denominado o *valor do problema de otimização*.

A região do domínio onde a função objetiva está definida e onde todas as restrições são satisfeitas é denominada de *região factível.* Um importante conceito é da *restrição ativa.* Uma restrição de desigualdade $g_i(\mathbf{w}) \leq 0$ é denominada *ativa* para um ponto factível $\mathbf{w}$ se $g_i(\mathbf{w}) = 0$ e *inativa* se $g_i(\mathbf{w}) < 0$. Por convenção, para qualquer restrição de igualdade $h_i(\mathbf{w}) = 0$, é classificada como *ativa*, para qualquer ponto factível $\mathbf{w}$.

A restrição ativa, para um ponto factível **w**, restringe o domínio de factibilidade nas vizinhanças de **w**, enquanto as outras restrições inativas não têm influência nas vizinhanças de **w**, conforme ilustrado na Fig 3.9, onde as propriedades locais para a solução ótima **w*** não dependem das restrições inativas $g_2(\boldsymbol{w})$ e $g_3(\boldsymbol{w})$.

Definição 3.2: (*Convexidade)* Uma função *f* definida em um conjunto convexo **Ω** é denominada *convexa* se, para cada w_1 e $w_2 \in \boldsymbol{\Omega}$ e cada α, $0 \leq \alpha \leq 1$, for satisfeita a relação[3-7]:

$$f(\alpha w_1 + (1-\alpha)\, w_2) \leq \alpha f(w_1) + (1-\alpha) f(w_2) \qquad [3\text{-}7]$$

Se, para cada $0 < \alpha < 1$ e $w_1 \neq w_2$, for satisfeita a relação[3-8], então *f* é denominada *estritamente convexa.*

$$f(\alpha w_1 + (1-\alpha)\, w_2) < \alpha f(w_1) + (1-\alpha) f(w_2) \qquad [3\text{-}8]$$

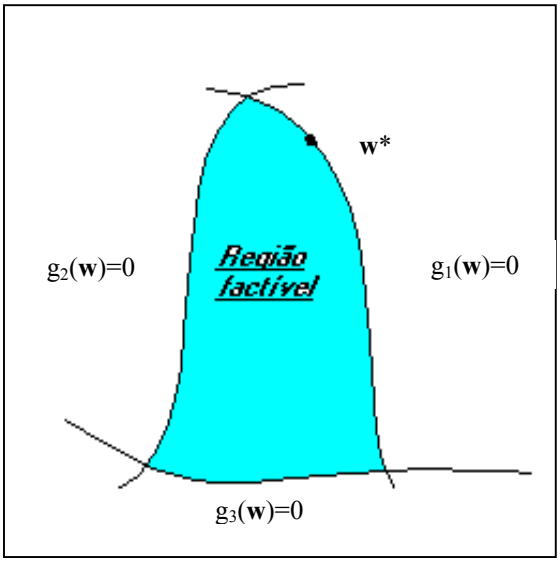


Figura 3.9 – Dependência entre a solução ótima e restrições ativas e inativas.

A solução ótima **w*** não depende das restrições inativas $g_2(\boldsymbol{w})$ e $g_3(\boldsymbol{w})$. Se for conhecida *a priori,* quais as restrições que são ativas para a solução do problema de otimização, a solução torna-se um ponto de mínimo local, definido ignorando as restrições inativas e tratando as restrições ativas como restrições de igualdade(Luenberger, 1973)

3.2.2 Teoria de Lagrange.

Na presença de problemas de otimização que não apresentam restrições, a solução da função objetiva pode ser caracterizada na forma proposta por Fermat, em 1629, e generalizada por Lagrange, em 1797(Cristianini, Taylor, 2002).

Teorema 3.1(Fermat) A condição necessária para que **w*** seja um mínimo de *f*(w), sendo *f* uma função quadrática convexa, é $\partial f(\boldsymbol{w}^*)/\partial \boldsymbol{w} = 0$, sendo, juntamente com a condição de convexidade de *f*, uma condição suficiente.

Para problemas com restrições é necessário definir uma função, conhecida por função Lagrangiana, definida pela soma da função objetiva e uma combinação linear da função de restrição, onde os coeficientes α e β são denominados de multiplicadores de Lagrange.

Definição 3.3: Dado um problema de otimização com função objetiva $f(\boldsymbol{w})$, e restrição de igualdade $h_i(\boldsymbol{w})=0$, i=1,2,....,m, define-se a função Lagrangiana, como:

$$L(w,\beta) = f(w) + \sum_{i=1}^{m} \beta_i h_i(w) \qquad [3\text{-}9]$$

Teorema 3.2 (Lagrange) A condição necessária para que um ponto normal $\boldsymbol{w}^*$ seja um mínimo de $f(\boldsymbol{w})$ sujeito à restrição $h_i(\boldsymbol{w})$, com $f, h_i \in C^1$, é:

$$\frac{\partial L(w^*, \beta^*)}{\partial w} = 0$$

$$\frac{\partial L(w^*, \beta^*)}{\partial \beta} = 0 \qquad [3\text{-}10]$$

A condição [3-10] é suficiente desde que a função $L(\mathbf{w}^*,\beta^*)$ seja função convexa de **w**. As duas condições geram dois sistemas de equações, que resolvidos juntos apresentam a solução do problema de otimização.

Definição 3.4 (Generalização de Lagrange) Dado o problema de otimização,

$$\begin{array}{lll} \text{minimizar} & f(\mathbf{w}) & \mathbf{w} \in \Omega \\ \text{sujeito às restrições} \begin{cases} & g_i(\mathbf{w}) \le 0, & i = 1,2,.....,k, \\ & h_i(\mathbf{w}) = 0 & i = 1,2,.....,m, \end{cases} \end{array} \qquad [3\text{-}11]$$

define-se a função Lagrangiana generalizada como:

$$\begin{aligned} L(w,\alpha,\beta) &= f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w) \\ &= f(w) + \alpha^t g(w) + \beta^t h(w) \end{aligned} \qquad [3\text{-}12]$$

Anteriormente definimos o problema de otimização primordial que apresenta uma função objetiva convexa e restrições lineares. Para um problema de otimização restrita como este é possível construir um outro problema chamado de *problema dual*, com a utilização dos multiplicadores de Lagrange para obter a solução ótima.

Definição 3.5 Dado o problema primordial da definição 3.1, representado pelas equações [3-6], o problema Lagrangiano dual se apresenta como:

$$\begin{array}{ll} \text{maximize} & \theta(\boldsymbol{\alpha},\boldsymbol{\beta}) = \inf_{w \in \Omega} L(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}) \\ \text{sujeito a} & \boldsymbol{\alpha} \ge 0 \end{array} \qquad [3\text{-}13]$$

A relação fundamental, entre os problemas primordial e dual, é dada pelo teorema da dualidade.

Teorema 3.3 Para uma solução viável **w** do problema de otimização primordial da Definição 3.1, e, para uma solução viável (α,β) do problema dual da Definição 3.5, existe a relação $f(w) \geq \theta(\alpha,\beta)$.

As condições para uma solução ótima de um problema de otimização geral são apresentadas pelo teorema de Kuhn-Tuchker.

Teorema 3.4 (Kuhn-Tucker) Dado um problema de otimização com domínio convexo $\Omega \subseteq R^n$,

$$\begin{array}{lll} \text{minimize} & f(w), & w \in \Omega, \\ \text{sujeito às restrições} & \begin{cases} g_i(\mathbf{w}) \leq 0, & i = 1,2,.....,k, \\ h_i(\mathbf{w}) = 0 & i = 1,2,.....,m, \end{cases} & \end{array} \qquad [3\text{-}14]$$

com $f \in C^1$ convexo e, g_i, h_i funções de restrição lineares, as condições necessárias e suficientes para que um ponto normal w^*, seja um ponto ótimo, é que devem existir valores α^* e β^*, tais que:

$$\begin{aligned} \frac{\partial L(w^*,\alpha^*,\beta^*)}{\partial w} &= 0 \\ \frac{\partial L(w^*,\alpha^*,\beta^*)}{\partial \beta} &= 0 \\ \alpha_i^* g_i(w^*) &= 0, \quad i = 1,...,k, \qquad [3\text{-}15] \\ g_i(w^*) &\leq 0, \quad i = 1,...,k, \\ \alpha_i^* &\geq 0, \quad i = 1,...,k. \end{aligned}$$

A terceira relação, é conhecida como condição complementar de Karush-Kuhn-Tucker(KKT), a qual implica que para restrições ativas $\alpha_i^* \geq 0$, enquanto que para restrições inativas $\alpha_i^* = 0$.

3.2.3 - Dualidade.

O método dual é baseado na idéia de que as variáveis duais são os fundamentos desconhecidos do problema(Cristianini, Taylor, 2002)
Podemos transformar o problema primordial em um problema dual, simplesmente igualando a zero o Lagrangiano da função objetiva com respeito às variáveis primordiais, substituindo as relações assim obtidas, novamente na função Lagrangiana, removendo a dependência com relação às variáveis primordiais. Isso corresponde explicitamente a computar a função $\theta(\alpha,\beta)= \inf_{w\in\Omega} L(w,\alpha,\beta)$. A função resultante contém somente variáveis duais e deve ser maximizada sujeita a restrição simples, estratégia adotada na teoria de máquinas de vetor de suporte.

A condição complementar KKT implica que somente restrições ativas terão variáveis duais diferentes de zero, sendo que a esses exemplos refere-se o termo *vetores suporte.*

3.3 Vetores suporte para regressão.

Com a finalidade de facilitar o entendimento sobre a natureza dos vetores suporte, será apresentada a forma de achar uma função linear aderente, capaz de aproximar com precisão ε, um conjunto de dados. A transformação em um problema de otimização convexa e a introdução de pequenas modificações, para adequar valores de erro superiores a ε, computando o problema dual de otimização convexa, resulta em um problema de programação quadrática(Smola,1998).

3.3.1 Formulação do problema primordial.

Dado um conjunto de treinamento $\{(x_1,y_1),....,(x_\ell,y_\ell)\} \subset \chi \in R$, onde χ é o espaço dos padrões de entrada, que pode assumir, por exemplo, dimensão R^d, obtido através de medidas seqüenciais independentes e identicamente distribuídos, de acordo com uma

probabilidade *dP(x,y)*. Um exemplo para o conjunto de dados pode ser o caso de valores de demanda de potência elétrica horária consumida em determinada região ou por um determinado tipo de consumidor.

O objetivo é apreender os valores do conjunto de treinamento como, por exemplo, achar uma função que apresente o menor erro quadrático, ou ainda, achar uma função *f(x)* que apresente no máximo um desvio ε em relação aos valores alvo desejados y_i, para todos os dados de treinamento do conjunto, e, ao mesmo tempo, apresente a máxima aderência possível(Smola, Schölkopf,1998). Esta função poderá apresentar erros no intervalo [y- ε , y+ ε], não sendo aceitável valores de erro superiores a ε.

A função linear *f*, toma a forma de um produto interno $\langle w, x \rangle$ em χ, como apresentado em [3-16]:

$$f(x) = \langle w, x \rangle + b \quad com \quad w \in \chi \quad e \quad b \in \mathrm{R} \qquad [3\text{-}16]$$

Para a solução do problema é necessária a transformação para um problema de otimização convexa com restrições. Determinar uma função *f* aderente quer dizer determinar um fator *w* menor possível, o que pode corresponder a minimizar a norma Euclidiana $\frac{1}{2}\|w\|^2$, satisfazendo ao mesmo tempo as restrições de erro $|f(x_i) - y_i| \leq \varepsilon$ *para todo* $i \in \{1,...,\ell\}$. O problema de otimização convexa pode ser escrito(Smola,1998):

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{sujeito à} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \qquad [3\text{-}17]$$

O equacionamento [3-17] supõe que existe uma função *f* que aproxima todos os pares (x_i,y_i), com erro máximo ε, ou em outras palavras, de que o problema de otimização convexa é factível. Algumas vezes quer-se a função *f* mesmo que com um certo grau de erro, o que pode ser obtido com a introdução de variáveis livres não negativas ξ_i e ξ_i^*, que levam em consideração aqueles pontos situados fora da margem $|f(x_i) - y_i| \leq \varepsilon$,

separando o conjunto de treinamento com um número mínimo de erros, mantendo de outra forma restrições não-factíveis do problema de otimização(Cortes, Vapnik, 1995). A formulação assume a forma(Vapnik,2000):

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\left(\xi_i + \xi_i^*\right)$$

$$\text{sujeito à} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i, & i = 1,\ldots,\ell, \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^*, & i = 1,\ldots,\ell, \\ \xi_i^* \ge 0, & i = 1,\ldots,\ell, \\ \xi_i \ge 0, & i = 1,\ldots,\ell. \end{cases} \qquad [3\text{-}18]$$

A constante C > 0 determina o compromisso entre a aderência de *f* e o montante de desvios maiores que ε tolerados.

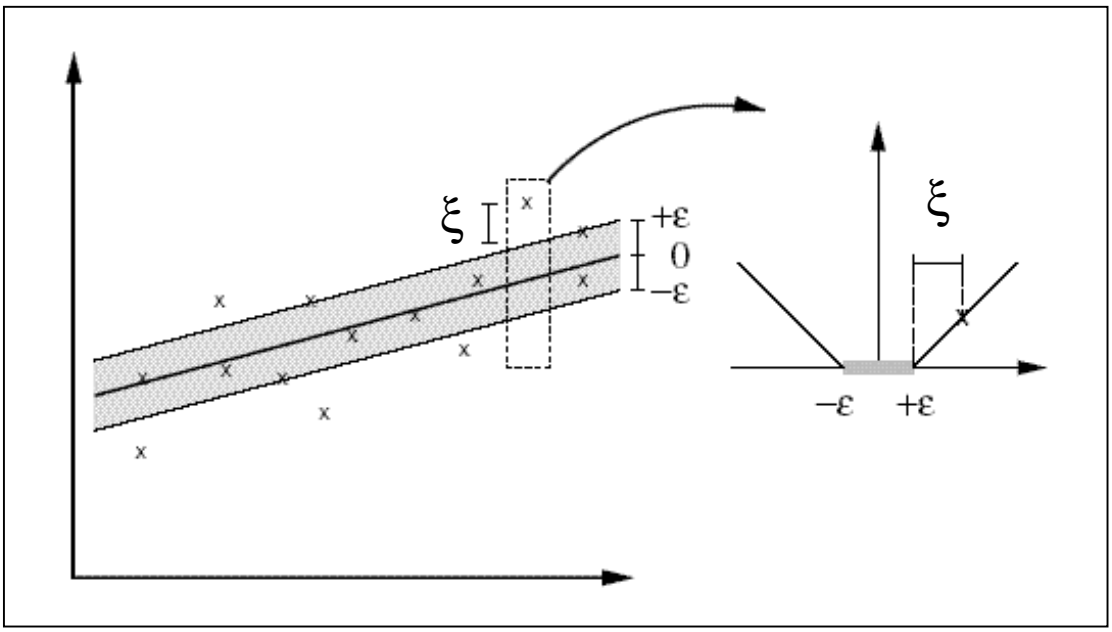


Figura 3.10 – Na regressão com vetores suporte uma precisão ε é especificada *a priori*, gerando uma região tubular com raio ε, em torno dos dados. O compromisso entre a complexidade do modelo e pontos situados fora da região tubular(com valor positivo para a variável livre ξ) é determinado pela minimização de [3.18](Schölkopf,1997).

A equação [3-18] corresponde à função de perda $|\xi|_\varepsilon$ denominada *função de perda insensível a* ε , descrita como:

$$|\xi|_\varepsilon = \begin{cases} 0 & se \quad |\xi| \le \varepsilon \\ |\xi| - \varepsilon & de \quad outra \quad forma \end{cases} \qquad [3\text{-}19]$$

A Figura 3.10 ilustra graficamente a situação apresentada pela função de perda insensível a ε, onde somente os pontos situados fora da região cinzenta contribuem para a extensão do custo, enquanto os desvios são penalizados de modo linear. A formulação apresentada em [3-18] é equivalente ao problema de determinar os valores w_ℓ e b_ℓ que minimizam o risco empírico:

$$R_{emp}(w,b) = \frac{1}{\ell}\sum_{i=1}^{\ell} |y_i - \langle w, x_i \rangle - b|_\varepsilon \qquad [3\text{-}20]$$

com a restrição $\langle w, x_i \rangle \le c_n$, que corresponde a determinar o par w e b que minimize as variáveis livres $\xi_i, \xi_i^*, i = 1,...,\ell$ em [3-18], que pode ser melhor resolvido na sua formulação dual, com a utilização de multiplicadores de Lagrange, podendo se estender para a solução de problemas não lineares.

3.3.2 – Formulação do problema dual e programação quadrática.

Como visto no item 3.2.2, a função de Lagrange é construída a partir da função objetiva a ser minimizada, menos a soma de todos os produtos entre as restrições e seus correspondentes multiplicadores de Lagrange. Otimização pode ser visto como sendo, a minimização da função de Lagrange com relação às variáveis primordiais, ou maximização em relação aos multiplicadores de Lagrange. No ponto de solução ótima,

existe um ponto de sela que, satisfaz tanto a função primordial, quanto, a função dual(Smola, Schölkopf,1998). A função Lagrangiana pode ser apresentada como:

$$\begin{aligned} L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{\ell}\alpha_i\left(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b\right) \\ - \sum_{i=1}^{\ell}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b\right) - \sum_{i=1}^{\ell}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) \end{aligned} \qquad [3\text{-}21]$$

onde α_i e α_i^* são os multiplicadores de Lagrange. O último termo no lado direito da equação [3.21], onde estão incluídas as variáveis η_i e η_i^*, é incluído para assegurar que as restrições de otimização sobre os multiplicadores de Lagrange α_i e α_i^* assumam formas variáveis. O objetivo é minimizar a função $L(w,\xi,\xi^*,\alpha,\alpha^*,\eta,\eta^*)$ em relação ao vetor w e ás variáveis livres não negativas ξ e ξ^* ,e, também maximizar em relação a α e α^*, e também em relação a η e η^* (Haykin, 2001). No ponto de sela as derivadas parciais da função *L* em relação às variáveis primordiais (w,b,ξ,ξ^*) devem ser nulas, pois é um ponto de mínimo. Assim pode-se escrever:

$$\begin{aligned} &\frac{\partial L}{\partial w} = w - \sum_{i=1}^{\ell}\alpha_i x_i + \sum_{i=1}^{\ell}\alpha_i^* x_i = 0 \quad \Rightarrow w = \sum_{i=1}^{\ell}\left(\alpha_i - \alpha_i^*\right)x_i \\ &\frac{\partial L}{\partial b} = -\sum_{i=1}^{\ell}\alpha_i + \sum_{i=1}^{\ell}\alpha_i^* = 0 \qquad \Rightarrow \sum_{i=1}^{\ell}\left(\alpha_i^* - \alpha_i\right) = 0 \\ &\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \qquad \Rightarrow \eta_i = C - \alpha_i \\ &\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \qquad \Rightarrow \eta_i^* = C - \alpha_i^* \end{aligned} \qquad [3\text{-}22]$$

Substituindo as equações [3-22] na equação [3-21], obter-se-á a função dual do problema de otimização.

$$\text{maximize} \quad Q(\alpha,\alpha^*) = \sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{\ell}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle$$

$$\text{sujeito a} \quad \begin{cases} \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C\,, \quad i = 1,\ldots,\ell \end{cases} \qquad [3\text{-}23]$$

De [3-22], reescrita, pode-se observar que o vetor ***w*** pode ser completamente descrito pela combinação linear dos padrões de treinamento x_i, que substituída na expressão [3-16], apresenta a equação [3-24] denominada de *expansão dos vetores suporte(Smola, 1998)*

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i \;\; em \;\; [3.16] \Rightarrow f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \qquad [3\text{-}24]$$

Na função aproximativa [3-24], os parâmetros livres ε e C, selecionados pelo usuário, controlam a dimensão VC do problema e devem ser sintonizados simultaneamente para otimização de problemas de regressão(Haykin,2001).

3.3.3 Determinação do bias *b*.

O valor do bias *b* pode ser determinado explorando-se as condições Karush-Kuhn-Tucker(KKT)(Karush, 1939 e Kuhn, Tucker, 1951 em Smola, Schölkopf,1998) a qual estabelece que, na solução ótima o produto entre variáveis duais e restrições desaparece, conforme mostrado na terceira equação de [3-15]. No caso de vetores suporte e considerando resultado obtido em [3-22], tem-se:

$$\begin{aligned} &\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) = 0 \quad e \quad \eta_i\xi_i = (C - \alpha_i)\xi_i = 0 \\ &\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) = 0 \qquad \eta_i^*\xi_i^* = (C - \alpha_i^*)\xi_i^* = 0 \end{aligned} \qquad [3\text{-}25]$$

Das relações [3-25] pode-se fazer várias conclusões. Somente exemplos (x_i,y_i) que correspondem a $\alpha_i = C \quad ou \quad \alpha_i^* = C$, que exigem que se tenha respectivamente $\xi_i \neq 0 \quad ou \quad \xi_i^* \neq 0$,estão locados externamente ao tubo de raio ε que gera a região factível ao redor da função *f*. O produto $\alpha_i \alpha_i^* = 0$,isto é, não se pode ter o conjunto de variáveis duais α_i, α_i^* simultaneamente não nulas, pois isto requer movimentação não nula, da função de otimização, em ambas as direções. Para que se tenha $0 \leq \alpha_i, \alpha_i^* \leq C$, $i = 1,...,\ell$ temos que ter $\xi_i = 0 \quad e \quad \xi_i^* = 0$ para satisfazer as duas equações situadas à direita de [3.25]. Dessa forma podemos computar *b* como:

$$\begin{aligned} b &= y_i - \langle w, x_i \rangle - \varepsilon \quad para \quad 0 < \alpha_i < C \\ b &= y_i - \langle w, x_i \rangle + \varepsilon \quad para \quad 0 < \alpha_i^* < C \end{aligned} \qquad [3\text{-}26]$$

De [3-25] concluímos que somente para pontos em que $\left| f(x_i) - y_i \right| \geq \varepsilon$ os multiplicadores de Lagrange α_i, α_i^* são não nulos, isto é, para todos os exemplos situados dentro do tubo de raio **ε**, mostrada na região cinzenta da Figura 3.10, os valores α_i, α_i^* desaparecem. Para $\left| f(x_i) - y_i \right| < \varepsilon$ ter-se-á $\alpha_i \xi_i = 0 \quad e \quad \alpha_i^* \xi_i^* = 0$ satisfazendo a condição KKT. Portanto, a expansão do vetor ***w*** com relação aos valores x_i do conjunto de treinamento, é esparsa, isto quer dizer, que não são necessários todos os valores x_i para descrever o vetor ***w***. Aqueles exemplos que apresentarem coeficientes não nulos são chamados vetores suporte(Smola, Schölkopf,1998).

3.3.4 Função de perda.

Para considerar dados de treinamento que apresentem ruído é necessário estabelecer funções de perda que penalizem estas singularidades. As funções devem ser convexas, de modo a não interferir na solução global do problema de convexidade da otimização, o que garante a unicidade da solução(Smola, 1996). Conforme visto no item 3.1, o risco

funcional (equação [3-1]) e o risco empírico (equação[3-3]) são definidos a partir de uma função de custo, dada por c*(y,f(x,w)),* que determina como se deseja penalizar erros de estimação, baseados nos dados de treinamento, obtidos a partir de medidas. Para muitas situações reais, os problemas de otimização a serem resolvidos são mal-representados, em que pequenos desvios da função alvo *f* geram grandes desvios na resposta do sistema, ou mesmo, para trabalhar com alguns dados em espaços dimensionais muito elevados, podendo gerar excesso de ajuste e propriedades de generalização, ruins. Essa questão é contornada com a introdução de um termo adicional na equação de risco, que apresente capacidade de controle e regularização, transformando o risco funcional e empírico(equações [3-1] e [3-3]) na equação [3-27] do risco funcional regularizado (Vapnik, 2000).

$$R_{reg}(f) = R_{emp}(f) + \frac{\lambda}{2}\|w\|^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} c\left(y_i, f(x_i, w)\right) + \frac{\lambda}{2}\|w\|^2 \qquad [3\text{-}27]$$

em que $\lambda > 0$ é a constante de regularização.

A questão posta na equação [3-27] diz respeito ao formato a ser assumido pela função custo c*(y,f(x,w)),* de forma que melhor represente as incertezas(ruídos) do problema de otimização proposto. Assim, considerando que os dados de treinamento são gerados segundo uma função de dependência somada a um ruído aditivo, da forma $y_i = f_{verdadeiro}(x_i, w_i) + \xi_i$ com densidade $p(\xi)$, a função custo ótima baseada no método da máxima probabilidade, é expressa por(Smola, 1998):

$$c(y, f(x,w)) = -\log p(f(x,w) - y) \qquad [3\text{-}28]$$

A função custo resultante desta argumentação talvez não seja convexa. Nesses casos deve-se achar uma função, já estudada, convexa, com a finalidade de tratar a situação eficientemente, qual seja, achar uma implementação eficiente para o problema de otimização convexa correspondente. A tabela de Figura 3.11 apresenta as principais funções de perda, definidas pela equação [3-28] e as funções densidade correspondentes, onde a única restrição imposta à função *c* é que para valores fixos de x_i

e y_i se tenha convexidade em $f(x_i,w)$, para garantir a existência e unicidade de um mínimo no problema de otimização(Fletcher, 1989 apud Smola, Schölkopf,1998).

Modelagem	Função de perda	Densidade de probabilidade do modelo
Insensível a ε	$c(\xi)=\lvert\xi\rvert_\varepsilon$	$p(\xi)=\frac{1}{2(1+\varepsilon)}\exp(-\lvert\xi\rvert_\varepsilon)$
Laplaciana	$c(\xi)=\lvert\xi\rvert$	$p(\xi)=\frac{1}{2}\exp(-\lvert\xi\rvert)$
Gaussiana	$c(\xi)=\frac{1}{2}\xi^2$	$p(\xi)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi^2}{2}\right)$
Huber robusta	$c(\xi)=\begin{cases}\frac{1}{2\sigma}(\xi)^2 & se \quad \lvert\xi\rvert\le\sigma \\ ou \quad \lvert\xi\rvert-\frac{\sigma}{2}\end{cases}$	$p(\xi)\propto\begin{cases}\exp\left(-\frac{\xi^2}{2\sigma}\right) & se \quad \lvert\xi\rvert\le\sigma \\ ou \quad \exp\left(\frac{\sigma}{2}-\lvert\xi\rvert\right)\end{cases}$
Polinomial	$c(\xi)=\frac{1}{p}\lvert\xi\rvert^p$	$p(\xi)=\frac{p}{2\Gamma(1/p)}\exp(-\lvert\xi\rvert^p)$
Trechos polinomiais	$c(\xi)=\begin{cases}\frac{1}{p\sigma^{p-1}}(\xi)^p & se \quad \lvert\xi\rvert\le\sigma \\ ou \quad \lvert\xi\rvert-\sigma^{\frac{p-1}{p}}\end{cases}$	$p(\xi)\propto\begin{cases}\exp\left(-\frac{\xi^p}{p\sigma^{p-1}}\right) & se \quad \lvert\xi\rvert\le\sigma \\ ou \quad \exp\left(\sigma^{\frac{p-1}{p}}-\lvert\xi\rvert\right)\end{cases}$

Figura 3.11 – Funções de perda e densidade de probabilidade dos modelos mais utilizados(Smola, Schölkopf,1998).

Assumindo que a função de custo *c* seja simétrica e que possa apresentar no máximo duas descontinuidades, em ±ε (caso da função de perda insensível a ε), e, ε≥0 na primeira derivada e igual a zero no intervalo [-ε,+ε]. Assim todas as funções de perda da tabela da Fig 3.11 pertencem a esta classe, e *c* toma a seguinte forma:

$$c(y, f(x,w)) = \begin{cases} 0 \quad para \quad |y - f(x)| \le \varepsilon \\ ou \quad \tilde{c}(|y - f(x) - \varepsilon|) \end{cases} \qquad [3\text{-}29]$$

A equação geral [3-29], se assemelha à função de perda insensível à ε (Vapnik, 2000), cujo equacionamento está apresentado em [3-19], e está representada graficamente na Fig 3.12, para um valor ε=±1.

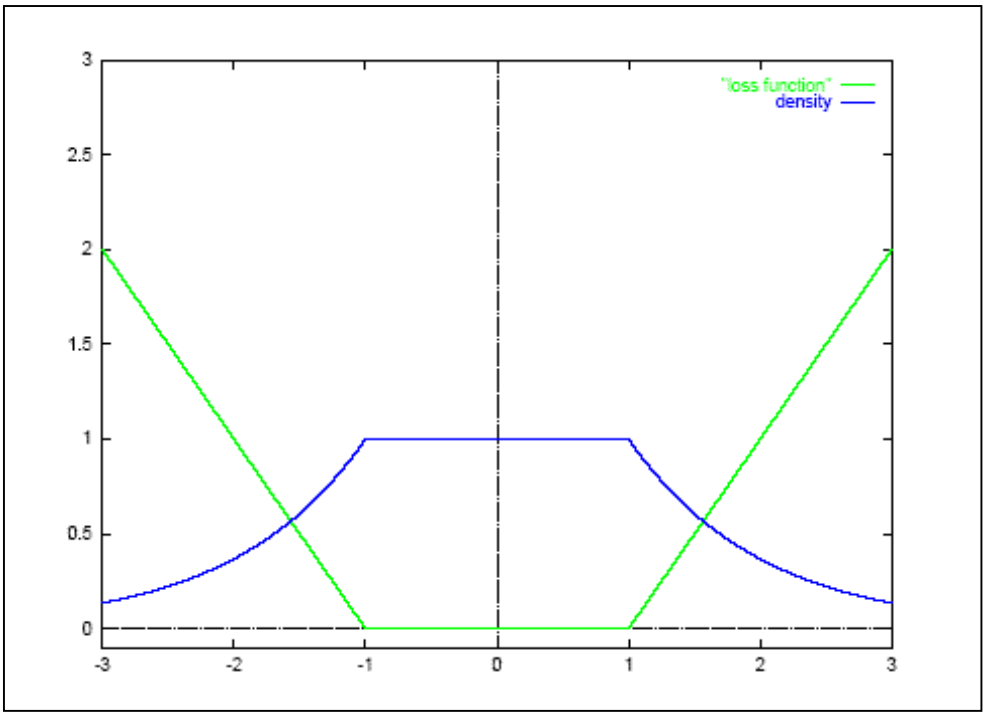


Figura 3.12 – Função de perda insensível a ε e densidade do modelo(Smola, 1996).

Para sistemas em que o ruído se distribui segundo uma distribuição normal, a função de perda pode ser representada pela distribuição de Gauss, conforme Fig 3.13.

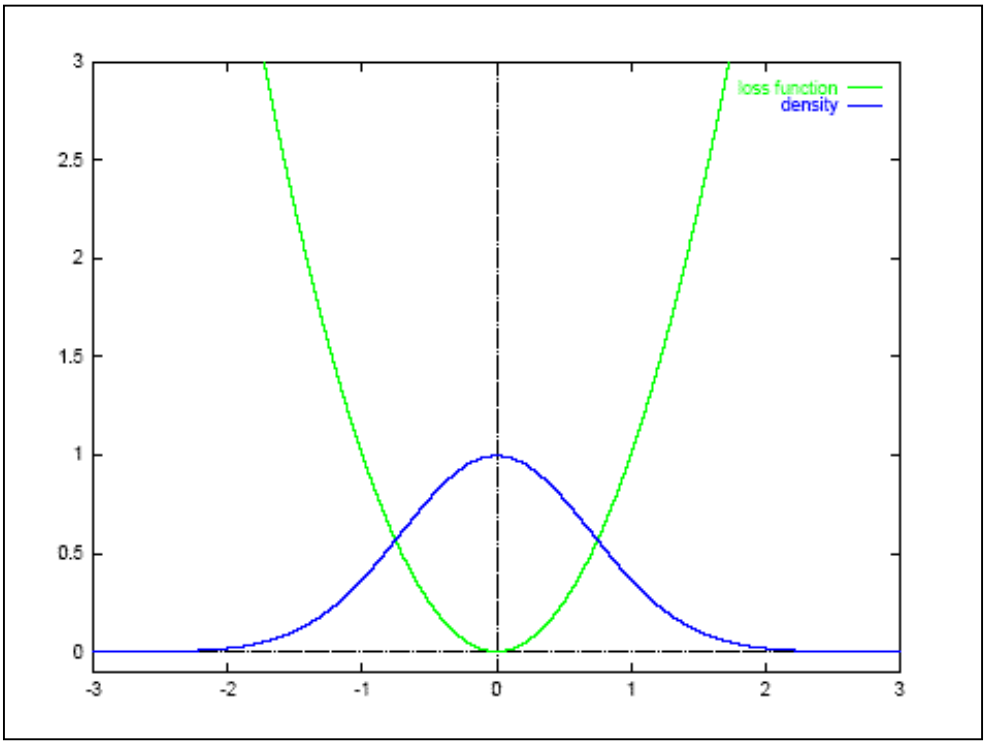


Figura 3.13 – Função de perda Gaussiana e densidade do modelo(Smola, 1996).

Para sistemas em que a variância dos dados se distribui de forma normal, utiliza-se a função de perda Laplaciana, conforme Fig 3.14, e para aqueles que apresentam dados situados entre a classe com densidade gaussiana e uma densidade simétrica arbitrária, aplica-se a função de perda Huber robusta, conforme Fig 3.15.

Da equação [3-27], utilizando *C,* ao invés de ℓ ou λ, podemos escrever o problema convexo a minimizar, de uma forma geral(Smola, Schölkopf,1998):

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\tilde{c}\left(\xi_i\right)+\tilde{c}\left(\xi_i^*\right)\right)$$

$$\text{sujeito à} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i, & i=1,\ldots,\ell, \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^*, & i=1,\ldots,\ell, \\ \xi_i^* \ge 0, & i=1,\ldots,\ell, \\ \xi_i \ge 0, & i=1,\ldots,\ell. \end{cases} \qquad [3\text{-}30]$$

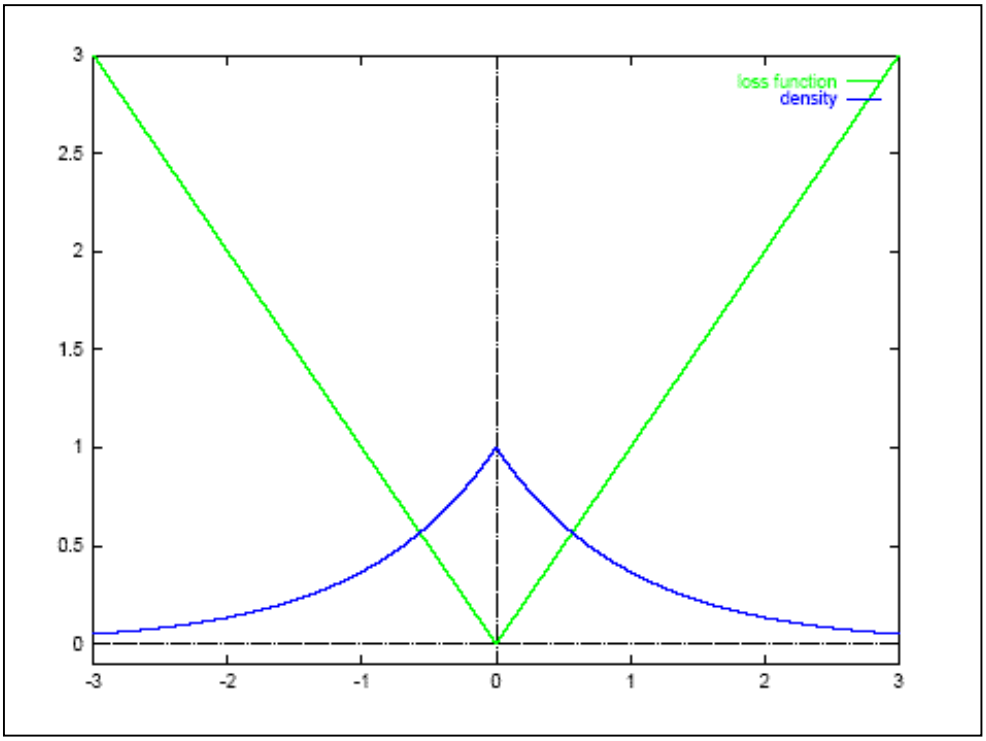


Figura 3.14 – Função de perda Laplaciana e densidade do modelo(Smola, 1996).

Computando a forma dual pela técnica dos multiplicadores de Lagrange, da mesma forma que obtido no item 3.3.2, obter-se-á o problema caracterizado pelo equacionamento [3-31], a otimizar.

$$\text{maximize} \begin{cases} Q(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i, x_j\rangle \\ \qquad\qquad + \sum_{i=1}^{\ell}\left[y_i(\alpha_i-\alpha_i^*) - \varepsilon(\alpha_i+\alpha_i^*) + C\left(T(\xi_i)+T(\xi_i^*)\right)\right] \end{cases}$$

$$\text{onde} \begin{cases} w = \sum_{i=1}^{\ell}(\alpha_i-\alpha_i^*)x_i \\ T(\xi) = \tilde{c}(\xi) - \xi\frac{\partial}{\partial\xi}\tilde{c}(\xi) \\ C\frac{\partial}{\partial\xi}\tilde{c}(\xi) = \alpha+\eta \end{cases} \text{ sujeito a } \begin{cases} \sum_{i=1}^{\ell}(\alpha_i-\alpha_i^*) = 0 \\ \alpha \le C\frac{\partial}{\partial\xi}\tilde{c}(\xi) \\ \xi = \inf\left\{\xi \middle| C\frac{\partial}{\partial\xi}\tilde{c}(\xi) \ge \alpha\right\} \\ \alpha,\alpha^*,\xi,\xi^* \ge 0 \end{cases} \quad [3\text{-}31]$$

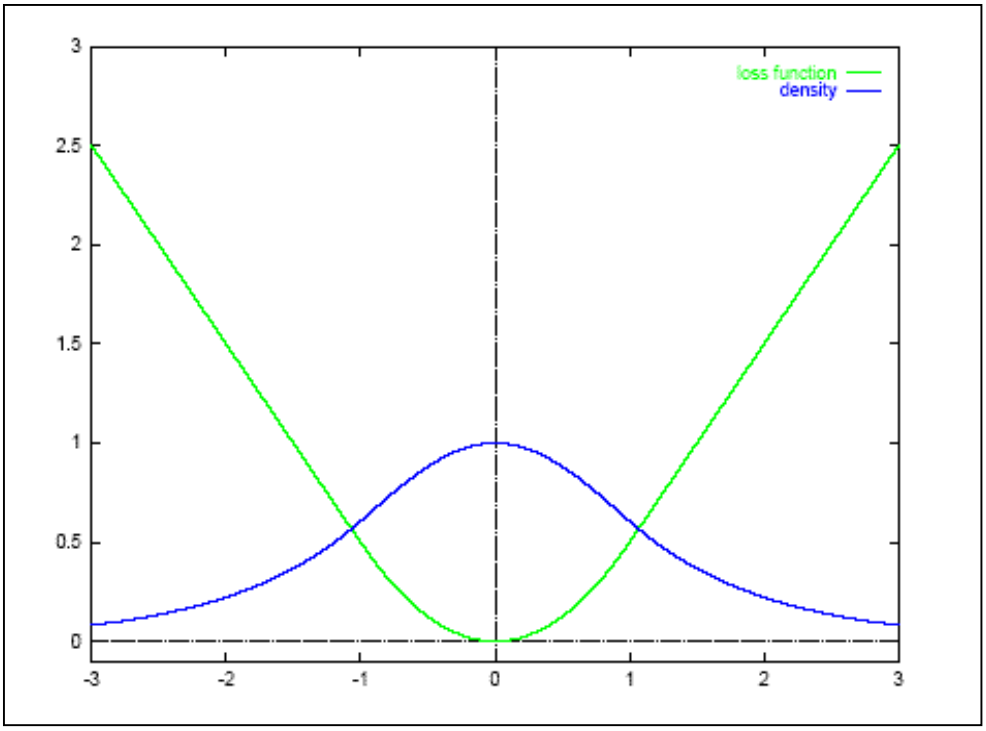


Figura 3.15 – Função de perda Huber robusta e densidade do modelo(Smola, 1996).

Para que a função objetiva da equação [3-31] possa ser resolvida, ter-se-á que simplificar a parte que depende de $\xi, \xi^*, \tilde{c}(\xi), \tilde{c}(\xi^*)$ e substituir os resultados obtidos novamente na função objetiva. Considerando somente os termos a serem substituídos, temos(Smola, 1996):

$$T(\xi) = \tilde{c}(\xi) - \xi \frac{\partial}{\partial \xi} \tilde{c}(\xi) \quad e \quad C \frac{\partial}{\partial \xi} \tilde{c}(\xi) = \alpha + \eta \quad com \quad \alpha, \xi, \eta \geq 0 \quad [3.32]$$

Para a função de perda insensível a ε, apresentada graficamente na Fig 3.12, e uma função custo, conforme tabela da Fig 3.11, na forma $c(\xi) = |\xi|$, obtém-se $T(\xi) = 0$ e $\partial \tilde{c}(\xi) / \partial \xi = 1$, e considerando a restrição de igualdade $\xi = \inf\{\xi | C \geq \alpha\} = 0$ conclui-se que $\alpha, \alpha^* \in [0, C]$. Com estas considerações e efetuando as substituições na função objetiva da equação [3-31], recaímos no problema dual de otimização da equação [3-23].

Para a função de perda Huber, mostrada na Fig 3.15, tem-se, conforme tabela da Fig 3.11, duas equações para expressar o custo. Assim para $|\xi| \le \sigma \Rightarrow \tilde{c}(\xi) = \frac{1}{2\sigma}(\xi)^2$ tem-se em [3-31]:

$$T(\xi) = \tilde{c}(\xi) - \xi \frac{\partial}{\partial \xi}\tilde{c}(\xi) = \frac{1}{2\sigma}(\xi)^2 - \xi \frac{\partial}{\partial \xi}\left(\frac{1}{2\sigma}(\xi)^2\right) = -\frac{(\xi)^2}{2\sigma}$$

$$De \quad C\frac{\partial}{\partial \xi}\tilde{c}(\xi) = \alpha + \eta \quad considerando \quad \xi = \inf\left\{\xi \middle| C\frac{\partial}{\partial \xi}\tilde{c}(\xi) \ge \alpha\right\} \quad tem-se$$

$$\eta = 0 \quad e \quad \xi = \frac{\alpha\sigma}{C} \quad que \quad aplicado \quad resulta \quad T(\xi) = T(\alpha) = -\frac{\sigma\,\alpha^2}{2\,C^2}.$$

Aplicando o resultado em [3-31] e considerando ε=0, tem-se:

$$\text{maximize} \quad \begin{cases} Q(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ +\sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) - \frac{1}{2C}\sum_{i=1}^{\ell}(\alpha_i^2\sigma_i + \alpha_i^{*2}\sigma_i^*) \end{cases}$$

[3-33]

$$\text{sujeito a} \quad \begin{cases} \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \\ \alpha, \alpha^* \in [0, C] \end{cases}$$

Para a função com trechos polinomiais, têm-se duas equações representativas dos trechos, conforme tabela da Fig 3.11.

$$No \quad primeiro \quad trecho \quad \tilde{c}(\xi) = \frac{1}{p\,\sigma^{p-1}}\xi^p \quad para |\xi| \le \sigma, \qquad então$$

$$T(\xi) = \tilde{c}(\xi) - \xi\frac{\partial}{\partial \xi}\tilde{c}(\xi) = \frac{1}{p\,\sigma^{p-1}}\xi^p - \xi\frac{\partial}{\partial \xi}\left(\frac{1}{p\,\sigma^{p-1}}\xi^p\right) = -\frac{p-1}{p}\sigma^{1-p}\xi^p$$

$$considerando \quad \xi = \inf\left\{\xi \middle| C\frac{\partial}{\partial \xi}\tilde{c}(\xi) \ge \alpha\right\} = \sigma\, C^{-\frac{1}{p-1}}\alpha^{\frac{1}{p-1}} \quad e \quad aplicando$$

$$resulta \quad T(\xi) = -\frac{p-1}{p}\sigma\, C^{-\frac{p}{p-1}}\alpha^{\frac{p}{p-1}}$$

No segundo trecho $\tilde{c}(\xi)=|\xi|-\sigma^{\frac{p-1}{p}}$ *para* $\xi \geq \sigma$, *então*

$$T(\xi)=\tilde{c}(\xi)-\xi\frac{\partial}{\partial\xi}\tilde{c}(\xi)=\left(|\xi|-\sigma^{\frac{p-1}{p}}\right)-\xi\frac{\partial}{\partial\xi}\left(|\xi|-\sigma^{\frac{p-1}{p}}\right)=\xi-\sigma^{\frac{p-1}{p}}-\xi=-\sigma^{\frac{p-1}{p}}$$

$$e\quad \xi=\left\{\xi|C\frac{\partial}{\partial\xi}\tilde{c}(\xi)\geq\alpha\right\}=\sigma\quad com\quad \alpha\in[0,C]$$

Os dois casos podem ser combinados, obtendo-se:

$$T(\alpha)=-\frac{p-1}{p}\sigma C^{-\frac{p}{p-1}}\alpha^{\frac{p}{p-1}}\quad com\quad \alpha\in[0,C]$$

Substituindo as expressões de *T(ξ)* na equação [3.31], obter-se-á uma função a maximizar na forma Q(α,α*), eliminando ξ e ξ*.

Modelagem	ε	α	$CT(\alpha)$
Insensível a ε	ε ≠ 0	α ∈ [0,C]	$CT(\alpha)=0$
Laplaciana	ε = 0	α ∈ [0,C]	$CT(\alpha)=0$
Gaussiana	ε = 0	α ∈ [0,∞)	$CT(\alpha)=-\frac{\alpha^2}{2C}$
Huber robusta	ε = 0	α ∈ [0,C]	$CT(\alpha)=-\frac{\alpha^2}{2C}$
Polinomial	ε = 0	α ∈ [0, ∞)	$CT(\alpha)=-\frac{p-1}{p}C^{-\frac{1}{p-1}}\alpha^{\frac{p}{p-1}}$
Trechos polinomiais	ε = 0	α ∈ [0,C]	$CT(\alpha)=-\frac{p-1}{p}\sigma C^{-\frac{1}{p-1}}\alpha^{\frac{p}{p-1}}$

Figura 3.16 – Termos da otimização convexa que dependem da escolha da função de perda(Smola, 1998).

Observa-se que a máxima inclinação de $\tilde{c}$ determina a região factível de α, pois $s=\sup_{\xi\in R^+}\frac{\partial}{\partial\xi}\tilde{c}(\xi)<\infty$, conduzindo a um intervalo compacto [0,C] para α. Assim

a influência de padrões isolados é limitada, propiciando estimadores robustos(Huber,1972 apud Smola, Schölkopf, 1998). Pode–se observar também que a performance da máquina de vetor de suporte depende significativamente da função de perda utilizada(Muller, Smola, Rätsch, Schölkopf, Kohlmorgen, Vapnik, 1997)(Smola, Schölkopf, Muller, 1998).

Da tabela da Fig 3.16 podemos observar a existência de dois casos diferentes. As funções de perda que se apresentam na forma de um problema de programação quadrática, tais como, as funções de Laplace, Gauss, Huber robusta e a função insensível a ε, que podem ser resolvidas por método de programação quadrática padronizado. As outras funções de custo conduzem a uma solução de problema com programação convexa, que apresenta um grau de dificuldade maior no processo de minimização(Smola, 1998).

3.4 Funções núcleo.

Para muitos problemas complexos do mundo real, requer-se hipóteses de espaço mais expressivas do que simples representação linear. Muitos problemas não podem ser resolvidos simplesmente através da combinação linear dos atributos(vetor de entrada), mas , a partir da aplicação de características mais abstratas aos dados a serem explorados(Cristianini, Taylor, 2000).

Uma boa performance na generalização de classificação de padrões pode ser determinada quando a capacidade da função de classificação é maior do que o tamanho do conjunto de treinamento. Classificar com um número elevado de parâmetros a ajustar e também com uma grande capacidade, provavelmente conduzirá a um processo de aprendizagem sem erro, mas, que apresentará um baixo desempenho de generalização. Por outro lado, classificar com capacidade insuficiente, o treinamento não consegue apreender todas as características do problema. No entanto, entre esses dois extremos existe um processo de classificação ótimo capaz de minimizar a generalização de erro esperada para um dado conjunto de dados de treinamento(Boser, Guyon, Vapnik, 1992).

A representação através de funções núcleo oferece uma alternativa de solução a partir da projeção dos dados em um espaço de características com alta dimensionalidade, incrementando a capacidade da máquina de aprendizagem linear, naquele espaço.

3.4.1 Espaço de características com alta dimensionalidade.

A complexidade da função alvo a ser apreendida depende da forma como é representada, e a dificuldade apresentada pelo processo de aprendizagem varia de acordo com essa representação. Assim, uma estratégia de pré-processamento nas máquinas de aprendizagem envolve introduzir modificações na representação dos dados, transformando os dados para um espaço de alta dimensionalidade, de uma forma não-linear, satisfazendo o teorema de Cover de separabilidade de padrões, que em termos qualitativos pode ser formulado como(Cover, 1965 apud Haykin, 2001):

Um problema complexo de classificação de padrões, disposto não linearmente, em um espaço de alta dimensionalidade tem maior probabilidade de ser linearmente separável do que em um espaço de baixa dimensionalidade.

Mapeando o espaço de vetores de entrada X dentro de um novo espaço $\vartheta = \{\phi(x) / x \in X\}$ onde $\phi : X \rightarrow \vartheta$ transforma o espaço de entrada de dimensão n em um novo espaço de alta dimensionalidade de dimensão N, conforme mostrado na Fig 3.17, que pode ser escrito:

$$x = (x_1, x_2, ..., x_n) \Rightarrow \phi(x) = \left(\phi_1(x), \phi_2(x), ..., \phi_N(x) \right) \qquad [3\text{-}34]$$

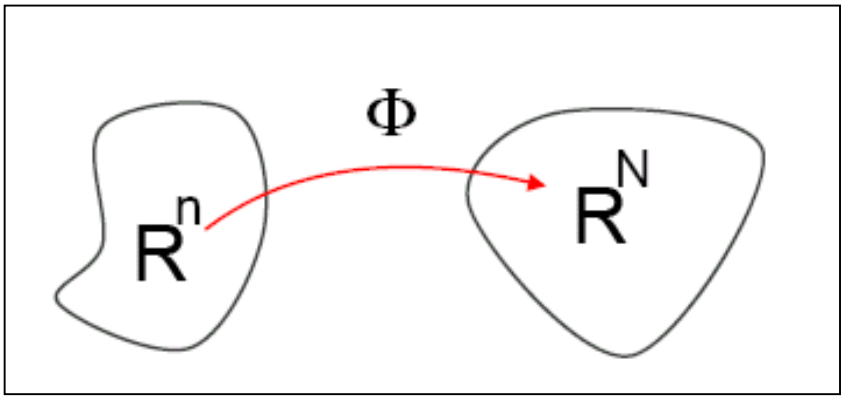


Figura 3.17 – Transformação dos dados de entrada do espaço de dimensão $\mathbf{R}^n$, para um espaço da alta dimensionalidade $\mathbf{R}^N$, através da aplicação de um mapeamento não linear **Φ**.(Smola, 1996).

Exemplificando de como através de uma determinada representação em outro espaço, um problema se transforma de não-linear em um problema de solução linear, considere-se a lei de Coulomb, dada por:

$$f(q_1, q_2, r) = K \frac{q_1 q_2}{r^2}$$

onde a lei é expressa pelas quantidades observáveis, cargas elétricas q_1, q_2 e distância entre as mesmas r. Efetuando uma transformação de coordenadas, $(q_1, q_2, r) \Rightarrow (x, y, z) = (\ln q_1, \ln q_2, \ln r)$ temos a representação nesse novo espaço: $g(x, y, z) = \ln f(q_1, q_2, r) = \ln K + \ln q_1 + \ln q_2 - 2 \ln r = k + x + y - 2z$ que pode ser apreendida por uma máquina linear.

A Fig 3.18 mostra uma ilustração de um mapeamento de características de um espaço de entrada de dimensão dois para um espaço de características da mesma dimensão, onde os dados no espaço de entrada não podem ser separados por função linear, no entanto permitem a separação linear no espaço de características.

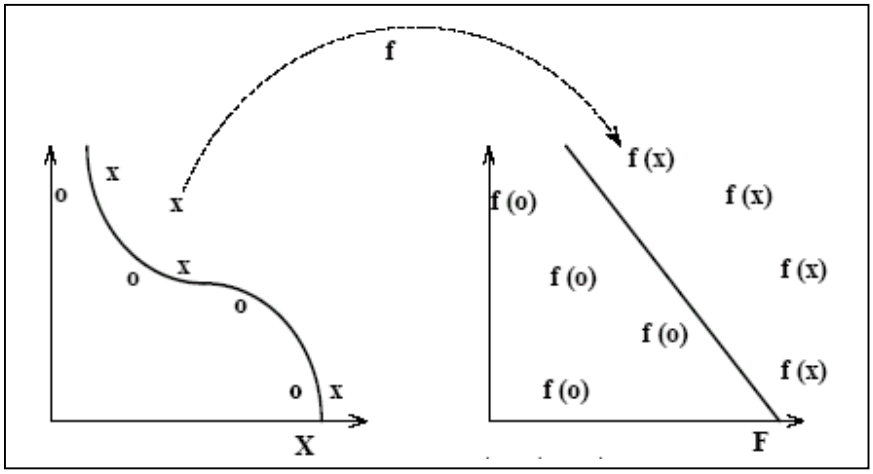


Figura 3.18 – Um mapeamento em um espaço de características pode simplificar o processo de separação de padrões(Cristianini, Taylor, 2000).

O espaço de características resultante da transformação **Φ** apresenta uma alta dimensionalidade, o que pode redundar em dificuldades na otimização do problema considerando o que se denomina de maldição da dimensionalidade, uma vez que para

funções complexas do espaço de entrada necessitar-se-á de pontos de amostra densos, para bem representá-las. A complexidade aumenta exponencialmente com a dimensionalidade, levando à deterioração das propriedades de preenchimento do espaço para pontos distribuídos aleatoriamente em espaços de dimensões mais elevadas(Haykin, 2001).

Apesar da maldição da dimensionalidade, apreender em um espaço de alta dimensionalidade, segundo a teoria estatística da aprendizagem, pode ser mais simples utilizando baixa complexidade, isto é, classes simples de regras de decisão(por exemplo classificação linear)(Muller, Mika, Rätsch, Tsuda, Schölkopf, 2001). Desta forma toda a variabilidade necessária para melhor representar os pontos de entrada e obter assim uma classe de funções capaz de representa-los, é introduzido através do mapeamento para o espaço de alta dimensionalidade **Φ,** isto é, o ponto de interesse não é a dimensionalidade, mas sim, a classe de funções(Vapnik, 2000). Intuitivamente pode-se ver na Fig 3.19, onde a superfície de decisão do espaço de dimensão dois é extremamente complicada e não-linear, enquanto no espaço de características de monômios de 2° grau a separação é um hiperplano linear.

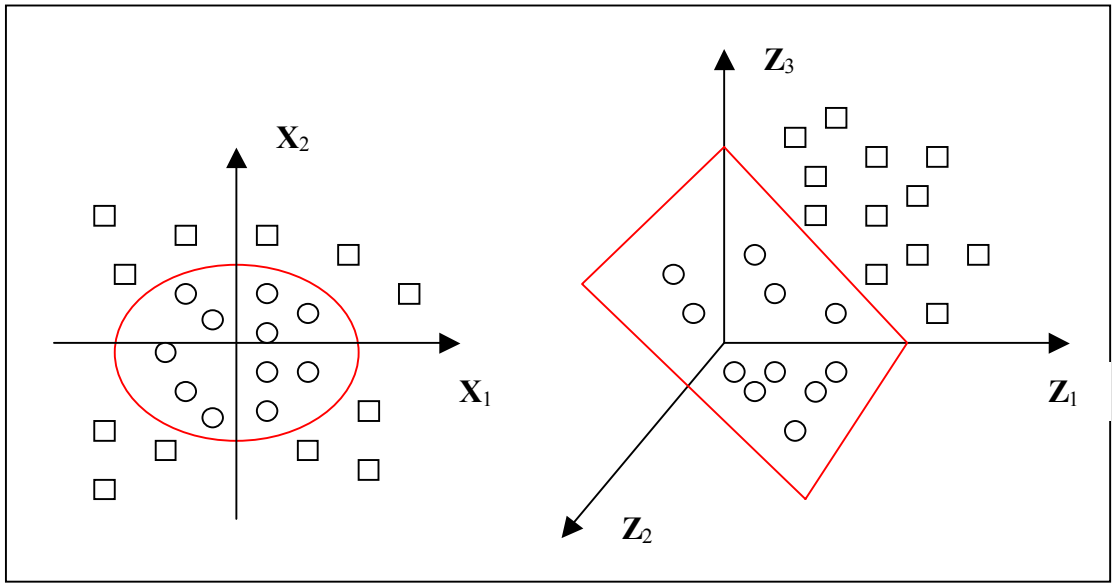


Figura 3.19 – Exemplo de classificação bidimensional. No espaço de entrada (esquerda) a superfície de decisão é elipsoidal e não-linear, enquanto, no espaço de características de monômios de 2°grau(direita), a superfície de decisão é um hiperplano linear(Muller, Mika, Rätsch, Tsuda, Schölkopf, 2001).

Considere-se o caso de um espaço de entrada com dimensão dois (x_1,x_2), e o conhecimento do problema sugere que informações relevantes podem ser obtidas na forma de monômios de 2°grau. Assim explicitando esta informação em um espaço de características, podemos obter o mapeamento Φ: $R^2 \Rightarrow R^3$ a seguir:

$$(x_1, x_2) \Rightarrow \phi(x) = \phi(x_1, x_2) = \left(x_1^2, \sqrt{2}\, x_1 x_2, x_2^2\right) \qquad [3\text{-}35]$$

Neste caso o espaço de características apresenta dimensão três, mas, para situações reais, a dimensão do espaço de características pode necessitar monômios de grau d, que para uma dimensão n do vetor de entrada, corresponde a uma dimensão do espaço de características dado por $\begin{pmatrix} d+n-1 \\ d \end{pmatrix}$, tornando intratável o problema, de controle estatístico da função, e de execução do algoritmo, nesse espaço(Runarsson, Sigurdsson, 2003).

Retomando o problema da equação [3.36], e efetuando o produto escalar entre dois vetores do espaço de entrada, $x=(x_1,x_2)$ e $z=(z_1,z_2)$, podemos escrever:

$$\begin{aligned} \langle x \cdot z \rangle^2 &= \langle (x_1, x_2) \cdot (z_1, z_2) \rangle^2 = \left(\sum_{i=1}^{2} x_i z_i \right)^2 = (x_1 x_2 + z_1 z_2)^2 \\ &= x_1^2 z_1^2 + 2\, x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= \left\langle \left(x_1^2, \sqrt{2}\, x_1 x_2, x_2^2\right) \cdot \left(z_1^2, \sqrt{2}\, z_1 z_2, z_2^2\right) \right\rangle \\ &= \langle \phi(x) \cdot \phi(z) \rangle \end{aligned} \qquad [3\text{-}36]$$

O exemplo pode ser generalizado para um espaço de entrada n-dimensional. Assim considere-se o mapeamento para um espaço de alta dimensionalidade(Taylor, Cristianini, 2004)

$$\Phi : x \to \Phi(x) = \left(x_i x_j\right)_{i,j=1}^{n} \in F = R^{n^2}$$

ter-se-á

$$\begin{aligned}\langle \phi(x), \phi(z) \rangle &= \left\langle \left(x_i x_j\right)_{i,j=1}^{n}, \left(z_i z_j\right)_{i,j=1}^{n} \right\rangle \\ &= \sum_{i,j=1}^{n} x_i x_j z_i z_j = \sum_{i=1}^{n} x_i z_i \sum_{j=1}^{n} x_j z_j \\ &= \langle x , z \rangle^2\end{aligned} \qquad [3\text{-}37]$$

Função núcleo: Função núcleo é definida como sendo uma função ***k(x,z)***, tal que para todo ***x*** e ***z*** pertencente a ***X,*** tem-se:

$$k(x,z) = \langle \phi(x), \phi(z) \rangle \qquad [3\text{-}38]$$

onde ϕ é o mapeamento de ***X*** para um(produto interno) espaço de características $\mathfrak{H}$ (Cristianini, Taylor, 2000).

A idéia da máquina de vetor de suporte é: mapear os dados de treinamento para um espaço de características com alta dimensionalidade, através da função ϕ, e construir o hiperplano de separação, ou efetuar a regressão linear, com a máxima separação, nesse espaço. Este procedimento resulta em uma superfície de decisão não linear no espaço de entrada. Utilizando a função núcleo ***k***(equação [4-5]), é possível computar o hiperplano de separação, ou a regressão linear, sem a necessidade de projetar o mapeamento no espaço de características(Smola, Bartlett, Schölkopf, Schuurmans, 2000).

Este é um ponto importante da função núcleo, pois o produto escalar é implicitamente efetuado no espaço de características, sem a necessidade de mapear ϕ explicitamente. Como conseqüência direta deste procedimento, tem-se(Schölkopf, Smola, Muller, 1998):*Todo algoritmo(linear) que somente utiliza produto escalar, pode implicitamente executa-lo no espaço de características pela utilização de funções núcleo, isto é, pode-se construir, de uma forma elegante, uma versão não-linear de um algoritmo linear.*

A máquina não-linear é construída em dois passos: utilizando um mapeamento fixo não-linear Φ, transforma-se os dados para o espaço de características, em seguida, se utiliza a máquina linear para classificar ou efetuar a regressão linear, no espaço de características.

Transformando a equação do espaço de entrada [3-16] para o espaço de características, através da função Φ(Runarsson, Sigurdsson, 2003)(Vapnik, 2000):

$$f(x) = \langle w, x \rangle + b = \sum_{i=1}^{N} w_i x_i + b \Rightarrow f(x) = \sum_{i=1}^{N} w_i \phi_i(x) + b \qquad [3\text{-}39]$$

Na forma dual não é necessário construir os pesos *w*, pois de [3-24], tem-se:

$$\begin{aligned} w &= \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \phi(x_i) \\ f(x) &= \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \phi(x_i) \phi_i(x) + b = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \langle \phi(x_i) \cdot \phi(x) \rangle + b \end{aligned} \qquad [3\text{-}40]$$

O produto escalar das características Φ não precisa ser construído, pois pode ser representado pela função núcleo ***K***:

$$f(x) = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad [3\text{-}41]$$

Todos os exemplos $i \notin SV$ *(vetores suporte)* são ignorados uma vez que α e $\alpha^*=0$

$$f(x) = \sum_{i \in SV} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad [3\text{-}42]$$

3.4.2 Propriedades das funções núcleo.

Como visto anteriormente, uma função que retorna o produto interno(Eq. 4-5) entre as imagens de dois valores de entrada em um espaço de características, é conhecida como função núcleo K. As propriedades que podem ser atribuídas às funções núcleo são aquelas que satisfazem as dos produtos internos(Taylor, Cristianini, 2004).

O espaço de produtos internos será um espaço vetorial X sobre os reais R, se existir um mapeamento bilinear de valor real, simétrico, que satisfaz:

$$\begin{aligned} \langle x, y \rangle &= \langle y, x \rangle \\ \langle x, x \rangle &\geq 0 \end{aligned} \qquad [3\text{-}43]$$

O produto interno será estrito se ⟨x,x⟩=0 se, e somente se, x=0.

Para o produto interno estrito pode-se definir a norma do espaço **X** como

$$\|x\|_2 = \sqrt{\langle x, x \rangle} \qquad [3\text{-}44]$$

A métrica associada é definida $d(x,z) = \|x - z\|_2$. Para o espaço R^n, a forma padronizada do produto interno é dado como

$$\langle x, z \rangle = \sum_{i=1}^{n} x_i z_i \qquad [3\text{-}45]$$

O espaço de produtos internos é algumas vezes referido como sendo espaço de Hilbert, requerendo adicionalmente as propriedades de serem completos e de separabilidade e às vezes requerem que a dimensão do espaço seja infinita. Formalmente tem-se:

Espaço de Hilbert ϑ : Espaço de Hilbert ϑ é um espaço de produtos internos com propriedades adicionais de serem completos e separáveis. Serem completos se refere à propriedade de cada seqüência de elementos de Cauchy $\{h_n\}_{n \geq 1}$ pertencentes a ϑ, convergir para um elemento h ∈ ϑ, sendo que a seqüência de Cauchy satisfaz a propriedade

$$\sup_{m>n} \|h_n - h_m\| \to 0 \quad n \to \infty \qquad [3\text{-}46]$$

Um espaço ϑ é separável se para cada $\varepsilon > 0$ existe um conjunto finito de elementos $h_1, h_2, \ldots, h_N$ de ϑ tal que para todo $h \in \vartheta$, tem-se

$$\min_i \|h_i - h\| < \varepsilon \qquad [3\text{-}47]$$

Exemplificando alguns casos de produtos internos.

Seqüência contável infinita - O espaço L_2 de X, constituído por todas as seqüências contáveis de números reais $\mathbf{x}=(x_1,x_2,...,x_n,...)$, tal que a soma $\sum_{i=1}^{\infty} x_i^2 < \infty$ e o produto interno entre duas seqüências **x** e **y** definido por $\langle x,z\rangle = \sum_{i=1}^{\infty} x_i y_i$.

Seqüência contável finita - Para as seqüências contáveis $\mathbf{x}=(x_1,....,x_n)^t$ e $\mathbf{z}=(z_1,.....,z_n)^t$ pertencentes a $\mathbf{X} = R^n$ e fixando um número positivo λ_i com i = 1,...,n, é válido como produto interno a expressão $\langle x,z\rangle = \sum_{i=1}^{n} \lambda_i x_i z_i = x^t \Lambda\, z$, onde $\mathbf{\Lambda}$ é uma matriz diagonal n × n com termos $\mathbf{\Lambda}_{ii}=\lambda_i$.

Funções integráveis - Tomando-se $\vartheta = L_2(X)$ como o espaço de vetores do quadrado das funções integráveis, tal que $L_2(X) = \left\{ f : \int_X f(x)^2\, dx < \infty \right\}$, se define como produto interno entre f e $g \in X$ $\langle f,g\rangle = \int_X f(x) g(x) dx$.

Proposição:Desigualdade de Cauchy-Schwartz- O espaço de produtos internos satisfaz a desigualdade de Cauchy-Schwartz $\langle x,\, z\rangle^2 \le \langle x \cdot x\rangle \langle z \cdot z\rangle = \|x\|^2 \|z\|^2$ sendo que a igualdade dar-se-á se e somente se x e z forem dependentes.

Ângulo entre dois vetores- O ângulo θ entre dois vetores x e z em um produto interno estrito é definido como $\cos\theta = \frac{\langle x\, , z\rangle}{\|x\|\|z\|}$, onde para θ=0, **x** e **z** são vetores paralelos e para θ=π, **x** e **z** são ortogonais. Um conjunto $S=\{x_1,....,x_\lambda\}$ de vetores de X é denominado ortonormal se $\langle x_i,x_j\rangle = \delta_{ij}$ (delta de Kronecker), satisfazendo $\delta_{ij}=1$ se i=j e $\delta_{ij}=0$ se i≠j.

Matriz de Gram G – Dado um conjunto de ℓ pontos com coordenadas vetoriais *n* dimensionais v_i, seja **X** a matriz $(n\times\ell)$ onde a coluna *j* é formada pelas coordenadas do vetor v_j, com $j=1,2...,\ \ell$. Então, define-se a matriz $(\ell\times\ell)$ de Gram como o produto interno $g_{ij}=v_i\cdot v_j$ obtendo(Weisstein, 2004)

$$\mathbf{G} = \mathbf{X}^t\mathbf{X} \qquad [3\text{-}48]$$

Assim, dado um conjunto $S=\{x_1,....,x_\ell\}$ de vetores pertencentes a um espaço de produto interno X, a matriz ***G*** de ordem $\ell \times \ell$ com componentes $\boldsymbol{Gij}=\langle x_i,x_j\rangle$ é definida matriz de Gram de *S.* Utilizando a função núcleo ***k*** para avaliar os produtos internos no espaço de características com mapeamento ϕ, a matriz de Gram apresentará componentes da forma:

$$G_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) \qquad [3\text{-}49]$$

A matriz resultante é denominada *matriz núcleo* e pode ser escrita na sua forma padronizada

K	1	2	.....	$\boldsymbol{\ell}$
1	$k(x_1,x_1)$	$k(x_1,x_2)$	.....	$k(x_1,x_\ell)$
2	$k(x_2,x_1)$	$k(x_2,x_2)$	.....	$k(x_2,x_\ell)$
:	:	:	:....	:
$\boldsymbol{\ell}$	$k(x_\ell,x_1)$	$k(x_\ell,x_2)$	.....	$k(x_\ell,x_\ell)$

[3-50]

A matriz de Gram é simétrica pois $\mathbf{G}_{ij}=\mathbf{G}_{ji}$, tornando $\mathbf{G}^t=\mathbf{G}$**.** A matriz de Gram contem toda a informação necessária para computar as distâncias entre pares de dados, perde , no entanto, a informação a respeito da orientação original do conjunto de dados com respeito à origem , uma vez que a matriz de produtos internos é invariante relativamente à rotações em torno da origem.

A *matriz núcleo* desempenha um papel central no desenvolvimento de algoritmos, pois, todas as informações que o algoritmo de análise de padrões pode extrair a respeito dos dados de treinamento e do espaço de características escolhido está contido na matriz, juntamente com qualquer informação de identificação dos dados.

Propriedades da matriz de Gram G(Brooks, 2004)**:**

- ***G****(X)* é uma matriz semi-definida hermitiana.

Uma matriz hermitiana A é uma matriz quadrada de valores reais que pode ser:
-Positiva semi-definida ou definida não negativa se $X^tAX \geq 0$ para todo X não nulo.
-Indefinida se $X^tAX > 0$ para alguns valores de X e <0 para outros valores de X.

- *Det(**G**(X))=0* se o determinante de um menor principal de ***G****(X)* for nulo.

Menor principal de uma matriz A é uma sub-matriz quadrada de ordem inferior ao da matriz A, cuja diagonal coincide com a diagonal da matriz principal A.

- *Posto(**G**(X))=Posto(X)*

Posto de uma matriz A (m#n) é o menor *r* para o qual existe $F_{[m\#r]}$ e $T_{[r\#n]}$ tal que A=FT.

- *Traço(**G**(X))=*$\left\|X\right\|_F$ (Norma da matriz de Frobenius).

Traço de uma matriz quadrada A é a soma dos elementos diagonais tr(A)=∑(diag(A)). A norma Euclidiana ou de Frobenius de uma matriz A é igual a $\left\|A\right\|_F = \sqrt{\sum\left|(A)^2\right|}$ e é sempre um número real.

- V_i é um autovetor de X^tX se e somente se XV_i for autovetor de XX^t. O autovalor é o mesmo para ambos os casos.

Determinar autovalores e autovetores de uma matriz A corresponde a efetuar a diagonalização da matriz. Sendo, a matriz diagonal de autovalores *D*, e a matriz de autovetores *V,* da matriz A, tem-se a relação denominada de *auto-decomposição da matriz A*.

$$[A][V] = [V][D] \qquad [3\text{-}51]$$

Desenvolvendo a expressão pode-se obter a *equação característica* para a determinação dos autovalores λ da matriz A

$$\det(A-\lambda I)=0 \qquad [3\text{-}52]$$

Matriz núcleo – As matrizes Gram e núcleo [4-17] são positivas semi-definidas, pois sendo a matriz núcleo $\boldsymbol{G}_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$ para i,j=1,..., ℓ, tem-se para qualquer vetor $\boldsymbol{v}$.

$$\begin{aligned} v^t G\, v &= \sum_{i,j=1}^{\ell} v_i v_j G_{ij} = \sum_{i,j=1}^{\ell} v_i v_j \left\langle \phi(x_i), \phi(x_j) \right\rangle \\ &= \left\langle \sum_{i=1}^{\ell} v_i \phi(x_i), \sum_{j=1}^{\ell} v_j \phi(x_j) \right\rangle = \left\| \sum_{i=1}^{\ell} v_i \phi(x_i) \right\|^2 \geq 0 \end{aligned} \qquad [3\text{-}53]$$

3.4.3 Caracterização de funções núcleo.

De acordo com [3-38] uma função núcleo define o produto interno de imagens dos dados de entrada, obtidas através de um mapeamento não-linear ϕ de dois pontos de dados, em um espaço de alta dimensionalidade.

$$k(x,z) = \langle \phi(x), \phi(z) \rangle \qquad [3\text{-}54]$$

Os dados podem ser representados na forma matricial através da composição de pares de dados de entrada e uma função núcleo adequada, resultando numa matriz núcleo (matriz de Gram) positiva semi-definida. A função núcleo define implicitamente o espaço de características de alta dimensionalidade de forma que em muitos casos dispensa a necessidade de construí-lo explicitamente, sugerindo que é possível construir núcleos sem a necessidade de construir explicitamente o espaço de características.Uma função que seja capaz de comparar dois dados de entrada a partir de um conhecimento prévio da sua aplicação, pode candidatar-se como função núcleo(Taylor, Cristinanini, 2004).

Assim, pode-se estabelecer uma característica geral para verificar se determinada função atende as propriedades de uma função núcleo, qual seja, constrói-se o espaço de características para o qual a função corresponde como uma primeira tentativa de mapeamento das características, então, efetua-se o produto interno das imagens.

Um método alternativo que pode definir se determinada função é função núcleo, utiliza a propriedade, positiva semi-definida, das funções simétricas.

Teorema de caracterização de núcleos – A função **k :** $\mathrm{X} \times \mathrm{X} \rightarrow \mathrm{R}$ a qual é contínua ou tem domínio finito, pode ser decomposta segundo $k(x,z) = \langle \phi(x), \phi(z) \rangle$, com um mapeamento de características ϕ em um espaço de Hilbert ϑ, aplicado aos dois argumentos, seguido pela avaliação do produto interno em ϑ, se e somente se, satisfizer a propriedade de ser finito positivo semi-definido.

Apreender em um espaço de características de alta dimensionalidade, é equivalente a identificar uma função naquele espaço, que pode ser da forma

$$\vartheta = \left\{ \sum_{i=1}^{\ell} \alpha_i k(x_i, x) : \ell \in N, x_i \in X, \alpha_i \in R, \ \mathrm{i} = 1,...,\ell \right\} \qquad [3\text{-}55]$$

Introduzindo os vetores *f* e *g* pertencentes a *F* dados por

$$f(x) = \sum_{i=1}^{\ell} \alpha_i k(x_i, x) \quad e \quad g(x) = \sum_{j=1}^{\ell} \beta_j k(z_j, x) \qquad [3\text{-}56]$$

define-se

$$\langle f,g\rangle=\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\beta_j k(x_i,z_j)=\sum_{i=1}^{\ell}\alpha_i g(x_i)=\sum_{j=1}^{\ell}\beta_j f(z_j) \qquad [3\text{-}57]$$

onde, a segunda e terceira igualdades são obtidas das definições de *f* e *g* em [3-55].

O produto interno $\langle f,g\rangle$, tem como resultado um valor real, simétrico e bilinear, tendo-se para todo $f\in \vartheta$

$$\langle f,f\rangle=\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j k(x_i,x_j)=\alpha^t K\alpha\geq 0 \qquad [3\text{-}58]$$

sendo $\boldsymbol{\alpha}$ um vetor com elementos α_i i=1,...,ℓ, e $\boldsymbol{K}$ a matriz núcleo construída com os valores $x_1,x_2,...,x_\ell$.

Fazendo $g=k(\boldsymbol{x},x_i)$ na equação [3-57] ter-se-á uma expressão conhecida como *propriedade de reprodução* do núcleo.

$$\langle f,g\rangle=\langle f,k(x,x_i)\rangle=\sum_{i=1}^{\ell}\alpha_i k(x_i,x)=f(x) \qquad [3\text{-}59]$$

Introduzindo as propriedades de serem completos e separáveis apresentadas em [3-46] e [3-47] ter-se-á obtido o espaço de Hilbert associado ao núcleo $\boldsymbol{k}$.

Assim a construção do espaço de características em um espaço de Hilbert, necessita que seja especificada a imagem de uma entrada **x** sob a função de mapeamento ϕ,

$$\phi: x\in X\rightarrow\phi(x)=k(x,x_i)\in \vartheta_K \qquad [3\text{-}60]$$

Avaliando o produto interno entre um elemento de ϑ_K e a imagem de uma entrada $\boldsymbol{x}$, utilizando a equação [3-59] te-se-á

$$\langle f,\phi(x)\rangle=\langle f,k(x,x_i)\rangle=f(x) \qquad [3\text{-}61]$$

A expressão [3-61] mostra que a função *f* pode realmente representar uma função linear definida por um produto interno, dela mesmo, no espaço de características ϑ_K.

Dada uma função *k* que satisfaz a propriedade finita positiva semi-definida, referir-se-á ao espaço correspondente ϑ_K como o Espaço de Hilbert de Núcleo Reproduzível*(Reproducing Kernel Hilbert Space –RKHS)*.

O teorema de Mercer é normalmente utilizado para a construção de um espaço de características para uma função núcleo válida. Com a construção a partir do RKHS, o teorema de Mercer completa a teoria de construção de núcleos, uma vez que ele define o espaço de características a partir de um vetor de características, ao invés da construção com RKHS.
O teorema de Mercer providencia a necessária caracterização de quando uma função do tipo ***k(x,z)*** é uma função núcleo. Formalmente pode ser apresentado como(Mercer, 1909 e Courant, Hilbert, 1970 apud Haykin, 2001):

Teorema de Mercer - *Seja* ***k(x,z)*** *um núcleo simétrico e contínuo que é definido no intervalo fechado* $\boldsymbol{a \le x \le b}$. *O núcleo* ***k(x,z)*** *pode ser expandido em série conforme*

$$k(x,z) = \sum_{i=1}^{N_\tau} \lambda_i \varphi_i(x) \varphi_i(z) \qquad [3\text{-}62]$$

onde $N_\tau \le \infty$ *é o número de autovalores positivos* λ_i *e* $\varphi_i(x)$ *são as autofunções da expansão . Para esta expressão ser válida e para convergir absoluta e uniformemente, é necessário e suficiente que a condição*

$$\int_b^a \int_b^a k(x,z) g(x) g(z) dx dz \ge 0 \qquad [3\text{-}63]$$

seja válida para toda função ***g(.)*** *para a qual*

$$\int_b^a g^2(x) dx < \infty \qquad [3\text{-}64]$$

Como os autovalores são todos positivos, significa que a função núcleo ***k(x,z)*** *é definida positiva.*

Uma das características importantes das funções núcleos é a sua modularidade. O mesmo algoritmo pode trabalhar com qualquer núcleo e em qualquer domínio. A matriz núcleo, **K,** que é totalmente desvinculada do processo de solução do problema, pode ser reutilizada para diferentes algoritmos, uma vez que os procedimentos de solução são adaptados para utilizar somente produtos escalares entre valores de entrada, e, a matriz núcleo é formada com os produtos escalares das imagens de dois valores no espaço de características, possibilitando a implementação do algoritmo em um espaço de alta dimensionalidade. A Fig 3.20 mostra os estágios envolvidos na implementação da análise de padrões através de núcleos. Os dados são processados utilizando núcleos que constroem a matriz núcleo, **K**, que por sua vez é utilizada pelo algoritmo de análise de padrões para gerar a função do padrão a ser analisado.

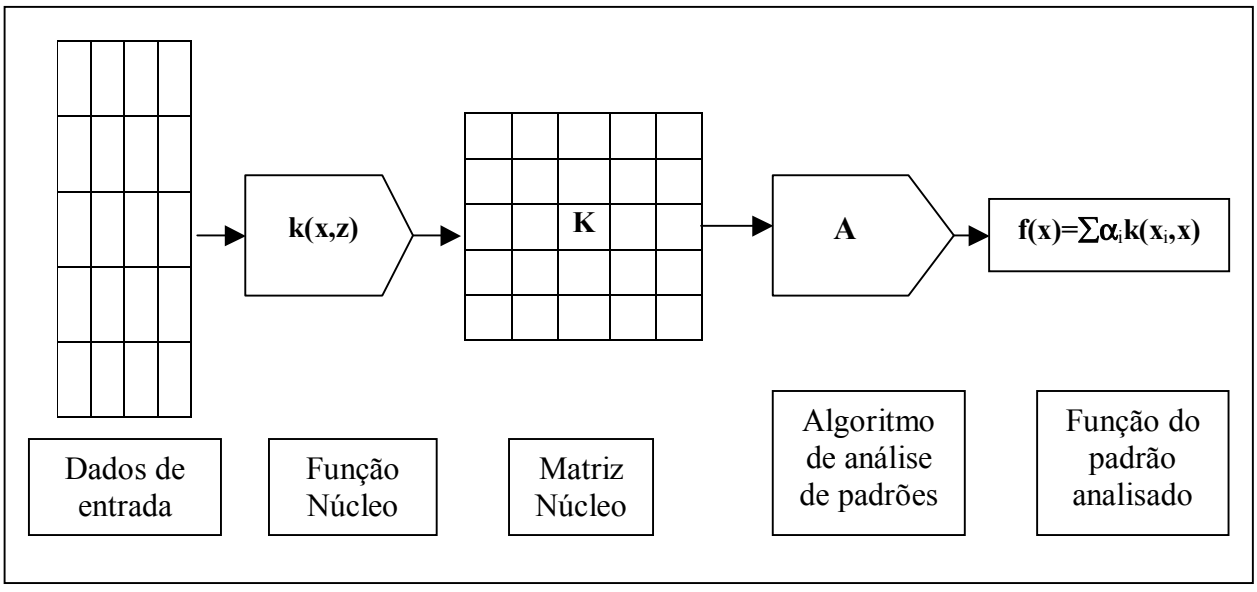


Figura 3.20 – Os estágios envolvidos na aplicação dos métodos com função núcleo(Taylor, Cristianini, 2004).

3.5 Construção de núcleos.

As propriedades apresentadas na seção anterior são válidas para decidir se determinada função pode ser uma função núcleo, ou seja, se cumpre com a propriedade de ser finita positiva semi-definida. Uma das conseqüências desse procedimento é que é possível gerar, a partir da combinação de núcleos simples, funções núcleo mais complexas. A geração de novas funções núcleo pode ser efetuada preservando a propriedade dos núcleos, de ser finita positiva semi-definida, determinando que a classe de funções núcleo assim obtidas é *fechada* sob tal operação.

Definindo operações com funções núcleos, pode-se gerar núcleos complexos que satisfazem várias propriedades de serem fechados.

Núcleos com propriedades de serem fechados: Seja k_1 e k_2 núcleos em $\mathbf{X} \times \mathbf{X}$, $\mathbf{X} \subseteq \mathrm{R}^n$, $a \in \mathrm{R}^+$, $f(\cdot)$ função de valor real em $\mathbf{X}$, ϕ: $\mathbf{X} \rightarrow \mathrm{R}^m$, com o núcleo k_3 definido em $\mathrm{R}^m \times \mathrm{R}^m$ e $\mathbf{B}$ matriz simétrica positiva (semi) definida de ordem $n \times n$. As seguintes funções são núcleos(Runarsson, Sigurdsson, 2003):

- $k(\mathbf{x},\mathbf{z}) = k_1(\mathbf{x},\mathbf{z}) + k_2(\mathbf{x},\mathbf{z})$
- $k(\mathbf{x},\mathbf{z}) = a + k_1(\mathbf{x},\mathbf{z})$
- $k(\mathbf{x},\mathbf{z}) = a\ k_1(\mathbf{x},\mathbf{z})$
- $k(\mathbf{x},\mathbf{z}) = k_1(\mathbf{x},\mathbf{z})\ k_2(\mathbf{x},\mathbf{z})$
- $k(\mathbf{x},\mathbf{z}) = f(\mathbf{x})\ f(\mathbf{z})$ [3-65]
- $k(\mathbf{x},\mathbf{z}) = exp(k_1(\mathbf{x},\mathbf{z}))$
- $k(\mathbf{x},\mathbf{z}) = p(k_1(\mathbf{x},\mathbf{z}))$ onde p é um polinômio com coeficientes positivos.
- $k(\mathbf{x},\mathbf{z}) = k_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$
- $k(\mathbf{x},\mathbf{z}) = \mathbf{x}^t \mathbf{B}\ \mathbf{z}$

3.5.1 – Núcleos para estimar funções de valor real.

Função núcleo polinomial: A função núcleo polinomial, derivada do núcleo k_1 da forma $k(\mathbf{x},\mathbf{z}) = p(k_1(\mathbf{x},\mathbf{z}))$, se refere geralmente ao caso especial da forma(Taylor, Cristianini, 2004)

$$k_d(x,z) = \left(\langle x,z \rangle + R\right)^d \qquad [3\text{-}66]$$

definido sobre um espaço $\boldsymbol{X}$ de dimensão n, gerando um espaço de características de dimensão $\binom{n+d}{d}$, onde R e d são parâmetros. Expandindo a função polinomial k_d na forma binomial ter-se-á

$$k_d(x,z)=\sum_{s=0}^{d}\binom{d}{s}R^{d-s}\langle x,z\rangle^{s} \qquad [3\text{-}67]$$

O parâmetro R exerce um certo controle relativo sobre os pesos dos diferentes graus dos monômios, uma vez que aumentando R, decresce o peso relativo para os termos de graus superiores do polinômio.

Pode-se recursivamente computar o núcleo polinomial de grau d em termos de núcleos de graus inferiores, utilizando a formulação

$$k_d(x,z)=k_{d-1}(x,z)\left(\langle x,z\rangle+R\right) \qquad [3\text{-}68]$$

Estendendo a formulação recursiva para a dimensionalidade n da entrada, pode-se escrever:

$$k_s^m(x,z)=\left(\langle x_{1:m},z_{1:m}\rangle+R\right)^s=s\,k_{s-1}^m(x,z)x_m z_m+k_s^{m-1}(x,z) \qquad [3\text{-}69]$$

sendo que, $\boldsymbol{x_{1:m}}$ indica a restrição em $\boldsymbol{x}$ para as m primeiras características, separando os termos em dois grupos conforme indicado na soma em [3-69].

Função núcleo gaussiana: Os núcleos gaussianos são os mais extensamente utilizados e se apresentam na forma:

$$k(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)=\exp\left(-\frac{k_1(x,x)-2k_1(x,z)+k_1(z,z)}{2\sigma^2}\right) \qquad [3\text{-}70]$$

O parâmetro σ controla a flexibilidade do núcleo da mesma forma que o grau d no núcleo polinomial. Valores pequenos de σ correspondem a valores elevados de d, como por exemplo, eles permitem efetuar o ajuste de qualquer rótulo sem o risco de haver ajuste em excesso(*overfitting*). Nesses casos, a matriz, núcleo $\boldsymbol{K}$, se aproxima da matriz identidade $\boldsymbol{I}$. Por outro lado valores elevados de σ, reduzem gradualmente o núcleo para uma função constante, tornando impossível apreender qualquer classificação não-trivial.

O espaço de características apresenta dimensão infinita para todos os valores de σ, mas para valores elevados, o peso decresce rapidamente para características de ordem elevada, mostrando que mesmo que a matriz, núcleo, se apresente cheia, para efeito prático, os pontos se localizam em um sub-espaço de baixa dimensão do espaço de características.

O problema de aproximação de funções de valor real é um problema mais difícil de ser efetuado do que o problema de classificação ou de aproximação de funções indicadas(Vapnik, 2000)(Haykin, 2000).

Muitos problemas de estimação de funções de valores reais necessitam de conjuntos de funções de aproximação, razão, pela qual torna-se necessária a construção de funções núcleo que apresentem propriedades especiais para aproximação dessas funções.

Considere-se o núcleo:

$$k(x,z)=\sum_{i=1}^{\infty} r_i \psi_i(x)\psi_i(z) \qquad [3\text{-}71]$$

sendo que r_i converge para zero com o acréscimo de i, que define a expansão de um núcleo polinomial regularizado.

Seja o exemplo dos polinômios uni-dimensionais de Hermite, que são conjuntos de polinômios ortogonais no domínio (-∞,+∞), com funções de peso e^{-x^2}, ilustrado na Fig 3.21 e equacionamento:

$$H_k(x)=(-1)^k \mu_k e^{x^2} \frac{d^k\left(e^{-x^2}\right)}{dx} \qquad [3\text{-}72]$$

onde μ_k é uma constante de normalização.

Deste polinômio pode-se obter o núcleo, tomando $r_i=q^i$, *com* $0 \le q \le 1$:

$$k(x,z)=\sum_{i=0}^{\infty} q^i H_i(x)H_i(z)=\frac{1}{\sqrt{\pi\left(1-q^2\right)}}\exp\left(\frac{2xzq}{1+q}-\frac{(x-z)^2 q^2}{1-q^2}\right) \qquad [3\text{-}73]$$

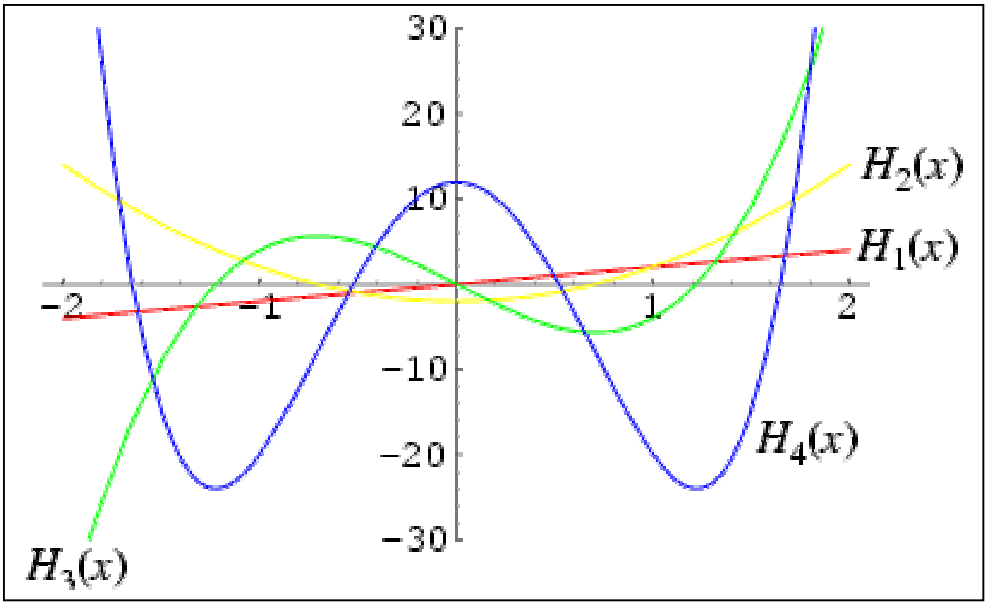


Figura 3.21 – Ilustração dos polinômios de Hermite para x ∈ [0,1] e n=1,2,3,4.

(Fonte: mathworld wolfram)

Para construir núcleos para a aproximação de funções multidimensionais definidas em um espaço em que todas as coordenadas estão localizadas no mesmo intervalo, finito ou infinito, temos o seguinte teorema como verdadeiro.

Núcleo de funções multidimensionais: *Seja um conjunto de funções multidimensionais, definidas pelas funções base, que são obtidas do produto tensorial das coordenadas das funções bases, então o núcleo que define o produto interno, na base dimensional n, é o produto de n núcleos unidimensionais (Vapnik, 2000).*

Aplicando o teorema ao nosso problema, e construindo a função para um espaço n-dimensional, pode-se escrever:

$$k(x,z)=\prod_{i=1}^{n}\frac{1}{\sqrt{\pi\left(1-q^2\right)}}\exp\left\{\frac{2x^i z^i q}{1+q}-\frac{\left(x^i-z^i\right)^2 q^2}{1-q^2}\right\}$$
$$=\frac{1}{\left[\pi\left(1-q^2\right)\right]^{n/2}}\exp\left\{\frac{2\langle x\cdot z\rangle q}{1+q}-\frac{|x-z|^2 q^2}{1-q^2}\right\} \qquad [3\text{-}74]$$

3.5.2 - Núcleos que geram superfícies geradoras flexíveis(splines).

Splines são trechos polinomiais construídos de forma que as peças são conectadas juntas de forma aderente suave. Para um spline de grau *p,* cada segmento, separado por nós, é um polinômio de grau *p*, necessitando-se de *p+1* coeficientes polinomiais para descrever cada trecho(Unser, 1999). A curva B-spline é obtida a partir da definição de um vetor de nós $T=\{t_0,t_1,....t_m\}$ que é uma seqüência não decrescente com $t_i \in [0,1]$ e de uma sucessão de pontos de controle $P_0,P_1....P_n$, conforme mostrado na Fig 3.22.

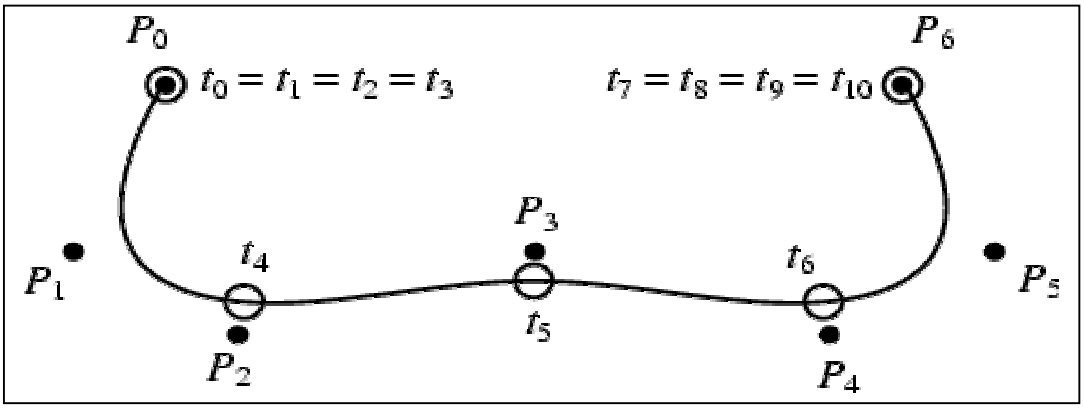


Figura 3.22 – Vetor de nós $T(t_1,....,t_{10})$ e pontos de controle $P_1,....,P_5$ da aproximação de uma curva B-spline entre os pontos P_0 e P_6 (Fonte: mathworld wolfram).

A função B-splines é construída a partir de blocos básicos splines, e provém da convolução de um pulso retangular (B spline de grau zero). Pode obter-se a forma do B-spline a partir da equação

$$B^n(x)=\frac{1}{n!}\sum_{k=0}^{n+1}\binom{n+1}{k}(-1)^k\left(x-k+\frac{n+1}{2}\right)^n \qquad [3\text{-}75]$$

A complexidade da solução computacional depende do número de vetores suporte, necessários para aproximar a função procurada com a precisão ε, ao invés da dimensionalidade do espaço ou do número de nós(Vapnik, 2000).

Seja construir um núcleo para aproximar a função uni-dimensional no intervalo [0,α] com superfícies flexíveis de ordem $d \geq 0$ com *m* nós igualmente espaçados

$$(t_1,...,t_m), \quad t_i = \frac{ia}{m}, \quad i = 1,..,m \qquad [3\text{-}76]$$

Por definição, uma aproximação com superfícies flexíveis têm a forma:

$$f(x) = \sum_{r=0}^{d} a_r^* x^r + \sum_{i=1}^{m} a_i (x - t_i)^d \qquad [3\text{-}77]$$

Seja o mapeamento da variável *x* unidimensional, em um vetor *u* de dimensão(m+d+1):

$$x \rightarrow u = (1, x, ..., x^d, (x - t_1)_+^d, ..., (x - t_m)_+^d)$$

$$onde \quad (x - t_k)_+^d = \begin{cases} 0 & se \quad x \le t_k \\ (x - t_k)^d & se \quad x > t_k \end{cases} \qquad [3\text{-}78]$$

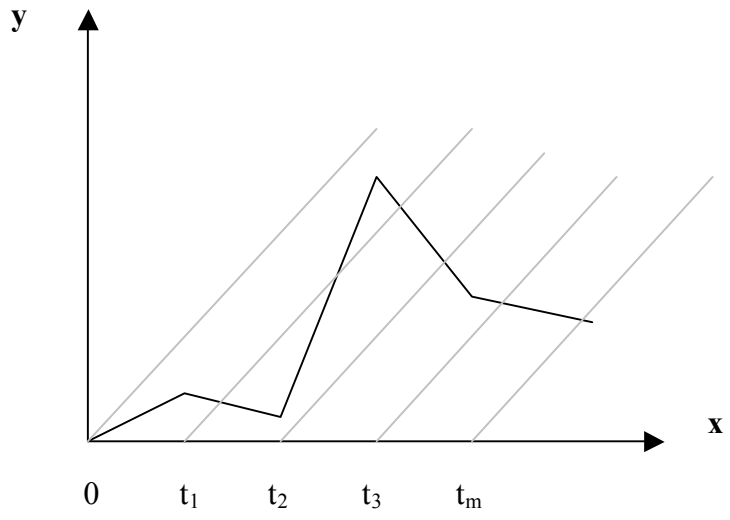


Figura 3.23 – Utilizando uma expansão em funções da forma *1, x, $(x-t_1)_+$,...$(x-t_m)_+$*, pode-se construir trechos lineares de aproximação da função. Da mesma forma uma expansão em funções da forma *1, x,...,x^d,$(x-t_1)_+^d$,....$(x-t_m)_+^d$* providencia aproximação na forma de trechos polinomiais (Vapnik,2000).

A equação [3-77] pode ser considerada como sendo o produto interno $f(x)=\langle a \cdot u \rangle$ e pode-se definir o núcleo que gera o produto interno no espaço de características como

$$k(x,z)=\langle u \cdot u_t \rangle=\sum_{r=0}^{d} x^r z^r+\sum_{i=1}^{m}(x-t_i)_+^d (z-t_i)_+^d \qquad [3\text{-}79]$$

Para a construção de núcleos que gerem superfícies flexíveis em um espaço n-dimensional efetua-se o produto, de n vezes, do núcleo unidimensional.

$$k(x,z)=\prod_{k=1}^{n} k(x^k, z^k) \qquad [3\text{-}80]$$

Para um número infinito de nós no intervalo (0,a), com $0 < a < \infty$ temos a expansão da função:

$$f(x)=\sum_{i=0}^{d} a_i x^i+\int_0^a a(t)(x-t)_+ dt \qquad [3\text{-}81]$$

sendo que, a_i é um valor desconhecido e $a(t)$ uma função desconhecida.

Pode-se construir o núcleo da função que gera superfícies flexíveis de ordem d.

$$k(x,z)=\sum_{r=0}^{d} \frac{\binom{d}{r}}{2d-r+1}(x \wedge z)^{2d-r+1}|x-z|^r+\sum_{r=0}^{d} x^r z^r \qquad [3\text{-}82]$$

sendo, **min(x,z)=(x ∧ z).**

Para obter o núcleo para superfícies n-dimensionais com número infinito de nós pode-se aplicar a equação [3-80].

3.5.3 - Núcleos que geram expansões da série de Fourier.

A série de Fourier representa uma ferramenta importante para a análise de sinais.

Analisando um sinal unidimensional em termos da expansão em série de Fourier mapeando a variável de entrada x em um vetor u de dimensão (2N+1), tem-se(Vapnik, 2000):

$$x \to u = \left({}^{1}\!/\!_{\sqrt{2}}, \text{sen}\, x, ..., \text{sen}\, Nx\ , \cos x\ , ..., \cos Nx \right) \qquad [3\text{-}83]$$

Pode-se escrever para cada valor fixo x o produto escalar:

$$f(x) = \langle a \cdot u \rangle = \frac{a}{\sqrt{2}} + \sum_{k=1}^{N} \left(a_k \,\text{sen}\, kx\ + b_k \cos kx\ \right) \qquad [3\text{-}84]$$

a função núcleo será dada por:

$$k_N(x,z) = \frac{1}{2} + \sum_{k=1}^{N} \left(\text{sen}\, kx\ \text{sen}\, kz\ + \cos kx\ \cos kz\ \right) \qquad [3\text{-}85]$$

que transformada considerando a função de Dirichlet:

$$k_N(x,z) = \frac{\text{sen}\left[\left(\frac{2N+1}{2} \right) (x - z) \right]}{\text{sen}\left(\frac{x - z}{2} \right)} \qquad [3\text{-}86]$$

Para a construção de núcleos da máquina de vetor de suporte para um espaço d-dimensional com espaço de entrada da forma $x=(x^1,x^2,....,x^n)$, aplica-se a equação [3-80].

Introduzindo termos para regularização, considerando que a expansão em Fourier, não processa boa aproximação, pode-se obter outras formas de núcleos que geram expansões em séries de Fourier, conforme a equação [3-87] com modo de regularização robusto mostrado na Fig 3.24.

$$k(x,z) = \frac{1}{2} + \sum_{k=1}^{\infty} q^k \cos k(x - z) = \frac{1 - q^2}{2\left(1 - 2q\cos(x - z) + q^2\right)} \qquad [3\text{-}87]$$

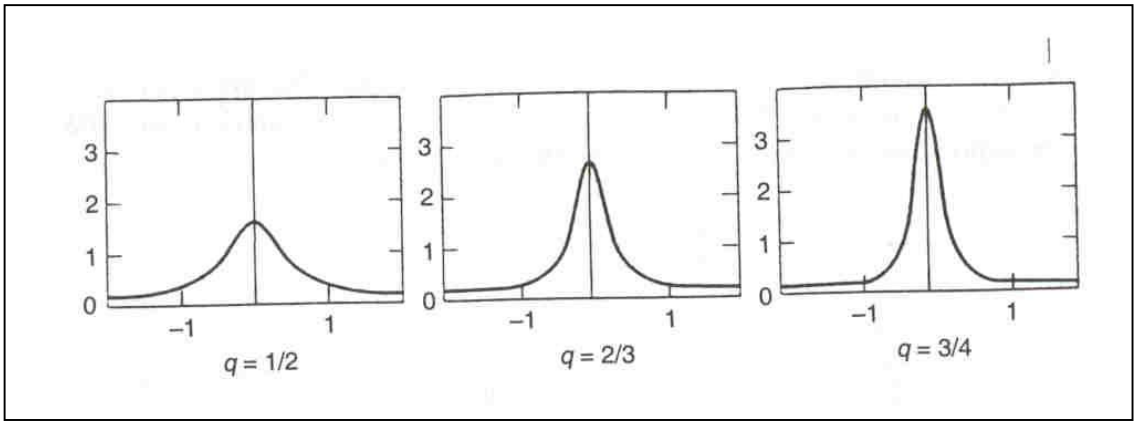


Figura 3.24 – Núcleos com modo de regularização robusto para vários valores de q(Vapnik, 2000)

A equação [3-88] introduz uma regularização fraca na expansão de Fourier conforme mostrado na Fig 3.25.

$$k(x,z) = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{\cos kx \cos kz + \operatorname{sen} kx \operatorname{sen} kz}{1+\gamma^2 k^2}$$

$$= \frac{\pi}{2\gamma} \frac{\cosh \dfrac{\pi - |x-z|}{\gamma}}{\operatorname{senh} \dfrac{\pi}{\gamma}} \qquad 0 \le |x-z| \le 2\pi \qquad \text{[3-88]}$$

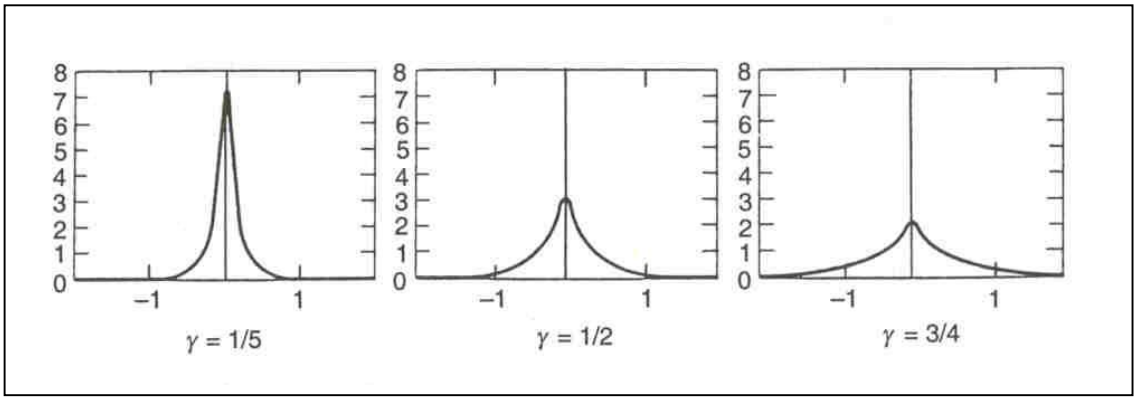


Figura 3.25 – Núcleos com modo de regularização fraco para vários valores de γ (Vapnik,2000).

3.5.4 – Funções núcleo ANOVA*(analysis of variances)*

O núcleo polinomial [3-66], e o núcleo formado por todos os subconjuntos polinomiais [3-69], apresenta controle limitado sobre qual característica utilizar e como atribuir pesos. O núcleo polinomial utiliza somente monômios de grau *d* com um parâmetro de controle *R*. Da mesma forma, o núcleo formado por subconjuntos polinomiais utiliza monômios que se restringem a possíveis subconjuntos das *n* características do espaço de entrada(Taylor, Cristianini, 2004).
O núcleo ANOVA de grau *d* se aproxima da formulação do núcleo formado por subconjuntos polinomiais [3-69], exceto de que está restrito aos subconjuntos de cardinalidade *d*.

Seja a função *n*-dimensional da forma $f(x)=f(x^1,....,x^n)$ definida em um espaço com intervalo finito ou infinito. A decomposição ANOVA da função *f(x)* é uma expansão da forma:

$$f(x^1,...,x^n) = F_0 + F_1(x^1,...,x^n) + F_2(x^1,...,x^n) + ... + F_n(x^1,...,x^n) \quad [3\text{-}89]$$

sendo

$$\begin{aligned}
F_0 &= C \\
F_1(x^1,...,x^n) &= \sum_{1\le k\le n} \phi_k\left(x^k\right) \\
F_2(x^1,...,x^n) &= \sum_{1\le k_1<k_2\le n} \phi_{k_1,k_2}\left(x^{k_1},x^{k_2}\right) \\
F_r(x^1,...,x^n) &= \sum_{1\le k_1<k_2<...k_r\le n} \phi_{k_1,k_2,...,k_r}(x^{k_1},x^{k_2},...,x^{k_r}) \\
F_n(x^1,...,x^n) &= \phi_{k_1,...,k_n}(x^1,...,x^n)
\end{aligned} \quad [3\text{-}90]$$

A construção clássica da decomposição ANOVA apresenta o problema do crescimento exponencial dos termos da somatória, com o crescimento da ordem da aproximação.

Para o caso de vetores suporte, que limita o número de pontos necessários para compor a função alvo, conforme mostrado na equação [3-42], esse problema não existe. Para construir o núcleo da decomposição ANOVA de ordem *d*, se efetua a soma dos produtos dos núcleos uni-dimensionais $k(x_i,z_i),\ i=1,2,..,n$

$$\begin{aligned} k_d(x,z) &= \sum_{1 \le i_1 < i_2 < \cdots < i_d \le n} k(x_{i_1}, z_{i_1}) k(x_{i_2}, z_{i_2}) \ldots . k(x_{i_d}, z_{i_d}) \\ &= \sum_{1 \le i_1 < i_2 < \cdots < i_d \le n} \prod_{j=1}^{d} k(x_{i_j}, z_{i_j}) \end{aligned} \qquad [3\text{-}91]$$

Pode-se introduzir um procedimento recursivo para computar [3-91]. Definindo

$$k^s(x,z) = \sum_{i=1}^{n} k(x^i, z^i) \qquad [3\text{-}92]$$

ter-se-á para o caso geral

$$\begin{aligned} k_0(x,z) &= 1 \\ k_d(x,z) &= \frac{1}{d} \sum_{s=1}^{d} (-1)^{s+1} k_{d-s}(x,z) k^s(x,z) \end{aligned} \qquad [3\text{-}93]$$

Utilizando a equação [3-93] em uma máquina de vetor de suporte com função de perda L_2, pode-se obter aproximações de qualquer ordem(Vapnik, 2000).

Na construção de diferentes máquinas de vetor de suporte encontrar-se-ão diferentes núcleos *k(x,z)* que satisfazem as condições do teorema de Mercer. A tabela da Fig 3.26 apresenta algumas formas de funções núcleo, com algumas considerações a respeito das constantes e limites das equações.

Tipo de máquina de vetor de suporte	Função núcleo $K(x,z)$	Comentários
Produto interno simples	$\langle x \cdot z \rangle$	
Máquina de aprendizagem polinomial	$(x \cdot z + R)^p$	A potência *p* é especificada *a priori* pelo usuário.
Polinômio real Vovk	$\frac{1-\langle x \cdot z \rangle^p}{1-\langle x \cdot z \rangle}$	A potência *p* é definida pelo usuário. **-1< ⟨x · z⟩< 1**
Polinômio real infinito Vovk	$\frac{1}{1-\langle x \cdot z \rangle}$	**-1< ⟨x · z⟩< 1**
Rede de função de base radial	$\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$	A largura σ^2 comum a todos os núcleos, é especificada *a priori* pelo usuário.
Multiquadrática inversa	$\frac{1}{\sqrt{\|x-z\|^2+c^2}}$	A constante *c* é definida pelo usuário.
Multiquadrática	$\sqrt{\|x-z\|^2+c^2}$	A constante *c* é definida pelo usuário
Splines de folha fina	$\|x-z\|^{2n+1}$ $\|x-z\|^{2n}\ln(\|x-z\|)$	Função núcleo da categoria de funções de base radial.
B-splines	$B_{2n+1}(x-z)$	B_n são trechos polinomiais de grau *n*.
Polinomial trigonométrica de grau *p*	$\frac{\text{sen}\left[\left(\frac{p+1}{2}\right)(x-z)\right]}{\text{sen}\left(\frac{x-z}{2}\right)}$	A ordem *p* da função trigonométrica é definida pelo usuário.
Gaussiano periódico	$\sum_{n=1}^{\infty} e^{-\frac{n^2\sigma^2}{2}}\cos(n(x-z))$	
Perceptron de duas camadas	$\tanh(\beta_0 x \cdot z + \beta_1)$	O teorema de Mercer é satisfeito apenas para alguns valores de β_0 e β_1.

Figura 3.26 – Resumo de algumas funções núcleo de produto interno(Haykin, 2001) (Evgeniou, Pontil, Poggio,2000)(Smola, 1998)(Vapnik, 2000)(Taylor, Cristianini, 2004)(Runarsson, Sigurdsson, 2003).

3.6 – Algoritmos para regressão com máquina de vetor de suporte.

Neste item serão apresentados os principais algoritmos capazes de determinar uma solução para os problemas de regressão de valores reais, obtidos a partir de medições efetuadas do fenômeno em estudo e apresentados na forma de pares de pontos de dado $(x_i,y_i),i=1,2,...,\ell$, com $x_i \in R^n$, e talvez, no espaço de características e $y_i \in R$.. Quando x_i está localizado no espaço de características, não há possibilidade de sabê-lo explicitamente, mas tão somente os valores do núcleo $k(x,z)$, correspondente.
A solução de um problema de regressão segue os passos apresentados na arquitetura de máquina de vetor de suporte, mostrada na Fig 3.27.

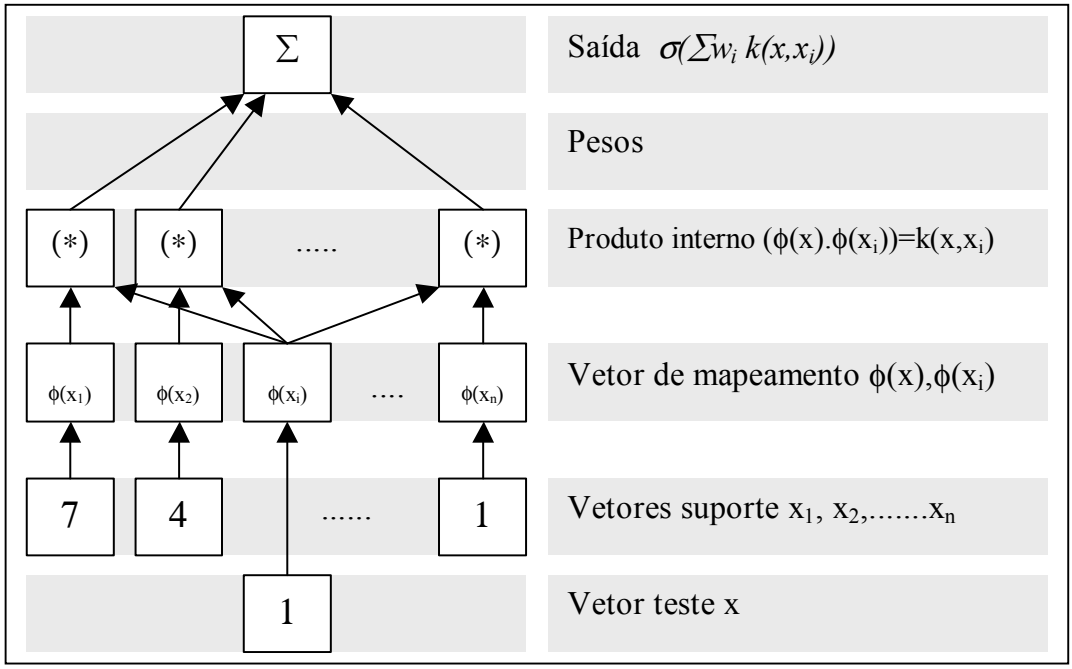


Figura 3.27 – Arquitetura de máquina de vetor de suporte(Smola, Bartlett, Schölkopf, Schuurmans, 2000).

A entrada x e o Vetor de Suporte x_i são mapeados não-linearmente (pela função ϕ) no espaço de características ϑ, onde são computados os produtos internos. Através do uso de funções núcleo k, as operações dessas duas camadas podem ser efetuadas em um só passo. Os resultados são linearmente combinados através dos pesos v_i, os quais são obtidos resolvendo um programa quadrático (para reconhecimento de padrões, $v_i=y_i\alpha_i$; para regressão $v_i=(\alpha_i-\alpha_i^*)$). A combinação linear é colocada dentro da função σ (para reconhecimento de padrões $\sigma(x)=sinal(\Sigma v_i k(x,x_i))$; para regressão $\sigma(x)= (\Sigma v_i k(x,x_i))$.

Iniciar-se-á a apresentação dos algoritmos com os métodos de aproximação através dos mínimos quadrados *(least squares regression)* e regressão de aresta *(ridge regression)*.

3.6.1 – Mínimos quadrados.

A apresentação clássica do método dos mínimos quadrados é da forma(Runarsson, Sigurdsson, 2003a)

$$\text{minimize} \quad \sum_{i=1}^{\ell} \xi_i^2$$

$$\text{sujeito a} \quad y_i - \langle w \cdot x_i \rangle - b = \xi_i, i = 1,2,...,\ell \qquad [3\text{-}94]$$

sendo que ξ é a distância entre o ponto medido e o calculado pela função de aproximação, isto é, é o erro da função linear para aquele exemplo particular de treinamento.

Fazendo $\mathbf{w}=[w\ \ b]^t$ e $\mathbf{X}=[x_i\ 1]$ pode-se colocar o vetor de discrepâncias na forma

$$\xi = y - Xw \qquad [3\text{-}95]$$

A função de perda a ser minimizada, pode ser escrita

$$L(\mathbf{w}, S) = \|\xi\|_2^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^t (\mathbf{y} - \mathbf{X}\mathbf{w}) \qquad [3\text{-}96]$$

Tomando a derivada parcial em relação à variável *w*, obtém-se a equação normal

$$X^t X w = X^t y \qquad [3\text{-}97]$$

cuja solução pode ser obtida pela inversão de $X^t X$

$$w = (X^t X)^{-1} X^t y \qquad [3\text{-}98]$$

3.6.2 – Regressão de aresta*(Ridge regression).*

O processo de otimização por regressão de aresta pode ser obtido resolvendo(Taylor, Cristianini, 2004)

$$\begin{aligned} &\min_w \quad \sum_{i=1}^{\ell} \xi_i^2 \\ &sujeito \quad a \quad y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle = \xi_i, i = 1,..., \ell \quad e \quad \|\mathbf{w}\| \le B \end{aligned} \qquad [3\text{-}99]$$

Aplicando a técnica dos multiplicadores de Lagrange obtém-se o lagrangeano

$$L(\mathbf{w}, \xi, \beta, \lambda) = \sum_{i=1}^{\ell} \xi_i^2 + \sum_{i=1}^{\ell} \beta_i [y_i - \langle \phi(\mathbf{x}_i), \mathbf{w} \rangle - \xi_i] + \lambda \left(\|\mathbf{w}\|^2 - B^2 \right) \qquad [3\text{-}100]$$

Derivando em relação às variáveis primordiais e efetuando as substituições em [3-99] e considerando $\alpha_i=\beta_i / (2\lambda)$ tem-se o lagrangeano na forma dual

$$\min_{\alpha} \quad L(\alpha, \lambda) = -\lambda \sum_{i=1}^{\ell} \alpha_i^2 + 2 \sum_{i=1}^{\ell} \alpha_i y_i - \sum_{i,j=1}^{\ell} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \qquad [3\text{-}101]$$

Diferenciando em relação aos parâmetros e igualando a zero, ter-se-á o equacionamento necessário para a implementação do algoritmo, dado por

$$\begin{aligned} &\boldsymbol{\alpha}* = (\mathbf{K} + \lambda I_{\ell})^{-1} \mathbf{y} \\ &f(x) = \sum_{j=1}^{\ell} \alpha_j^* k(x_j, x) \\ &\mathbf{w} = \sum_{j=1}^{\ell} \alpha_j^* \phi(x_j) \end{aligned} \qquad [3\text{-}102]$$

A otimização do problema de regressão ocorre a partir do ajuste do parâmetro λ, que corresponde a diferentes escolhas do valor *B*, o que equivale "variar λ corresponde a variar *B*". O processo de regressão de aresta apresenta a desvantagem de que a solução do vetor $\alpha*$ não é esparsa, assim para avaliar a função de aprendizagem para um novo exemplo, deve-se avaliar o núcleo para cada exemplo de treinamento.

3.6.3 – Regressão com função de perda insensível a ε.

Com a finalidade de introduzir esparsidade na solução da regressão, definem-se funções de perda, que envolvam desigualdades, permitindo que erros menores que um valor ϵ sejam ignorados, reduzindo os pontos a serem considerados, na avaliação da solução do problema de aproximação da função de regressão.

A Fig 3.28 ilustra um exemplo de função de regressão unidimensional com uma banda insensível ao ajuste, com valor atribuído ε. Os valores ξ medem o custo do erro dos exemplos de treinamento.

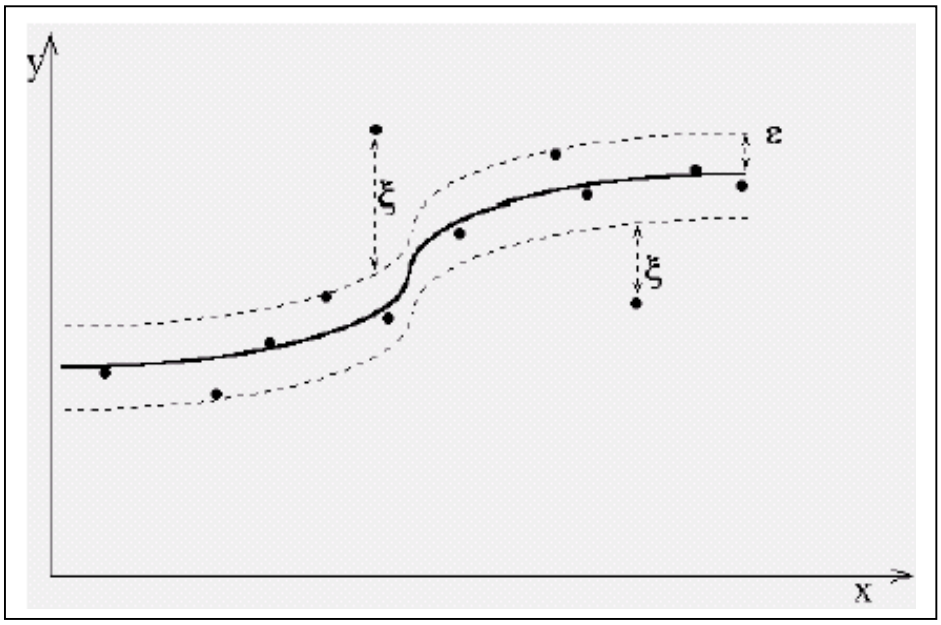


Figura 3.28 – Regressão utilizando a função de perda insensível a ε(Taylor, Cristianini, 2004).

Perda insensível a ε, quadrática: O vetor peso w e o bias *b*, para a solução do problema de regressão com função de perda insensível a ε, quadrática, da máquina de vetor de suporte, podem ser encontrados otimizando o problema

$$\begin{aligned} &\min_{w,b,\xi,\xi*} && \|w\|^2 + C\sum_{i=1}^{\ell}\left(\xi_i^2 + \xi_i^{*^2}\right) \\ &\textit{sujeito} \quad a && \begin{cases} \left(\langle w, \phi(x_i)\rangle + b\right) - y_i \le \varepsilon + \xi_i, i = 1,...,\ell \\ y_i - \left(\langle w, \phi(x_i)\rangle + b\right) \le \varepsilon + \xi_i^*, i = 1,...,\ell \end{cases} \end{aligned} \qquad [3\text{-}103]$$

A formulação dual pode ser obtida com o método padronizado considerando $\xi_i\xi_i^*=0$, obtendo-se a relação com multiplicadores de Lagrange

$$\max_{\alpha,\alpha*} \sum_{i=1}^{\ell} y_i\left(\alpha_i^* - \alpha_i\right) - \varepsilon\sum_{i=1}^{\ell}\left(\alpha_i^* + \alpha_i\right) - \frac{1}{2}\sum_{i,j=1}^{\ell}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_i^* - \alpha_j\right)\left(k(x_i, x_j) + \frac{1}{C}\delta_{ij}\right)$$

$$\textit{sujeito} \quad a \quad \sum_{i=1}^{\ell}\left(\alpha_i^* - \alpha_i\right) = 0 \qquad \alpha_i^* \ge 0, \alpha_i \ge 0, i = 1,...,\ell$$

$$\textit{condições} \quad KKT \quad \textit{complementares}$$

$$\begin{cases} \alpha_i\left(\langle w, \phi(x_i)\rangle + b - y_i - \varepsilon - \xi_i\right) = 0, i = 1,...,\ell \\ \alpha_i^*\left(y_i - \langle w, \phi(x_i)\rangle - b - \varepsilon - \xi_i^*\right) = 0, i = 1,...,\ell \\ \xi_i\xi_i^* = 0, \alpha_i\alpha_i^* = 0, i = 1,...,\ell \end{cases} \qquad [3\text{-}104]$$

Perda insensível a ε, **linear:** O vetor de pesos w e o bias *b*, podem ser encontrados com a formulação já apresentada no capítulo III em [3-23], e aqui repetida

$$\begin{aligned} &\min_{w,b,\xi,\xi*} && \|w\|^2 + C\sum_{i=1}^{\ell}\left(\xi_i + \xi_i^*\right) \\ &\textit{sujeito} \quad a && \begin{cases} \left(\langle w, \phi(x_i)\rangle + b\right) - y_i \le \varepsilon + \xi_i, i = 1,...,\ell \\ y_i - \left(\langle w, \phi(x_i)\rangle + b\right) \le \varepsilon + \xi_i^*, i = 1,...,\ell \\ \xi_i, \xi_i^* \ge 0, i = 1,...,\ell \end{cases} \end{aligned} \qquad [3\text{-}105]$$

A formulação dual pode ser obtida com o método padronizado, obtendo-se a relação com multiplicadores de Lagrange

$$\max_{\alpha,\alpha*} \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \varepsilon \sum_{i=1}^{\ell}(\alpha_i^* + \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i, x_j)$$

$$\textit{sujeito} \quad a \quad \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i) = 0 \qquad 0 \le \alpha_i^*, \alpha_i \le C, \quad i = 1,..., \ell$$

condições KKT complementares

$$\begin{cases} \alpha_i(\langle w, \phi(x_i)\rangle + b - y_i - \varepsilon - \xi_i) = 0, i = 1,..., \ell \\ \alpha_i^*(y_i - \langle w, \phi(x_i)\rangle - b - \varepsilon - \xi_i^*) = 0, i = 1,..., \ell \\ \xi_i\xi_i^* = 0, \alpha_i\alpha_i^* = 0, i = 1,..., \ell \\ (\alpha_i - C)\xi_i = 0, (\alpha_i^* - C)\xi_i^* = 0 \end{cases} \qquad [3\text{-}106]$$

Considerando a banda ±ε em torno da função de saída do algoritmo de aprendizagem, os pontos que não são estritamente internos ao tubo, são vetores suporte. Aqueles que não estão situados na superfície tubular ±ε, terão valor absoluto, do correspondente α_i, igual a C.

Regressão com vetores suporte, com parâmetro ν: O parâmetro utilizado no algoritmo de perda insensível a ε, não é sempre conhecido, a priori, com o necessário nível de precisão. A modificação introduzida, através do parâmetro ν, computa automaticamente o parâmetro ε (Runarsson, Sigurdsson, 2003a).

O vetor de pesos **w** e o bias *b* para a regressão com vetores suporte e o parâmetro ν, podem ser obtidos otimizando

$$\min_{w,b,\xi,\xi^*} \quad \|w\|^2 + C\left(\nu\varepsilon + \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\xi_i + \xi_i^*\right)\right)$$

$$sujeito\ \ a \quad \begin{cases} \left(\langle w, \phi(x_i)\rangle + b\right) - y_i \le \varepsilon + \xi_i, i = 1,...,\ell \\ y_i - \left(\langle w, \phi(x_i)\rangle + b\right) \le \varepsilon + \xi_i^*, i = 1,...,\ell \\ \xi_i, \xi_i^* \ge 0, i = 1,...,\ell \end{cases} \qquad [3\text{-}107]$$

Utilizando o procedimento usual, qual seja, encontrar o ponto de sela do lagrangiano, e introduzindo-o novamente no lagrangiano, obter-se-á a formulação dual do problema

$$\max_{\alpha,\alpha^*} \sum_{i=1}^{\ell} y_i\left(\alpha_i^* - \alpha_i\right) - \varepsilon\sum_{i=1}^{\ell}\left(\alpha_i^* + \alpha_i\right) - \frac{1}{2}\sum_{i,j=1}^{\ell}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)k\left(x_i, x_j\right)$$

$$sujeito\ \ a \begin{cases} \sum_{i=1}^{\ell}\left(\alpha_i^* - \alpha_i\right) = 0, \quad \sum_{i=1}^{\ell}\left(\alpha_i + \alpha_i^*\right) \le C\nu \\ -\frac{C}{\ell} \le \alpha_i^*, \alpha_i \le \frac{C}{\ell}, \quad i = 1,...,\ell \end{cases}$$

condições KKT complementares

$$\begin{cases} \alpha_i\left(\langle w, \phi(x_i)\rangle + b - y_i - \varepsilon - \xi_i\right) = 0, i = 1,...,\ell \\ \alpha_i^*\left(y_i - \langle w, \phi(x_i)\rangle - b - \varepsilon - \xi_i^*\right) = 0, i = 1,...,\ell \\ \xi_i\xi_i^* = 0, \alpha_i\alpha_i^* = 0, i = 1,...,\ell \\ \left(\alpha_i - \frac{C}{\ell}\right)\xi_i = 0, \left(\alpha_i^* - \frac{C}{\ell}\right)\xi_i^* = 0 \end{cases} \qquad [3\text{-}108]$$

O parâmetro ν controla a fração de erro, no sentido de que há mais $\nu\ell$ pontos de treinamento que caem fora do tubo, enquanto que, no mínimo $\nu\ell$ do total de pontos de treinamento, são vetores suporte, e assim, estão localizados ou na parte de fora do tubo ou na sua superfície(Taylor, Cristianini, 2004).

Como exposto, com a finalidade de estimar a função regressão de determinado conjunto de dados de treinamento, ter-se-á que especificar três parâmetros livres: o valor do raio do tubo ε (função de perda insensível a ε), o parâmetro de regularização C e o parâmetro de núcleo(a ordem do polinômio para núcleos polinomiais, o parâmetro de

distância para funções núcleo de base radial, a ordem da superfície flexível *(spline)* ou outras funções possíveis)(Vapnik,2000).

3.7 Síntese.

Neste capítulo foram apresentadas as teorias nas quais se baseia a aprendizagem com a utilização de vetores suporte. A teoria estatística da aprendizagem providencia o necessário entendimento do funcionamento das máquinas de aprendizagem, baseado na teoria da minimização estrutural do risco a partir do controle da complexidade da classe de funções através da dimensão Vapnik-Chervonenkis. A teoria da otimização permite estabelecer o equacionamento necessário para a solução do problema, baseado em funções objetivas e restrições, apresentando a importante conceituação de dualidade, que permite otimizar a solução de um problema convexo pela introdução de uma formulação dual com a utilização de multiplicadores de Lagrange. A introdução dos vetores suporte para regressão, a partir da formulação primordial do problema e da correspondente formulação dual, com a utilização de funções de perda para penalizar pontos situados fora da região factível, das quais depende significativamente a performance da máquina de vetor de suporte, permite solucionar problemas de natureza linear. A não-linearidade das aplicações complexas do mundo real requer hipóteses mais expressivas do que funções lineares. A introdução de uma apropriada função núcleo(kernel), permite implicitamente, a partir de um mapeamento não-linear, representar o problema em um espaço de características de alta dimensionalidade, onde é possível formular o problema com uma máquina linear.

A característica do mapeamento é tal que não é necessária a definição da função de mapeamento, pois uma função núcleo K definida em um espaço de produtos internos, permite mapear diretamente os dados exemplos de treinamento, para o espaço de alta dimensionalidade.

A função núcleo assim definida apresenta como propriedade principal de ser positiva semidefinida, em um espaço de Hilbert e satisfazer as condições do teorema de Mercer. Uma característica extremamente importante da construção de núcleos é a sua completa modularidade, que permite a sua utilização com qualquer algoritmo, com total independência dentro dos passos de execução da solução do problema. Outra característica essencial para a manipulação dos dados de entrada em um espaço de alta dimensionalidade, é a representação matricial da função núcleo, a partir da utilização da

formulação da matriz de Gram, permitindo juntamente com a conceituação de vetores suporte, reduzir de forma substancial o número de pontos necessários(ordem da matriz) para a solução do problema de regressão. A construção de funções núcleo deve seguir a propriedade de serem positivas semi-definidas, o que permite construir uma infinidade de funções, dentre as quais se destacam as funções polinomiais que apresentam uma formulação através da análise de variâncias, denominada ANOVA.

Os algoritmos desenvolvidos para determinar os pesos e o bias, necessários à formatação da função de saída, são definidos a partir da função de Lagrange construída a partir da função objetiva com as respectivas restrições. A principal restrição atribuída aos dados de entrada é de que aqueles pontos situados em uma região definida por uma superfície tubular de raio pré-fixado $\boldsymbol{\epsilon}$ ao redor da curva solução, não contribuem para a componente do erro global do processo, relaxação necessária para prover o problema de um fator de flexibilidade para permitir a solução através de vetores suporte, evitando o excesso de ajuste(*overfitting*). São os métodos conhecidos como de perda insensível a $\boldsymbol{\varepsilon}$**.**

Com base na formulação de vetores suporte, e, mapeamento através de funções núcleo em um espaço de alta dimensionalidade, far-se-á um estudo de regressão não-linear, definindo os parâmetros que melhor se adaptem à série em estudo: $\boldsymbol{\varepsilon}$ do método de perda insensível, **C** parâmetro de regularização e **k** função núcleo, comparando com uma rede neural do tipo *RBF(Radial Basis Function)*.

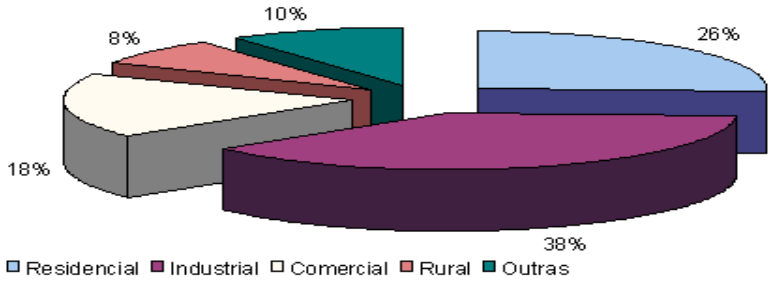
CAPÍTULO IV – METODOLOGIA.

A série temporal escolhida para efetuar o estudo de caso apresenta o perfil de demanda de potência elétrica(valor total da potência elétrica instantânea medida a cada 30 min) da região geográfica que abrange o Estado do Paraná.
Neste capítulo apresentar-se-á a metodologia utilizada para efetuar o estudo de caso mostrando as características da curva de demanda de potência elétrica.

4.1 Características de consumo da região abrangida pelo estudo.

A Fig. 4.1 apresenta a estrutura percentual da classificação de consumo de energia e o percentual de crescimento experimentado no primeiro trimestre de 2005 relativamente ao mesmo período do ano anterior.

Consumo de Energia- GWh	Acumulado Até		
Classes	mar/05	mar/04	%
Residencial	1.160	1.122	3,4
Industrial	1.727	1.727	0,0
Comercial	821	777	5,7
Rural	362	349	3,8
Outras	432	422	2,4
Total	4.502	4.397	2,4



Figua 4.1 – Percentual de consumo por categoria de consumidor da Copel(Fonte: Copel)

A carga total em uma área de concessão de energia elétrica é composta de diferentes unidades de consumo. Grande parte da eletricidade consumida é devida a atividades industriais, enquanto outra parte substancial é consumida por pessoas na forma de aquecimento, iluminação, preparação de alimentos, limpeza de roupas e outras atividades domésticas. Uma categoria igualmente importante é o consumo comercial, com características notadamente urbanas, apresentando um percentual significativo na

matriz de consumo de energia. O Estado do Paraná apresenta uma estrutura de produção agrícola bastante desenvolvida, resultando em uma categoria de consumidor classificada como rural, atendida por uma rede de sub-transmissão bastante extensa apresentando um consumo de energia significativo. Dentro da categoria classificada como "Outros", situam-se os consumos relacionados com os serviços públicos e também as perdas do sistema.

Os fatores que podem afetar o consumo são particulares para cada unidade de consumo. Assim o consumo industrial está relacionado com o nível de produção industrial e está intimamente relacionado com o momento econômico que o país ou região está atravessando. A dependência do consumo industrial com a produção pode ser aferida para os diferentes tipos de produtos, no entanto, as unidades industriais normalmente introduzem incertezas no prognóstico de carga advindas das paradas ocasionadas por quebra de máquinas ou greves, introduzindo grandes distúrbios não previstos no nível de demanda de potência.

Para o caso residencial, os fatores que determinam o consumo são de definição mais difícil, uma vez que cada pessoa decide por seus próprios meios e a decisão de consumo está relacionada com a psicologia do comportamento dos indivíduos. Muitos fatores sociais e ambientais podem influenciar o consumo residencial, tais como, grandes eventos, feriados, programas especiais de TV e outros.

Para o caso a ser estudado não será considerada a variação de temperatura ambiental como fator que influencia o consumo de energia e, portanto, não afetando a curva de demanda de potência. Tal condição foi também adotada em trabalho apresentado na EUNITE(EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems)(Chen, Chang, Lin, 2001) que concluiu que para um experimento para prognosticar valores de pico de demanda de potência em um horizonte de 30 dias, os resultados encontrados sem a utilização da modelagem com temperatura foram melhores, uma vez que para um período de curto prazo, em que a temperatura não apresenta variações significativas, prognosticá-la e incluí-la na modelagem não conduz a uma melhoria nos resultados.

Para o caso a ser estudado será considerado um período máximo de prognostico de uma semana, baseado em valores ocorridos no passado recente para promover a aprendizagem da máquina de vetor de suporte.

4.2 Características da curva de demanda.

Os dados de demanda de potência são apresentados na forma de valores reais sucessivos, sendo um valor para cada 30 minutos, representativo da média do período, totalizando 48 medidas diárias, constituindo uma série temporal discreta.

Os valores de demanda de potência a serem utilizados no trabalho traduzem a totalidade das medidas efetuadas em cada horário, efetuadas por sistemas automatizados de aquisição de dados e integralizam todas as categorias de consumo anteriormente descritas, geograficamente circunscritas no estado do Paraná.

Os valores utilizados no trabalho, situados dentro do período correspondente ao ano de 2000, foram normalizados, dividindo cada valor por um valor maior que o máximo atingido na série, sendo no caso apresentado na Fig 4.2, correspondente a 4000MW.

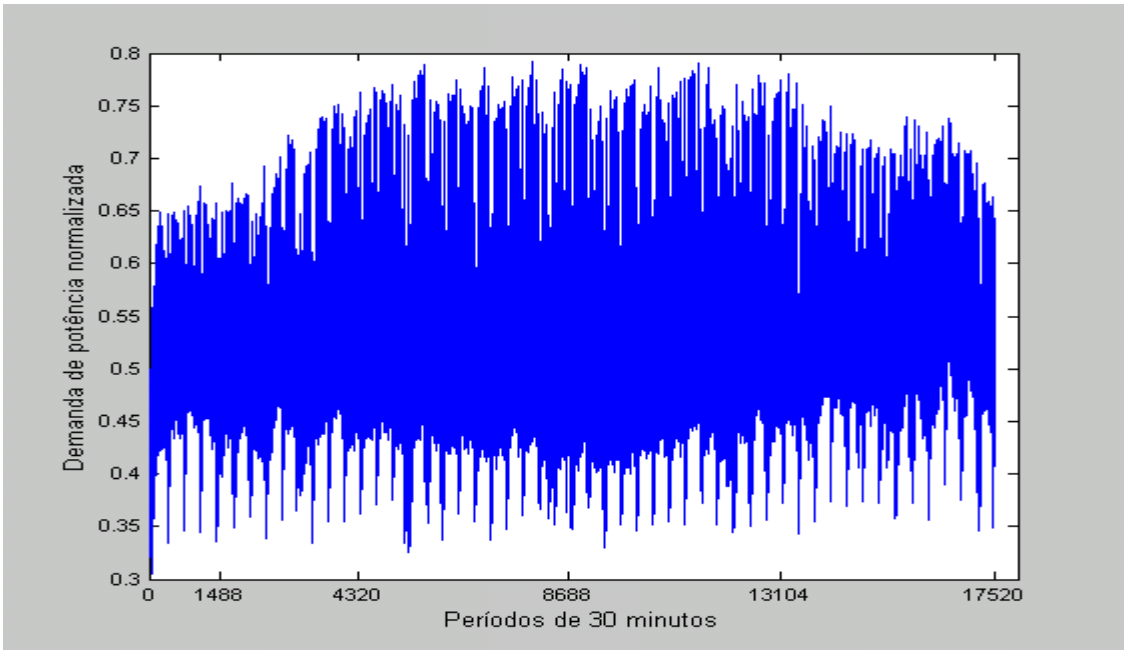


Figura 4.2 – Demanda de potência do período 1 de janeiro a 31 de Dezembro de 2000(Fonte: Copel).

4.3 Análise do comportamento do perfil de demanda.

A curva de demanda apresenta um pequeno crescimento sazonal no meio do ano que pode ser atribuído ao decréscimo de temperatura no inverno, no entanto, não representa uma variação substancial em relação aos meses com temperaturas mais elevadas.

Os picos inferiores correspondentes aos menores valores, apresentados na parte inferior da curva, correspondem a valores que ocorrem nos finais de semana e feriados totalizando aproximadamente 52 eventos(semanas anuais).

Os valores de pico superiores da curva, situados entre dois picos inferiores, correspondem aos valores experimentados no ritmo semanal do consumo de energia.

A variação cíclica semanal da demanda de potência pode ser, melhor observada através da função de autocorrelação, conforme mostrado na Fig 4.3. Os picos que ocorrem a cada 48 períodos, correspondentes a 24 h, se referem ao ritmo diário de utilização de energia, e os picos que ocorrem a cada 336 períodos, correspondentes aos 7 dias da semana, está a indicar que existe um ritmo semanal de utilização de energia.

Durante a semana as atividades sociais e de trabalho são mais intensas, razão pela qual o consumo de energia é mais elevado do que aquele dos finais de semana e feriados.

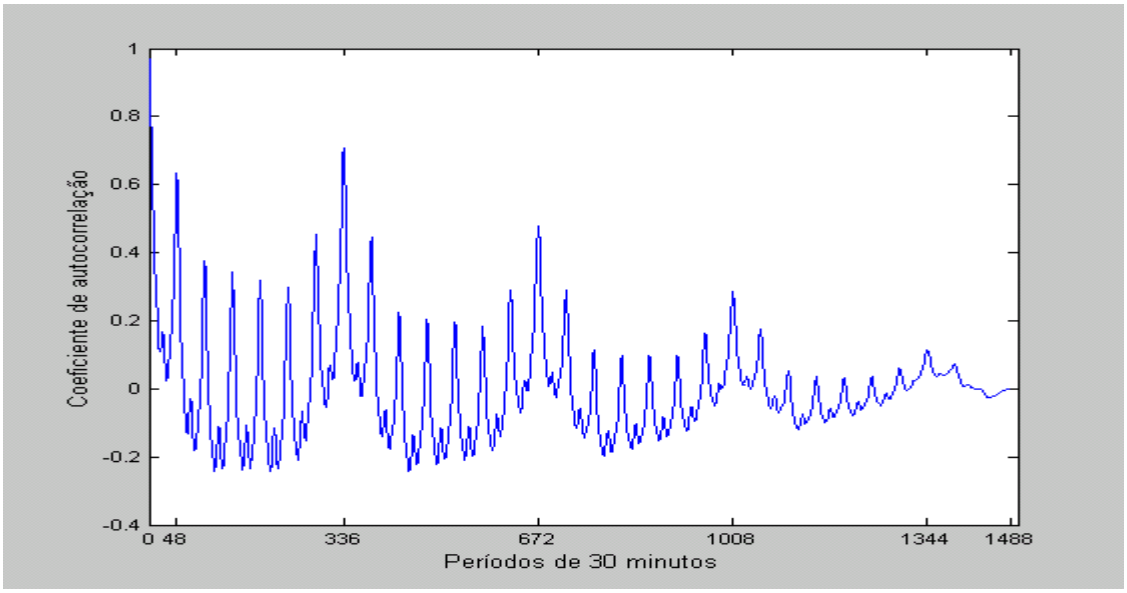


Figura 4.3 – Exemplo de autocorrelação dos dados do mês de Janeiro/2000

Na curva de demanda de potência apresentada na Fig 4.4, pode-se observar uma regularidade no padrão de comportamento para os dias de semana(períodos situados entre as abcissas 96 a 336 e 432 a 672), onde as cinco curvas correspondem às curvas de segunda à sexta-feira, em um período que não apresenta a ocorrência de feriados.

Da mesma forma, as curvas de demanda de potência dos finais de semana (sábado e domingo - entre abcissas 0 - 96, 336 – 432 e 672 - 768) apresentam uma regularidade no padrão de comportamento.

Pode-se observar, também, que o comportamento inicial da curva de Segunda-feira, nas primeiras horas da manhã(iniciando na abcissa 96 e 432) é diferente do início das curvas dos demais dias da semana, o que leva a considerar esse dia pertencente a uma outra família de curvas de demanda.

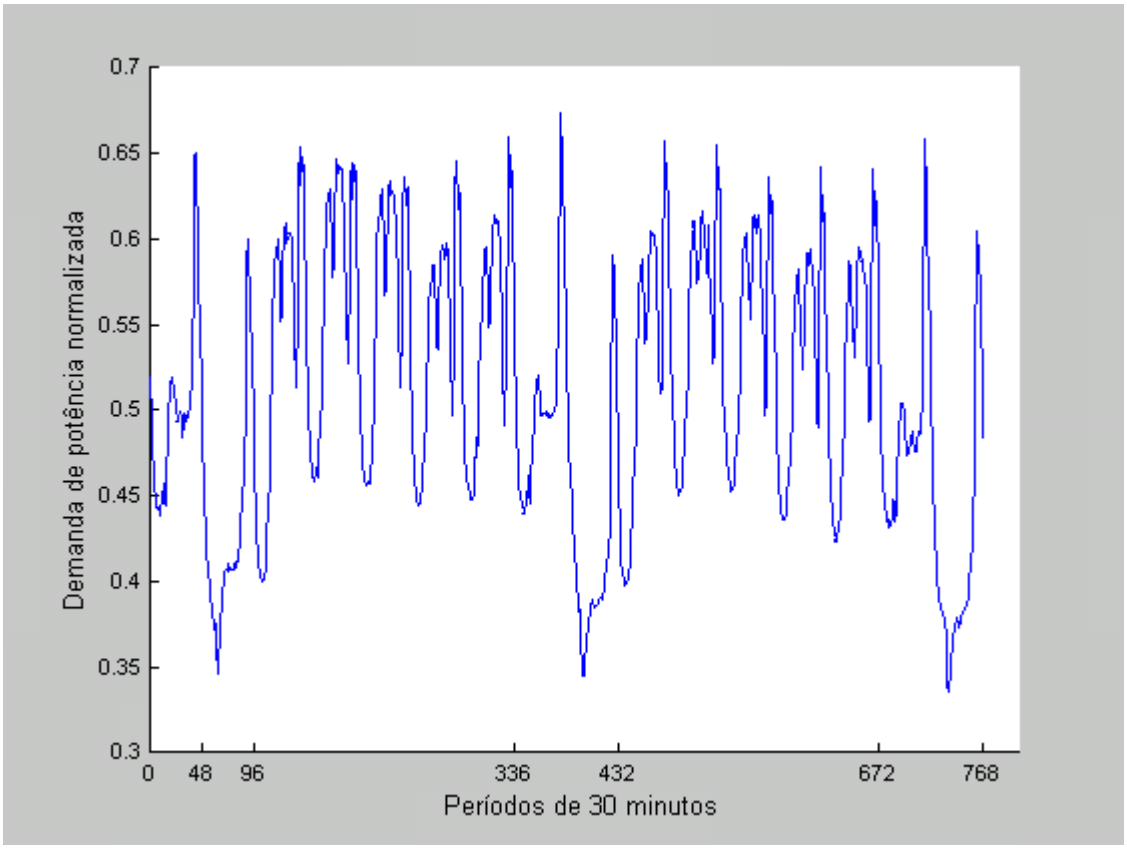


Figura 4.4 – Curva de demanda de potência no período de 15.01.2000 a 30.01.2000(Fonte:Copel).

O comportamento diário do consumo de energia, e, conseqüentemente da demanda de potência instantânea está relacionado com o desenvolvimento das atividades de rotina das pessoas quando em trabalho e tempo de lazer. A Fig 4.5 mostra as curvas de demanda de 4 meses do mesmo ano e para um mesmo dia da semana(quarta-feira).

Até às 16:00h todas as curvas apresentam aproximadamente o mesmo comportamento. A partir das 16:00h o comportamento dos meses de abril e julho se apresentam da mesma forma, sendo que os dos meses de janeiro e outubro diferem destes e também entre si.

O mês de janeiro normalmente tem as atividades comercial e industrial reduzidas principalmente por ser um mês de férias coletivas nos mais diversos setores da economia, justificando a redução de consumo.

Existe um defasamento de aproximadamente 1 hora entre os picos de demanda dos meses de abril e julho em relação ao do mês de outubro. Isto se deve ao comportamento da demanda, introduzido pela adoção do horário de verão(iniciado em 8 de outubro de 2000), que efetua o deslocamento de parte da demanda de pico, reduzindo o consumo de energia no período.

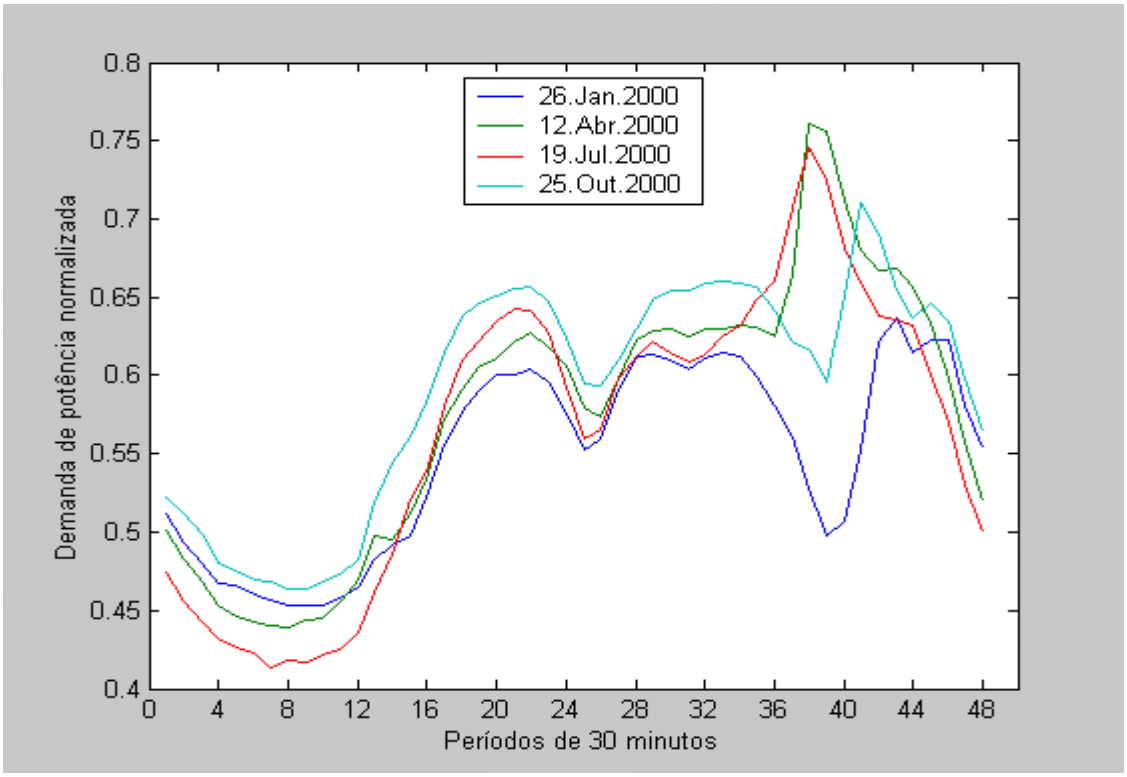


Figura 4.5 – Curvas de demanda de quatro quartas-feiras no ano de 2000.(Fonte:Copel).

O consumo de energia nos dias de feriados apresenta uma similaridade com o comportamento verificado nos finais de semana, no entanto, sofre variações com a posição do feriado dentro da semana. Assim um feriado que ocorre no meio da semana, por exemplo, numa quarta-feira, não introduz alterações significativas na curva de demanda de potência dos dias que antecedem e sucedem o evento.

A Fig 4.6 mostra o comportamento da demanda de potência no feriado de 15.11.2000 que ocorreu em uma quarta feira (período de 48 a 96 em verde), onde, o dia que antecede e o que sucede o feriado apresentam uma curva típica de dia de semana.
A curva em azul apresenta o comportamento do final de semana anterior ao feriado, mostrando a similaridade entre a curva do feriado(dia 15) e a do domingo(dia 12), indicando uma possível direção para prognosticar esse tipo de feriado.

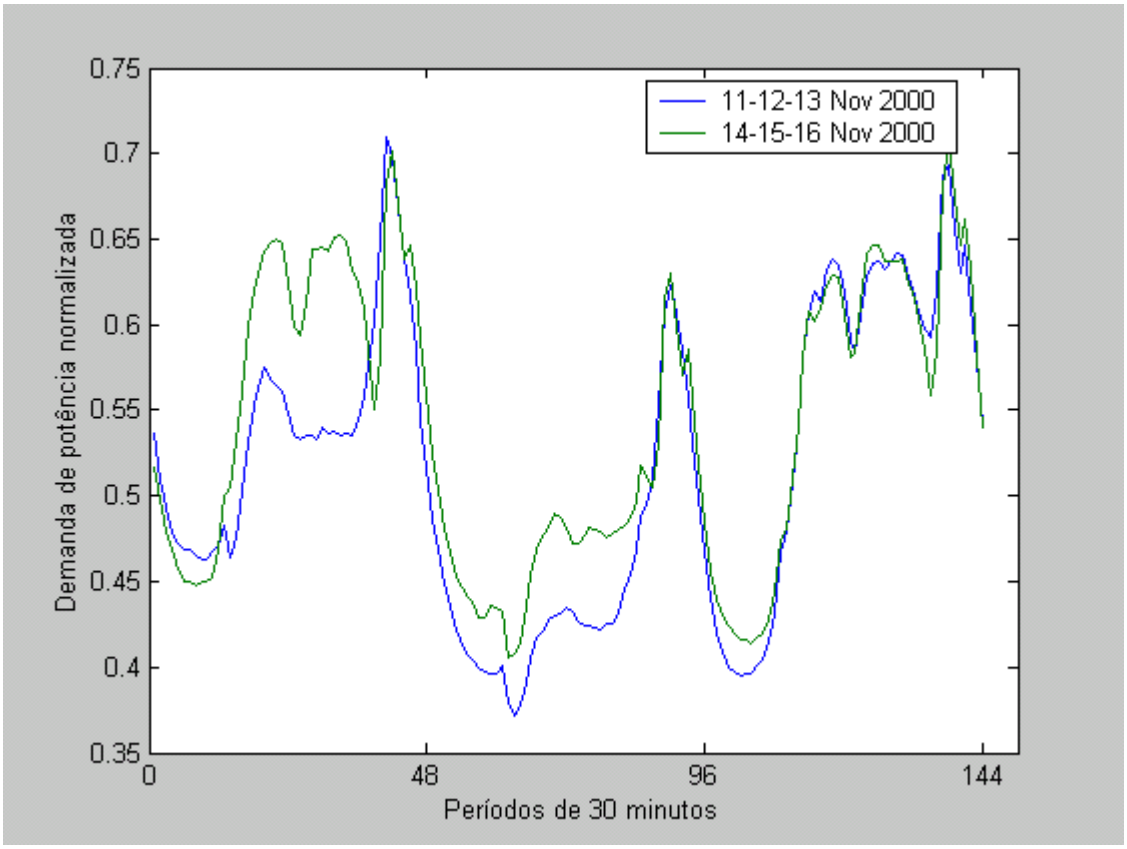


Figura 4.6 – Curva de demanda, em feriado, ocorrida no meio de semana(Quarta-feira) (Fonte:Copel)

A ocorrência de feriado no início ou final de semana altera substancialmente o comportamento do consumo de energia, pelo fato de que cada unidade de consumo, principalmente as comerciais, industriais e setor público, não apresentam uma uniformidade na decisão quanto ao expediente de trabalho, entre os dias normais situados entre o final de semana e o feriado.

A Fig 4.7 mostra o comportamento de um feriado situado em uma segunda-feira (dia 01 de maio), comparativamente com o final de semana subseqüente, sem feriado na segunda-feira. Observa-se que o consumo no feriado (período de 96 a 144 em azul) acompanha aproximadamente o comportamento do consumo do domingo anterior e posterior, o que permite estabelecer procedimentos para prognosticar a curva de demanda desse tipo de feriado.

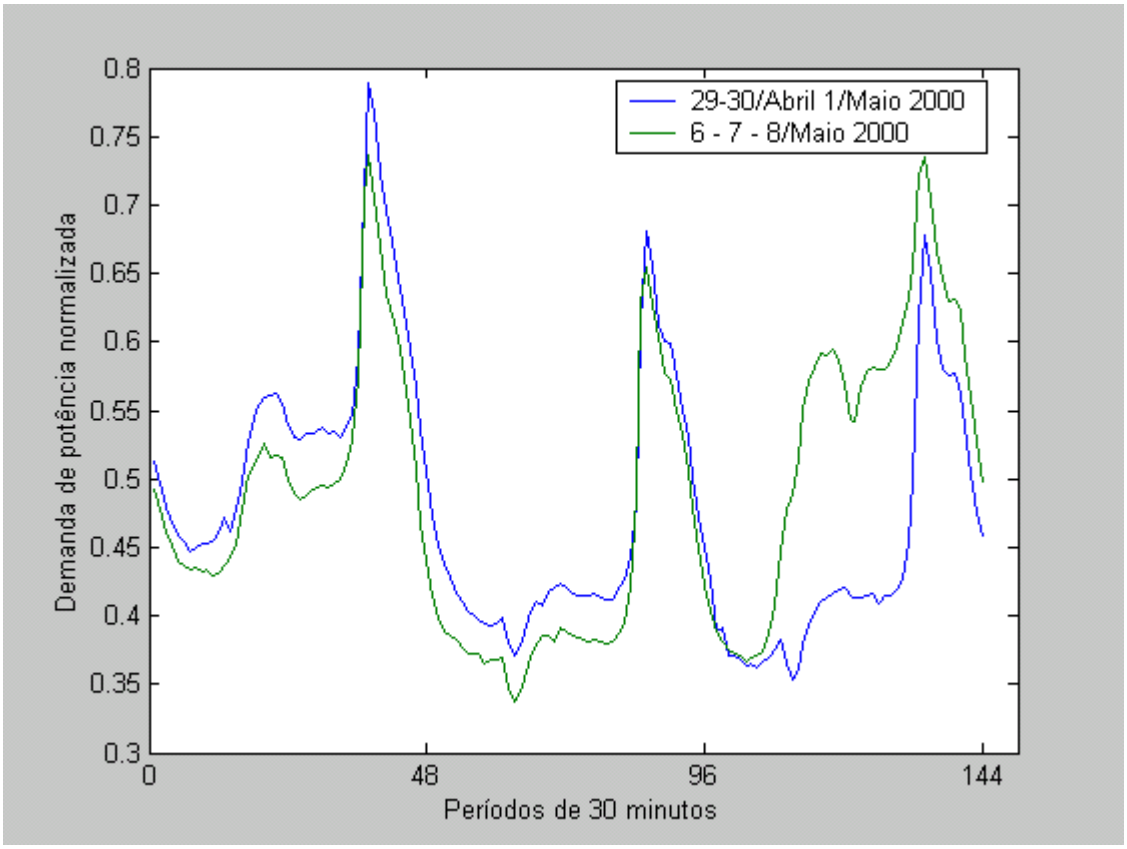


Figura 4.7 - Curva de demanda, em feriado, ocorrida no inicio de semana(Segunda-feira) (Fonte:Copel)

A Fig 4.8 apresenta um caso em que o feriado ocorre próximo do final de semana(7 de setembro, quinta feira – período de 48 a 96 em azul) em contraposição com idêntico período no final de semana subseqüente(em verde).

Observa-se que o dia útil(8 de setembro, sexta-feira – período de 96 a 144 em azul) apresenta um comportamento bastante próximo do sábado de semana normal(16 de setembro, sábado - período de 144 a 192 em verde) e no sábado propriamente dito(dia 9 de setembro, sábado – período de 144 a 192 em azul) a curva de demanda ao longo do dia apresenta uma redução no consumo, relativamente a um sábado dentro de uma semana normal.

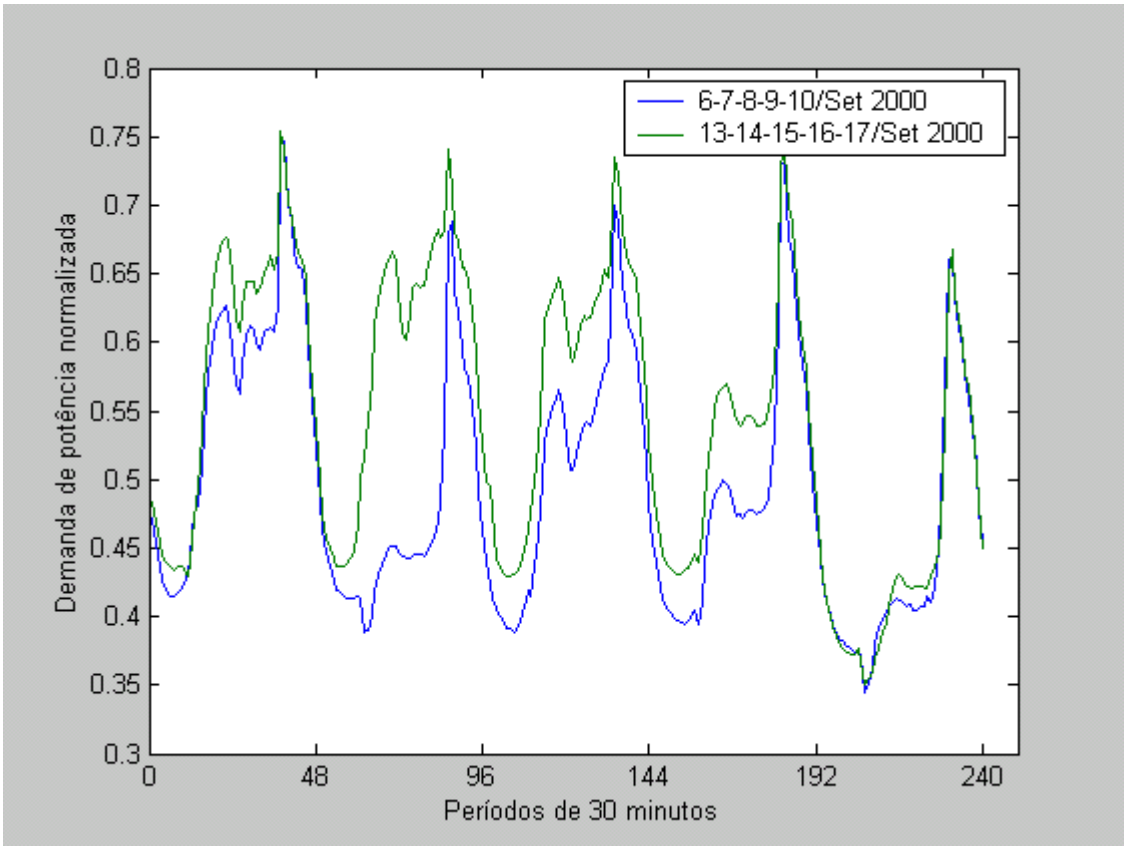


Figura 4.8 - Curva de demanda, em feriado, ocorrida próximo a final de semana (Quinta-feira) (Fonte:Copel)

Pelas observações expostas nas diversas curvas de demanda apresentadas pode-se classificar o comportamento do consumo de energia em várias classes, a saber: 1)Segundas, 2)Terças a Sextas, 3)Sábados, 4) Domingos e 5)Classes com feriados.
A escolha dos conjuntos de treinamento e de prognóstico para o estudo de caso dar-se-á com base no comportamento diferenciado da curvas de demanda de potência elétrica.

4.4 Recursos computacionais.

O algoritmo utilizado para efetuar o estudo de caso foi o de regressão com função de perda insensível a ε linear, que requer a definição prévia, pelo usuário, dos parâmetros C e ε, bem como da função núcleo k.
O equipamento utilizado foi um computador com processador Pentium 4, 3.2GHz, 512Mb e 120Gb. O estudo da série foi constituído da montagem do conjunto de treinamento, processamento do treinamento com programação quadrática e obtenção do prognóstico em ambiente MATLAB.

4.5 Estrutura do conjunto de treinamento.

Inicialmente foi estruturado um modelo constituído por um conjunto de treinamento, sempre com 48 componentes, representando os valores de demanda a cada 30min de um dia, e a saída, pelo valor imediatamente subseqüente a cada vetor de entrada, conforme mostrado no esquema da Fig 4.9.

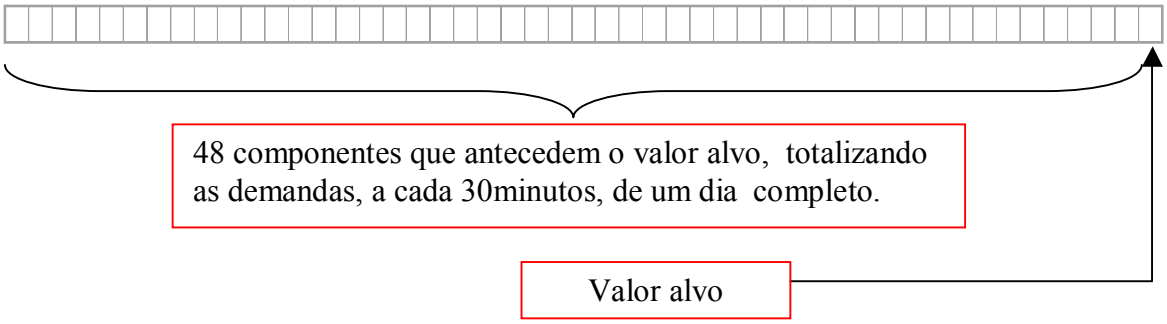


Figura 4.9 – Estrutura do vetor de treinamento com 48 componentes.

A razão da escolha do conjunto de treinamento da Fig 4.9 se deve ao fato de que, observando o perfil da curva de demanda de potência elétrica, de um dia para outro, não

apresentar variações excessivas, a não ser os valores de finais de semana, o conjunto de treinamento assim constituído poderia ser capaz de reproduzir o mesmo perfil, quando em presença de novos valores de entrada. Esta hipótese não foi confirmada, conforme apresentado no caso I, a seguir.

Com a finalidade de obter um melhor resultado de prognóstico foi estruturado uma nova modelagem do conjunto de treinamento. O vetor de entrada do modelo é constituído por valores de demanda de potência elétrica ocorridos em oito instantes do passado recente, configurando uma variável *L(i)*, onde *i* é o período de 30 minutos que corresponde ao valor discreto da demanda de potência integralizada, representando a média do período. A saída é constituída por um único valor correspondente ao valor da demanda no instante *i*, conforme esquema da Fig 4.10.

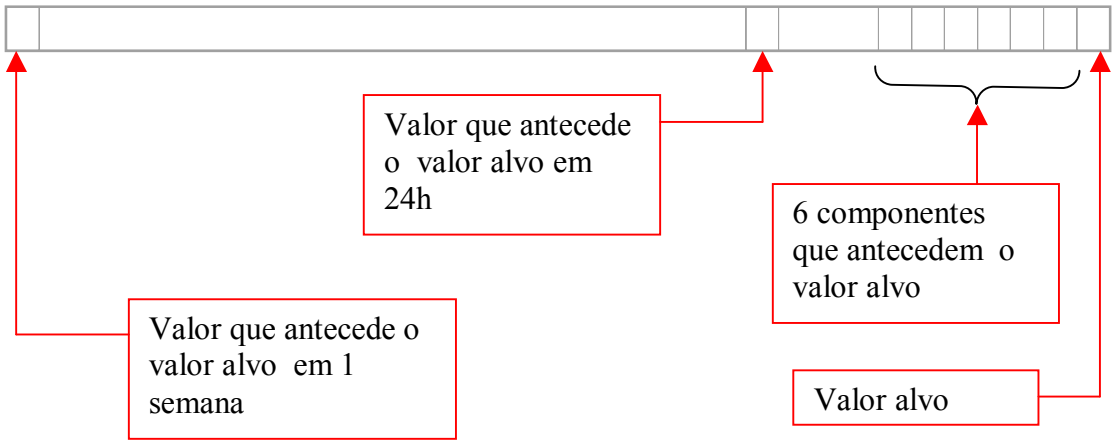


Figura 4.10 - Estrutura do vetor de treinamento com 8 componentes.

Assim, o vetor de entrada com 8 componentes, conduziu a resultados de prognóstico de valores futuros mais compatíveis com as necessidades dos sistemas elétricos, se mostrou adequado para o problema, apresentando a estrutura:

$$L_1 = L(i-336)$$

$$L_2 = L(i-48)$$

$$L_k = L(i-9+k), \text{ onde } k=3,4,5,6,7,8$$

sendo:

L_1 = Valor de demanda de potência, ocorrido no mesmo período de 30 minutos, que antecede 336 períodos(1 semana) do valor L_i*(valor de saída do treinamento)*.

L_2= Valor de demanda de potência, ocorrido no mesmo período de 30 minutos, que antecede 48 períodos(1 dia) do valor L_i*(valor de saída do treinamento)*.

L_k= Valores de demandas de potência, ocorridos nos 6 períodos de 30 minutos, que antecedem o valor L_i*(valor de saída do treinamento)*.

Assim, cada vetor de entrada é constituído por 8 componentes, e os vetores de entrada subseqüentes são deslocados de um período de 30 minutos, de forma a constituir um conjunto de treinamento, como mostrado na Fig 4.11, a seguir:

i	Vetores de entrada								Saída
1	$L(i_1\text{-}336)$	$L(i_1\text{-}48)$	$L(i_1\text{-}6)$	$L(i_1\text{-}5)$	$L(i_1\text{-}4)$	$L(i_1\text{-}3)$	$L(i_1\text{-}2)$	$L(i_1\text{-}1)$	$L(i_1)$
2	$L(i_1+1\text{-}336)$	$L(i_1+1\text{-}48)$	$L(i_1+1\text{-}6)$	$L(i_1+1\text{-}5)$	$L(i_1+1\text{-}4)$	$L(i_1+1\text{-}3)$	$L(i_1+1\text{-}2)$	$L(i_1+1\text{-}1)$	$L(i_1+1)$
....	....	....	....	....	....	....	....	....	....
ℓ	$L(i_1+\ell\text{-}337)$	$L(i_1+\ell\text{-}49)$	$L(i_1+\ell\text{-}7)$	$L(i_1+\ell\text{-}6)$	$L(i_1+\ell\text{-}5)$	$L(i_1+\ell\text{-}4)$	$L(i_1+\ell\text{-}3)$	$L(i_1+\ell\text{-}2)$	$L(i_1+\ell\text{-}1)$

Figura 4.11 – Estrutura do conjunto de treinamento da rede e saída correspondente.

O critério adotado para verificação do erro no resultado foi a determinação do erro médio absoluto (*Mean Absolute Error -MAE*) dos valores prognosticados, relativamente aos valores reais ocorridos no período correspondente ao vetor de saída da rede, conforme equação [4.1], onde y_R e y_P são os valores reais e prognosticados da saída, dos N valores do conjunto.

$$MAE\% = \frac{1}{N}\sum_{i=1}^{N}\left\|1-\frac{y_{P_i}}{y_{R_i}}\right\|100 \qquad [4.1]$$

4.6 Descrição dos estudos de caso.

Foram efetuados nove estudos de caso, sendo que as diferenças, entre eles, estão relacionadas ao conjunto de treinamento utilizado(48 componentes ou 8 componentes), ao tamanho do conjunto de treinamento(número de vetores de entrada), à classificação do dia de semana do conjunto de treinamento(meio ou fim de semana), ao tipo de função núcleo ou RBF e variações nos parâmetros C e ε da máquina de vetor de suporte.

A Fig 4.12 mostra de forma condensada a estrutura dos estudos de caso efetuados, sendo PI, PL e BR funções núcleo de produto interno, polinomial e da base radial, respectivamente, e, RBF rede de função de base radial.

O "Número de componentes do vetor" de entrada está relacionado com as Figs 4.9 e 4.10, o "Número de dias" corresponde à totalidade de dias que são assumidos pelo conjunto de treinamento, o "Número de vetores" corresponde ao produto Número de dias x 48(sendo 48 a quantidade diária de valores da série).

CASO	CONJUNTO DE TREINAMENTO				DIAS DA SEMANA NO CONJUNTO DE TREINAMENTO
	Número de com-ponentes do vetor	Número de dias	Número de vetores	Função núcleo	
I	48	7	336	PI	Todos os dias
II	8	7	336	PI	Todos os dias
III	8	14	672	PI	Todos os dias
IV	8	8	384	PI	Terça a sexta
V	8	12	576	PI	Terça a sexta
VI	8	16	768	PI,PL,BR	Terça a sexta
VII	8	16	768	RBF	Terça a sexta
VIII	8	15	720	PI,PL,BR	Sáb, Dom, Seg
IX	8	15	720	RBF	Sáb, Dom, Seg

Figura 4.12 - Estrutura dos conjuntos de treinamento dos estudos de caso.

4.6.1 Características do estudo de caso I.

A função núcleo de produto interno foi escolhida para iniciar o estudo de casos por se tratar da forma mais simples de computar a matriz, núcleo, dos vetores de entrada. Em todos os casos estudados, a curva de prognóstico foi obtida utilizando os valores calculados pela rede treinada, isto é, cada vetor de entrada foi constituído por valores medidos e pelos valores já prognosticados em fases anteriores. As características do estudo são:

- Conjunto de treinamento com 48 componentes em cada vetor, representando o perfil de demanda de 1 dia que antecede o valor alvo. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla uma semana completa de valores e foi aplicado à SVM com função núcleo de produto interno.
- Hipótese: Perfil semanal da demanda de potência, de uma semana para outra, apresenta pequenas variações nos valores, mantendo o formato geral e, portanto, a SVM treinada seria capaz de reproduzir o perfil de demanda.
- Resultado da escolha: A hipótese não se confirmou quando aplicado para prognóstico do mesmo conjunto de treinamento.

4.6.2 Características do estudo de caso II.

Este caso II difere do caso I na composição dos vetores de entrada, que no caso I é de 48 componentes e no caso II, 8 componentes.

A escolha do mesmo conjunto utilizado para treinamento, para obter o prognóstico, se deve pelo fato de que uma vez que houve convergência do processo de treinamento, o prognóstico deve reproduzir o perfil da curva de demanda. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla uma semana completa de valores e foi aplicado à SVM com função núcleo de produto interno.

- Hipótese: Os 6 valores que antecedem o valor alvo são suficientes para indicar a amplitude do ponto seguinte, e a presença dos valores a 1 semana e a 24horas do valor alvo em cada vetor providenciam o balizamento necessário para reproduzir o perfil da curva.
- Resultado da escolha: A hipótese se confirmou, apresentando uma curva de "prognóstico", do mesmo conjunto de treinamento, com o mesmo formato da curva real.

4.6.3 Características do estudo de caso III.

O Caso III acrescenta uma semana a mais ao conjunto de treinamento e faz o prognóstico de um período diferente daquele utilizado no treinamento. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla duas semanas completas de valores e foi aplicado à SVM com função núcleo de produto interno.
- Hipótese sobre as razões da escolha: O acréscimo de vetores no conjunto de treinamento permite representar melhor as variações do perfil de demanda e, portanto, melhorar o prognóstico.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real.

4.6.4 Características do estudo de caso IV.

O caso IV utiliza para o conjunto de treinamento os valores de meio de semana(terça a sexta-feira), mantendo as demais características do caso anterior. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de terça a sexta de 2 semanas e foi aplicado à SVM com função núcleo de produto interno.

- Hipótese: A retirada dos finais de semana (sábado, domingo e segunda), que tem perfil de demanda com formato diferente, deve melhorar o prognóstico de demanda.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real e redução no erro médio absoluto.

4.6.5 Características do estudo de caso V

O Caso V acrescenta uma semana de dados ao conjunto de treinamento procurando obter uma redução no erro de prognóstico. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de terça a sexta de 3 semanas e foi aplicado à SVM com função núcleo de produto interno.
- Hipótese: O acréscimo de vetores no conjunto de treinamento pode melhorar o resultado do prognóstico.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real e redução no erro médio absoluto.

4.6.6 Características do estudo de caso VI.

O caso VI acrescenta uma semana aos dados de treinamento e efetua o processamento com a utilização de outras funções núcleo, buscando uma redução no erro de prognóstico de valores futuros. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de terça a sexta de 4 semanas e foi aplicado à SVM com função núcleo de produto interno, polinomial e de base radial.
- Hipótese: O acréscimo de vetores no conjunto de treinamento, bem como a aplicação de outras funções núcleo, podem melhorar o resultado do prognóstico.

- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real e redução no erro médio absoluto para as três funções núcleo.

4.6.7 Características do estudo de caso VII.

O caso VII utiliza uma RBF para os dados do caso VI para permitir comparar resultados de desempenho. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de terça a sexta de 4 semanas e foi aplicado a uma RBF.
- Hipótese: Comparar os resultados com a SVM do caso VI.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real e erro médio absoluto maior que no caso VI.

4.6.8 Características do estudo de caso VIII.

O caso VIII utiliza para dados de entrada valores dos finais de semana(sábado, domingo e segunda-feira) processadas com a três funções núcleo. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de sábado, domingo e segunda-feira de 5 semanas e foi aplicado à SVM com função núcleo de produto interno, polinomial e de base radial.
- Hipótese: Verificar o comportamento da SVM quando aplicado a perfis não uniformes.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo

formato da curva real e aumento do erro médio absoluto para as três funções núcleo, quando comparados ao caso, com conjunto de treinamento, de valores de meio de semana.

4.6.9 Características do estudo de caso IX.

O caso IX foi processado com RBF para efetuar estudos comparativos de desempenho com SVM. As características do estudo são:

- Conjunto de treinamento com 8 componentes em cada vetor. Vetores de entrada subseqüentes defasados de 30minutos.
- Conjunto de treinamento contempla os perfis de demanda de sábado, domingo e segunda-feira de 5 semanas e foi aplicado a uma RBF.
- Hipótese: Comparar os resultados com a SVM do caso VIII.
- Resultado da escolha: A hipótese confirmou-se parcialmente, apresentando curva de prognóstico para um período diferente do período de treinamento, do mesmo formato da curva real e aumento do erro médio absoluto em relação à SVM.

4.7 Síntese.

Neste capítulo foi apresentada a metodologia empregada para obtenção dos resultados do estudo de caso, bem como as principais características da série de demanda de potência.

O estudo do conjunto de treinamento conduziu a adoção de 8(oito) componentes para cada vetor de entrada, sendo, um valor com uma semana de antecedência , um outro valor com 24h de antecedência e seis valores imediatamente anteriores ao valor alvo, representando o possível universo de valores que podem influenciar o valor futuro procurado.

Os estudos de caso refletem a gama de variações que podem ser introduzidas na formulação da máquina, sendo as principais o número de componentes do conjunto de treinamento, a composição do conjunto de treinamento, as funções núcleo e a aplicação da rede de função de base radial à mesma base de dados.

Os resultados quantitativos dos estudos de caso estão apresentados no capítulo que segue, com as considerações sobre o desempenho das configurações.

CAPÍTULO V – PROGNÓSTICO DE DEMANDA DE POTÊNCIA ELÉTRICA.

A obtenção de melhores resultados fez com que fosse necessária a análise de diversas situações de treinamento, uma vez que não há uma regra definida para o estabelecimento das variáveis que compõe o algoritmo de solução do problema de programação quadrática. Neste capítulo estão apresentados os resultados obtidos na aplicação da máquina de vetor de suporte para a determinação de prognóstico de demandas de potência elétrica, com base em valores ocorridos no passado.

5.1 – Resultados da aplicação do modelo.

Os resultados obtidos conforme mostrado de forma condensada na Fig 5.1 permitiram visualizar o comportamento dos diversos parâmetros envolvidos na definição da máquina de vetor de suporte, sendo PI, PL, BR e RBF, as funções núcleo de produto interno, polinomial de base radial e rede de função de base radial, respectivamente. A coluna "Dias de prognóstico", diz respeito ao número de dias futuros para os quais foi efetuado o prognóstico de demanda de potência, sendo que somente para os casos I e II foi utilizado o mesmo conjunto de treinamento para efetuar o prognóstico, nos demais casos o período, de prognóstico, é subseqüente ao período de treinamento.

CASO	CONJUNTO DE TREINAMENTO				Dias de prognóstico	MAE%
	Número total de vetores	Número total de dias	Dias da semana	Função Núcleo		
I	336	7	Todos os dias	PI	7	10,9324*
II	336	7	Todos os dias	PI	7	1,9492*
III	672	14	Todos os dias	PI	7	3,8160
IV	384	8	Terça a sexta	PI	4	2,8984
V	576	12	Terça a sexta	PI	4	2,8523
VI	768	16	Terça a sexta	PI	4	1,8364
				PL		2,0555
				BR		2,0188
VII	768	16	Terça a sexta	RBF	4	2,1945
VIII	720	15	Sáb, Dom, Seg	PI	3	2,8328
				PL		3,9487
				BR		2,5878
IX	720	15	Sáb, Dom, Seg	RBF	3	4,1293

*Utilizado o mesmo conjunto de treinamento, para prognóstico.

Figura 5.1 – Resultados dos estudos de caso.

1)Caso I

O estudo de caso I, em que o conjunto de treinamento é composto por vetores com 48 componentes que antecedem o valor alvo, foi efetuado com a finalidade de pesquisar uma estrutura adequada para a modelagem da máquina de vetor de suporte.

Foram utilizados valores de demanda de potência de uma semana completa(7 dias), separados em vetores de entrada com 48 componentes cada que abrangem 24h de dados integralizados a cada 30min. Vetores de entrada subseqüentes se diferenciam entre si pelo deslocamento horário de valores a cada 30min, constituindo um vetor de treinamento para cada valor de demanda do período do estudo, totalizando para o período do estudo um conjunto de 336 vetores de entrada. Para conjunto de prognóstico foram utilizados os mesmos dados do conjunto de treinamento, conforme mostrado na tabela da Fig 5.2.

Conjunto de treinamento	Vetores de entrada, constituídos por conjuntos de 48 componentes cada um, sendo que cada vetor corresponde aos 48 valores que antecedem o valor alvo, abrangendo o período de 14.01.2000 a 20.01.2000 (7 dias), totalizando 336 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor de cada entrada.
Conjunto de prognóstico	O mesmo que o do conjunto de treinamento.
Função núcleo	$K(x,z)=\langle x\cdot z\rangle$

Figura 5.2 – Estrutura do caso I.

A Fig 5.3 apresenta de forma gráfica o erro médio absoluto como função das variáveis C e ε, quando da aplicação, do conjunto de prognóstico, ao algoritmo de programação quadrática. O valor MAE aqui obtido resulta da aplicação da equação [4.1], efetuando o prognóstico de valores futuros, substituindo os valores reais do conjunto de prognóstico pelos resultados encontrados na simulação de valores futuros.

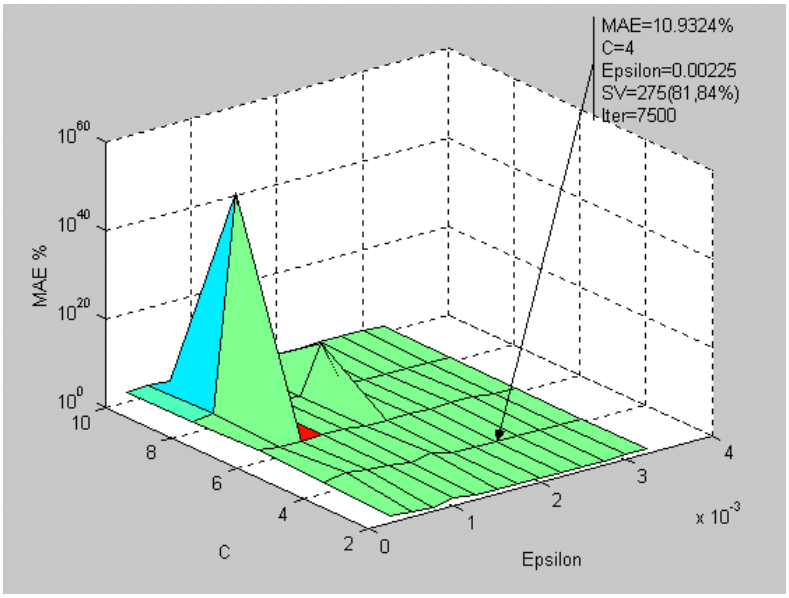


Figura 5.3 –Erro MAE=f(C,ε) do prognóstico, com utilização de valores calculados para obtenção de valores futuros, do caso I.

O valor médio do erro MAE para valores futuros sem a utilização de valores obtidos de prognóstico é da ordem de 1,35%. No entanto, efetuando o prognóstico com a utilização de valores calculados, para definição dos novos valores futuros, ter-se-á uma discrepância muito grande do erro MAE, que assumem valores da ordem de 10% até 2,4E+51, dificultando a obtenção de uma possível relação entre as variáveis do modelo com os vetores de entrada.

A Fig 5.4 mostra graficamente as curvas de demanda de potência, real e duas com valores de prognóstico. A curva em cor verde, que se aproxima bastante da curva com valores reais, foi obtida efetuando-se o prognóstico de valores futuros, utilizando o mesmo conjunto de vetores do conjunto de treinamento, ou seja, sem efetuar a substituição dos valores de demandas reais pelos valores de demanda calculados. A curva em vermelho foi obtida, utilizando nos vetores de entrada da simulação, os valores mais recentes das componentes, ou seja, efetuando a substituição dos valores das demandas reais pelos valores prognosticados.

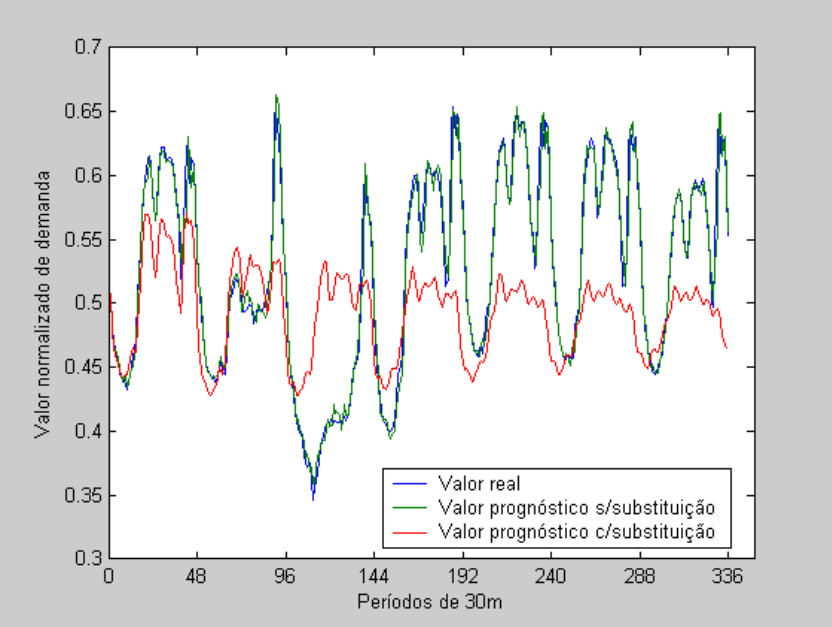


Figura 5.4 -Demanda real e prognóstico, sem e com utilização dos novos valores, para determinação de valores futuros, do período de 14.01.2000 a 20.01.2000, do caso I.

A Fig 5.5 mostra o erro de prognóstico para os casos da Fig 5.4, mostrando que utilizando valores calculados anteriormente para a determinação dos valores futuros, o erro não mantém uma regularidade no período do estudo, conduzindo o estudo a pesquisar outras formatações do conjunto de treinamento, que apresentem melhores resultados.

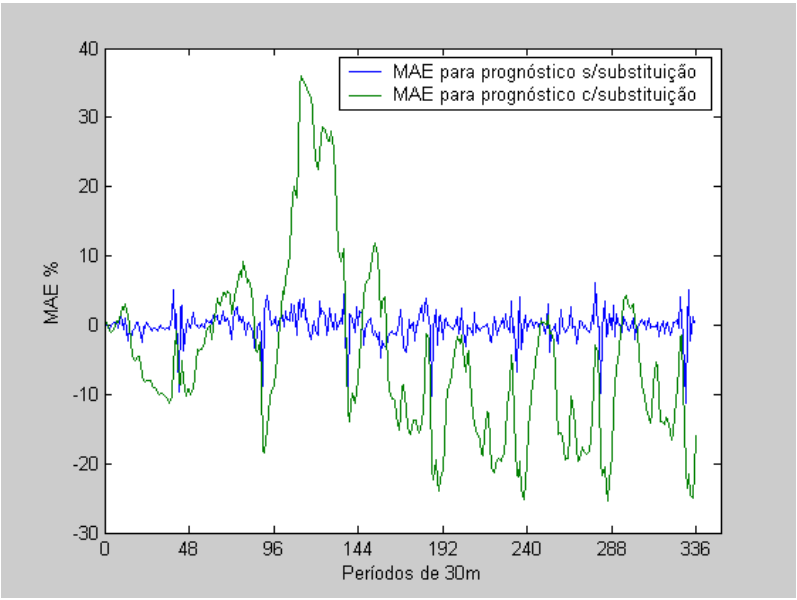


Figura 5.5 – Erro das curvas de prognóstico do caso I, da Fig 5.4.

Para uma função núcleo de produto interno foram processados vários casos, variando-se os parâmetros de entrada C e ε, procurando aqueles que apresentassem o menor erro MAE quando aplicados para prognóstico do mesmo conjunto de entrada, efetuando a substituição dos valores reais pelos valores obtidos na simulação, para determinar os valores futuros subseqüentes.

Dentre os valores processados, utilizando os valores prognosticados para obter valores futuros, o menor valor MAE=10,9324% foi encontrado para C=4 e ε=0,00225. Comparando a curva de demanda real com a respectiva curva de prognóstico(Fig 5.4) observa-se que não há um bom ajuste, apresentando uma curva do erro(Fig 5.5) para este caso com valores pontuais na ordem de 35%, entre o ponto 96 e 144 do eixo das abcissas, que corresponde ao domingo da semana em estudo. Pode-se concluir que para este formato do conjunto de entrada, considerando a diferença de comportamento da curva de demanda entre os dias de semana e finais de semana, não é possível obter um resultado satisfatório para o erro de prognóstico. De maneira geral não foi possível estabelecer uma possível relação entre os parâmetros C e ε e valores de entrada, do conjunto de treinamento, da máquina de vetor de suporte, de modo a obter uma melhor performance no resultado de prognóstico, exigindo o processamento de inúmeros casos para efetuar a escolha da melhor parametrização, sem, no entanto, apresentar garantia de tratar-se da melhor solução do problema.

2)Caso II.

O caso II teve como objetivo efetuar alterações no conjunto de treinamento e conseqüentemente no modelo de máquina de vetor de suporte visando contornar as dificuldades encontradas no caso I. O conjunto de treinamento abrange o mesmo período do caso anterior, no entanto, cada vetor de entrada é constituído por 8 componentes, mostrando o comportamento do ponto da curva de demanda nos seis valores que antecedem, o valor ocorrido 24h antes e o valor ocorrido 336h(1 semana) antes do valor alvo procurado, conforme apresentado na tabela da Fig 5.6.

Esta forma de composição do conjunto de treinamento particulariza cada valor alvo com um determinado conjunto de entrada, apresentando uma melhor performance da máquina de vetor de suporte.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme modelo da Fig 4.11, abrangendo o período de 14.01.2000 a 20.01.2000 (7 dias), totalizando 336 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	O mesmo que o do conjunto de treinamento.
Função núcleo	$K(x,z)=\langle x \cdot z \rangle$

Figura 5.6 – Estrutura do caso II

A Fig 5.7 apresenta de forma gráfica o erro médio absoluto como função das variáveis C e ε, quando da aplicação do conjunto de prognóstico ao algoritmo de programação quadrática.

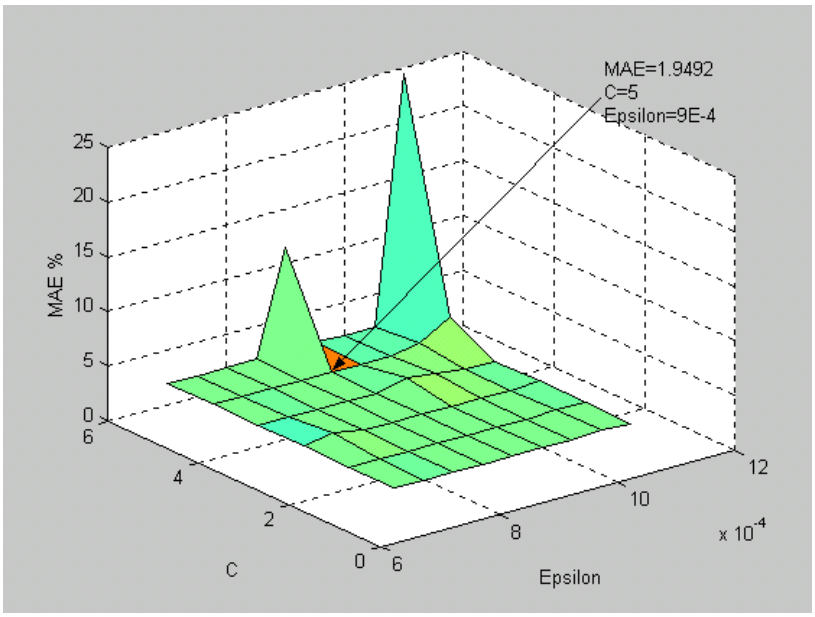


Figura 5.7 –Erro MAE=f(C,ε) do prognóstico, com utilização de valores calculados para obtenção de valores futuros, do caso II.

Os valores do erro MAE apresentam um valor médio em torno de 2%, podendo chegar em alguns casos a valores da ordem de 25%, não permitindo estabelecer uma possível relação entre as variáveis. O menor valor de erro encontrado, entre os casos processados, foi MAE=1,9492%, para C=5 e ε =0,0009.

A Fig 5.8 apresenta percentualmente o número de vetores suporte, definidos pelo estudo, para cada par de valores C e ε, onde se pode observar, da mesma forma, a não linearidade do processo. O número de vetores suporte encontrado para o menor valor MAE foi de 313, correspondendo a 93,15% do conjunto de vetores de treinamento.

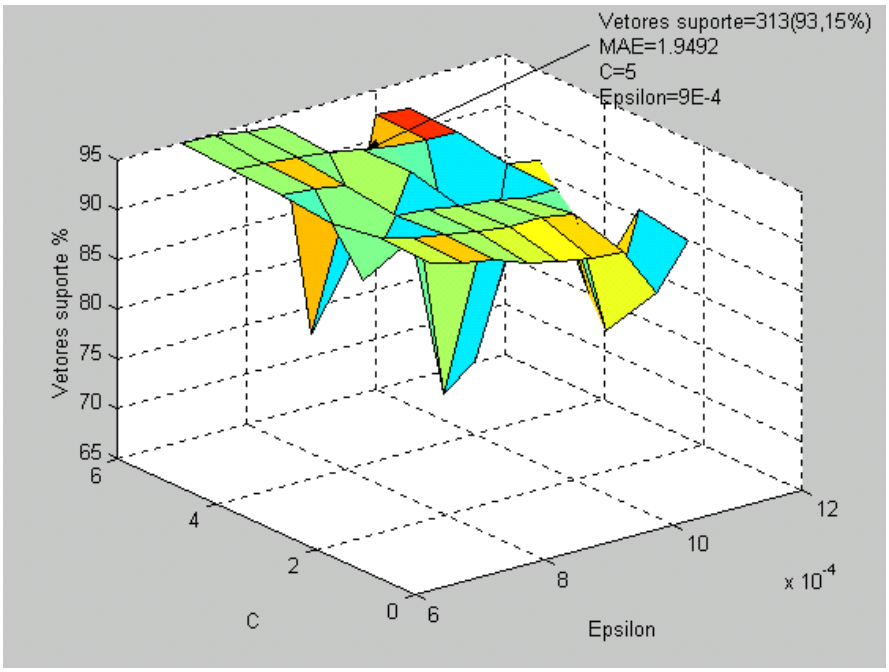


Figura 5.8 – Perfil do número de vetores suporte SV=f(C,ε) para o caso II.

A Fig 5.9 mostra graficamente as curvas de demanda de potência, real e duas com valores de prognóstico, sendo, com e sem a utilização dos valores obtidos no algoritmo para a definição dos valores subseqüentes.

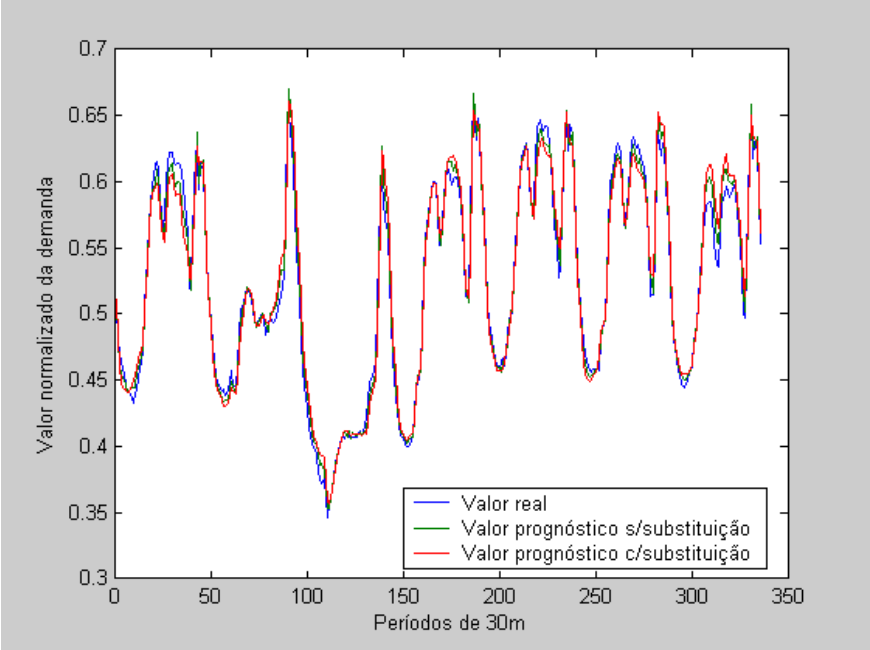


Figura 5.9 –Demanda real e prognóstico, sem e com utilização dos novos valores, para determinação de valores futuros do período de 14.01.2000 a 20.01.2000, do caso II.

A Fig 5.10 mostra o erro MAE para os casos da Fig 5.9, mostrando que com a utilização de valores calculados anteriormente para a determinação dos valores futuros, o erro(na ordem de 7% na pior condição) reduz-se substancialmente relativamente ao caso I, demonstrando uma melhora na estrutura do conjunto de treinamento da rede.

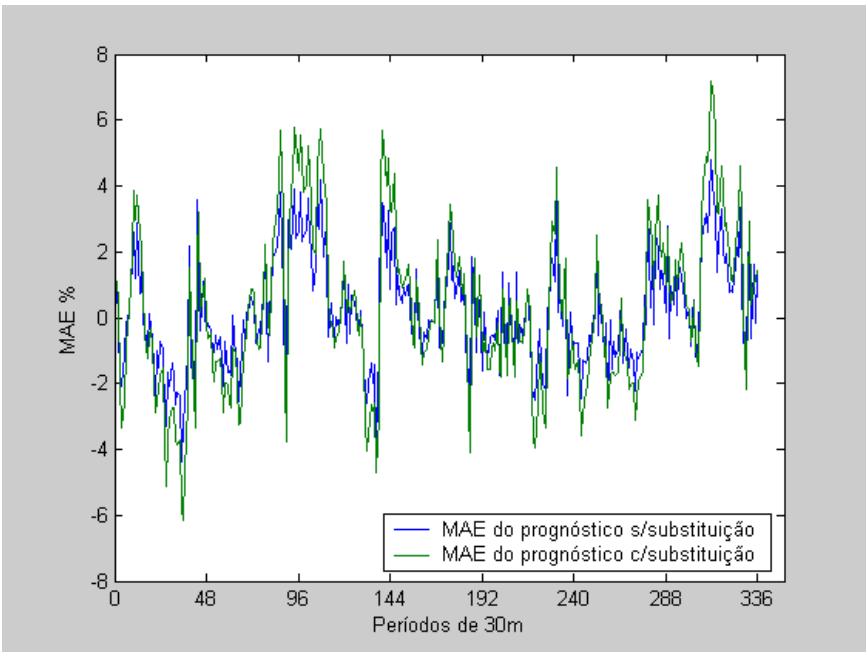


Figura 5.10 – Erro das curvas de prognóstico do caso II, da Fig 5.9.

Para uma função núcleo de produto interno e utilizando o mesmo conjunto para prognóstico, obteve-se uma redução significativa do valor do erro MAE=1,9492%, para valores C=5 e ε=0,0009 dos parâmetros, e uma aderência muito boa entre as curvas real e de prognóstico, da demanda(Fig 5.9), apresentando um erro pontual , para o mesmo ponto do caso anterior, da ordem de 6%(Fig 5.10), mostrando a melhora na escolha da modelagem do conjunto de treinamento e prognóstico. Os dois primeiros casos processados utilizaram como conjunto para diagnóstico o mesmo conjunto de treinamento, não caracterizando um estudo de caso de prognóstico de demandas futuras, o que será efetuado em todos os casos que se seguem, tendo por base a estruturação dos conjuntos de treinamento e de prognóstico da forma do caso II, qual seja, valores de demandas de potência componentes do vetor de entrada, antecedendo o valor alvo, de uma semana, um dia e os seis valores anteriores, nessa ordem.

3)Caso III.

O caso III foi estruturado mantendo a base de dados do conjunto de treinamento com 14 dias(2 semanas) completos, totalizando uma entrada de 672 vetores, deslocados de 30min no tempo, entre si. Neste caso o conjunto de prognóstico foi constituído pela semana subseqüente às do vetor de entrada, conforme apresentado na tabela da Fig 5.11.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme modelo da Fig 4.11, abrangendo o período de 10.01.2000 a 23.01.2000 (14 dias), totalizando 672 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme modelo da Fig 4.11, abrangendo o período de 24.01.2000 a 30.01.2000 (7 dias), totalizando 336 vetores.
Função núcleo	$K(x,z)=\langle x\cdot z\rangle$

Figura 5.11 – Estrutura do caso III.

Neste caso o conjunto de prognóstico é separado do conjunto de treinamento e apresenta um grau de dificuldade maior para obter um resultado satisfatório para o erro MAE, no prognóstico da semana subseqüente ao conjunto de treinamento.

Observou-se que, durante o processamento do algoritmo de programação quadrática, para um número de iterações diferentes, o resultado do erro, para valores iguais dos parâmetros(C e ε), apresenta valores diferentes do erro MAE, bem como do número de vetores suporte SV, demonstrando que há a necessidade de se efetuar validação da aproximação efetuada pelo processo, durante a execução do mesmo, o que pode ser obtido por um procedimento denominado validação cruzada(*cross validadtion)*, que divide o conjunto de treinamento em partes e efetua o teste de prognóstico dentro do próprio conjunto.

Nos estudos efetuados, foram processados vários casos com um número diferente de iterações do processamento, de forma a procurar aquele que apresentasse menor erro MAE no prognóstico, com utilização dos valores calculados para obtenção dos valores subseqüentes. A Fig 5.12 apresenta a variação verificada no percentual de vetores suporte e do valor do erro médio absoluto MAE, em função da variável ε para um valor C=8.

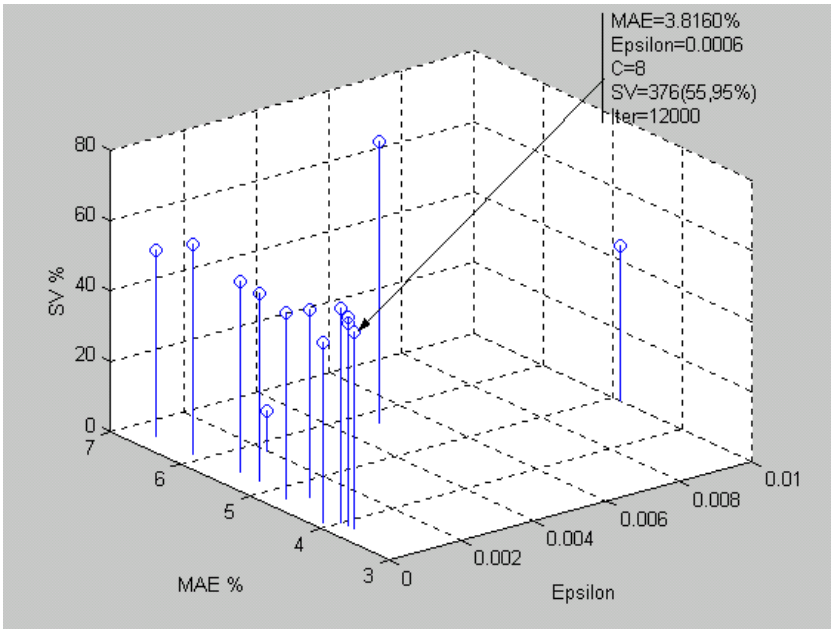


Figura 5.12 – Gráfico SV=f(MAE,ε) para C=8, do caso III.

A Fig 5.13 apresenta as curvas de demanda de potência, real e prognosticada efetuando a atualização de valores no vetor de entrada, para o caso indicado na Fig 5.12.

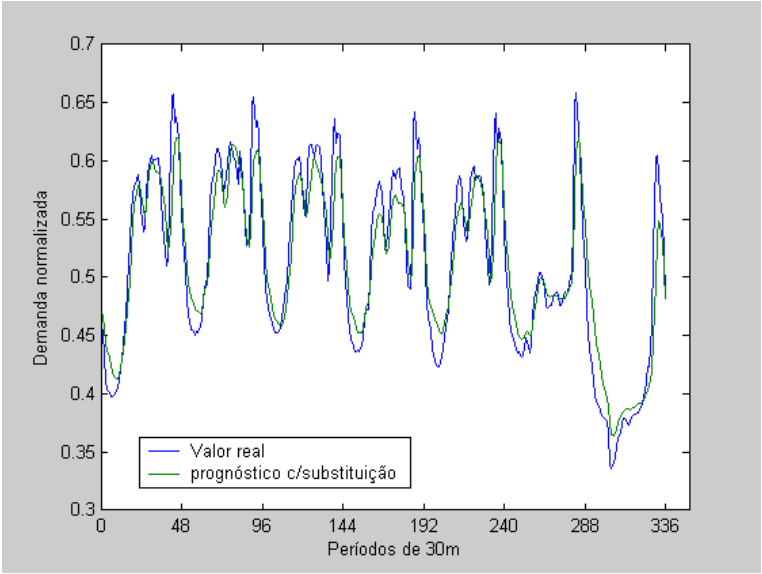


Figura 5.13 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 24.01.2000 a 30.01.2000, do caso III.

O melhor valor encontrado para o erro médio MAE foi 3,8160%, e a curva do erro da Fig 5.14 mostra que o ajuste não é satisfatório, principalmente nos valores de pico, tanto superiores quanto inferiores.

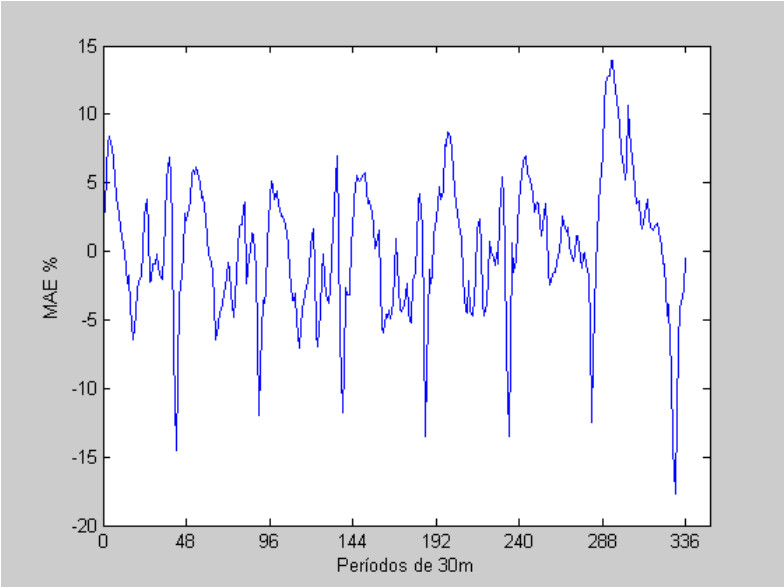


Figura 5.14 – Erro de prognóstico do caso III apresentado na Fig 5.13.

No processamento dos vários casos observou-se que atribuindo valores diferentes para o número de iterações, critério de parada para a convergência do processo de aproximação dos dados, obteve-se resultados MAE diferentes para o mesmo conjunto de parâmetros C e ε. O caso que apresentou o melhor resultado, foi obtido com 12000 iterações, parâmetros C= 8, ε= 0,0006, conduzindo a um valor MAE=3,8160% para uma função núcleo de produto interno. O ajuste entre as curvas real e de prognóstico da semana subseqüente(Fig 5.13) não ocorreu, observando-se diferenças significativas de valores tanto nos dias de semana(pontos situados entre 0 e 288 do eixo das abcissas) bem como no domingo(pontos situados entre 288 e 336). Os maiores erros pontuais(Fig 5.14) ocorreram nos pontos de demandas de pico e de mínima das curvas de demanda, com valores máximos da ordem de ±15%, observados na proximidade de 0h e 24h de domingo, respectivamente.

O acréscimo no erro deste caso, relativamente ao caso anterior, se deve ao fato de que a semana de prognóstico não faz parte do conjunto de treinamento, o que não ocorria no caso anterior. Para melhorar o desempenho da máquina de vetor de suporte(redução do erro de prognóstico), far-se-á estudos de caso com a separação dos dados de entrada em conjuntos que apresentem o mesmo comportamento da curva de demanda, assim, o conjunto de treinamento será constituído pelos dados da curva de demanda ocorridos de terça-feira até sexta-feira, que apresentam comportamentos similares.

4)Caso IV.

O caso IV adota a premissa descrita acima, tomando para conjunto de treinamento somente os dados de terça a sexta, do mesmo período(2 semanas) utilizado para o processamento do caso III. O conjunto para prognóstico utilizou dados de terça a sexta da semana subseqüente, permitindo uma comparação direta dos resultados obtidos com os do caso anterior por utilizarem o mesmo conjunto de dados para treinamento e prognóstico, conforme apresentada na tabela da Fig 5.15.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 11.01.2000 a 14.01.2000(terça a sexta-feira) e 18.01.2000 a 21.01.2000(terça a sexta-feira) (8 dias), totalizando 384 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, conforme estrutura da Fig 4.11, abrangendo o período de 25.01.2000 a 28.01.2000(terça a sexta) (4 dias), totalizando 192 vetores.
Função núcleo	$K(x,z)=\langle x \cdot z \rangle$

Figura 5.15 – Estrutura do caso IV.

A Fig 5.16 apresenta resultados obtidos com a estrutura pesquisada no caso IV, obtendo-se para menor valor do erro MAE=2.8984%, para C=15 e ε=0.0007.

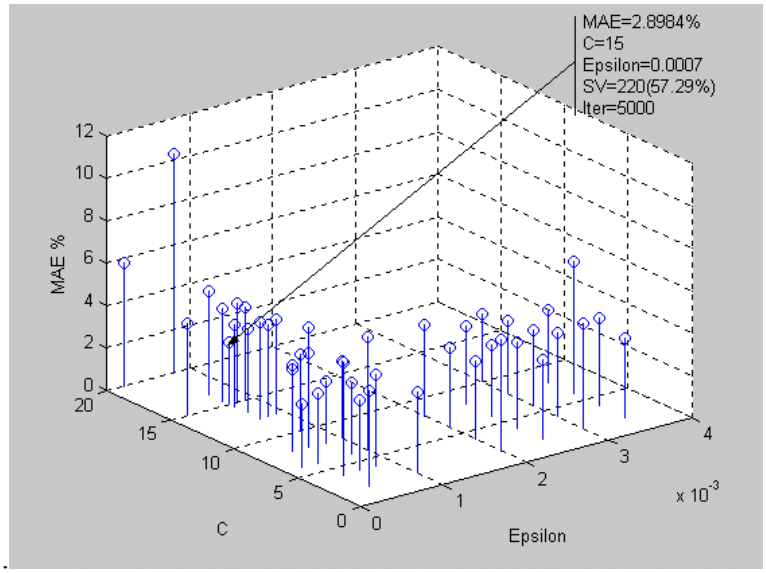


Figura 5.16 – Gráfico MAE=f(C,ε) do caso III.

A Fig 5.17 apresenta as curvas de demanda de potência, real e prognosticada efetuando a atualização de valores no vetor de entrada, para o caso IV.

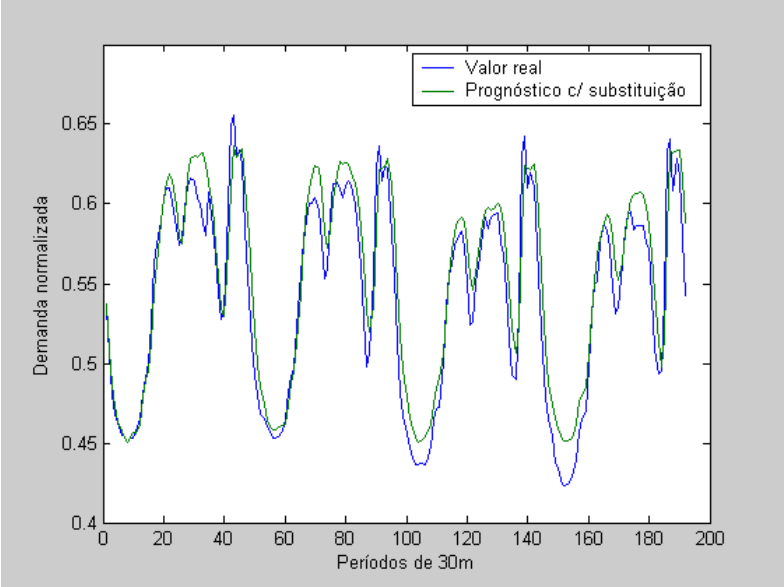


Figura 5.17 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 25.01.2000 a 28.01.2000, do caso IV.

A Fig 5.18 mostra o erro de prognóstico pontual onde para alguns valores de demanda de potência o ajuste apresenta erro na ordem de 8%, que pode conduzir a projeções de demandas não compatíveis com as necessidades do sistema elétrico em estudo.

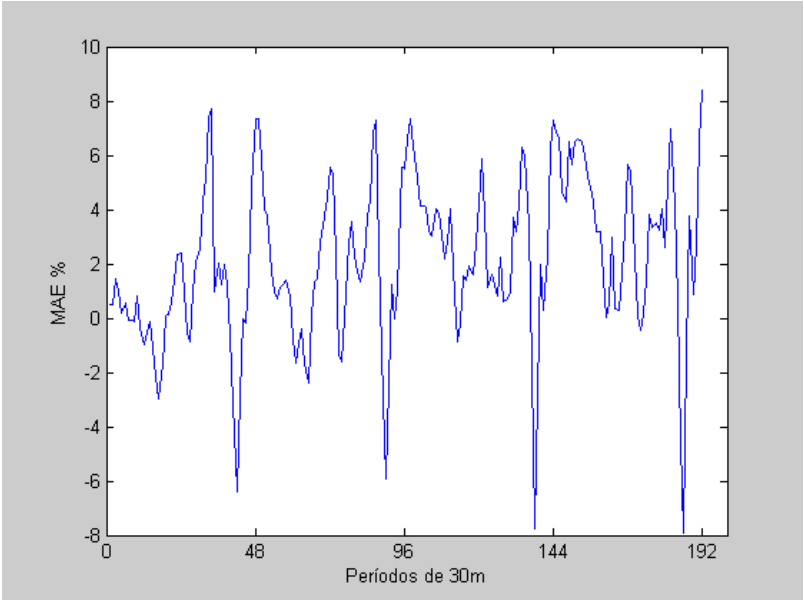


Figura 5.18 – Erro de prognóstico do caso IV apresentado na Fig 5.17.

O melhor resultado de erro MAE=2,8984% foi obtido com C=15 e ε=0,0007, representando uma redução de aproximadamente 25% do erro MAE em relação ao caso em que os conjuntos de treinamento e prognóstico incluem dados dos finais de semana(sábado,domingo,segunda). O ajuste das curvas(Fig 5.17) ainda não se apresenta visualmente satisfatória, com diferenças significativas nos picos de máxima e mínima, com erros pontuais na ordem de 8%(Fig 5.18), no entanto, relativamente ao caso anterior(Fig 5.14), apresenta uma redução expressiva, de quase 50% no erro, observado nos picos negativos da curva de erros pontuais.

5)Caso V.

Para o caso V foi mantida a separação dos dados de entrada entre meio de semana e fim de semana, sendo considerados, 12 dias, correspondendo a dados de 3 semanas de valores de terça a sexta-feira para o conjunto de treinamento, e uma semana subseqüente para conjunto de prognóstico, conforme apresentado na tabela da Fig 5.19.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 11.01.2000 a 14.01.2000(terça a sexta-feira), 18.01.2000 a 21.01.2000(terça a sexta-feira) e 25.01.2000 a 28.01.2000(terça a sexta) (12 dias), totalizando 576 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 01.02.2000 a 04.02.2000(terça a sexta) (4 dias), totalizando 192 vetores.
Função núcleo	$K(x,z)=\langle x \cdot z \rangle$

Figura 5.19 – Estrutura do caso V.

A separação dos dados entre valores de meio de semana e de final de semana já apresentou uma melhora no resultado de prognóstico, conforme mostrado no caso IV, e, com a finalidade de reduzir o erro de prognóstico, foi efetuado um acréscimo de semanas, ao conjunto de treinamento, visando incluir no projeto da máquina de vetor de suporte, o maior número de características diferentes de comportamento das curvas de demanda, para permitir um melhor ajuste da curva entre valores reais e de prognóstico.

A Fig 5.20 apresenta resultados obtidos com a estrutura pesquisada no caso IV, sendo obtido para menor valor do erro MAE=2.8523%, para C=15, ε=0.0008 e 7500 iterações do algoritmo, conforme indicado.

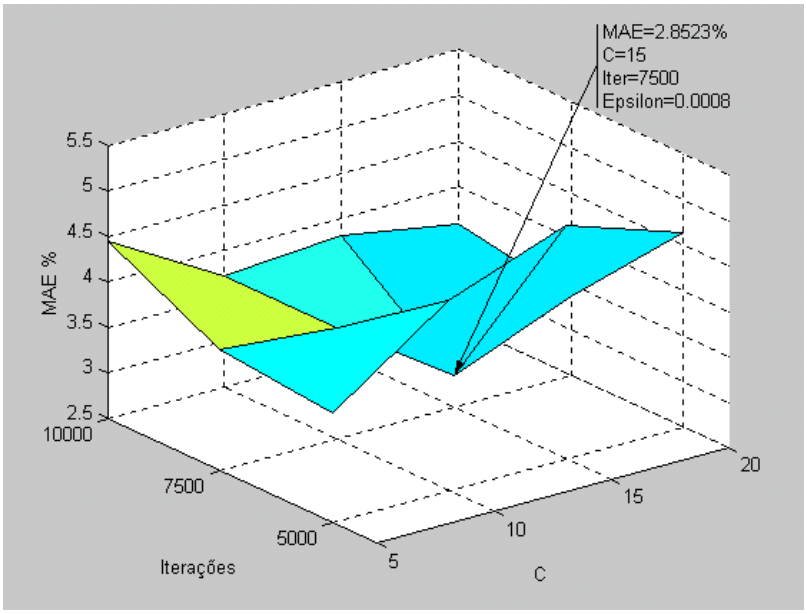


Figura 5.20 – Gráfico MAE=f(C,Iterações) com ε=0.0008 para o caso V.

A Fig 5.21 apresenta as curvas de demanda de potência, real e prognosticada com atualização de valores no vetor de entrada e a Fig 5.22 apresenta o erro de prognóstico, para o caso V.

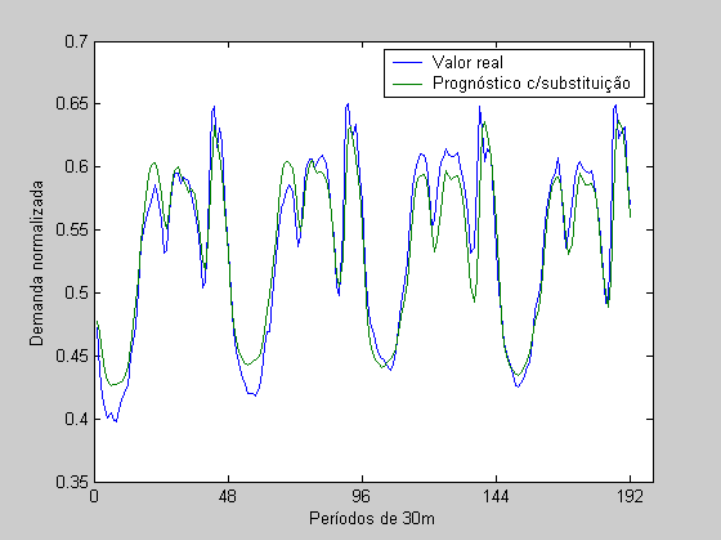


Figura 5.21 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 01.02.2000 a 04.02.2000, do caso V.

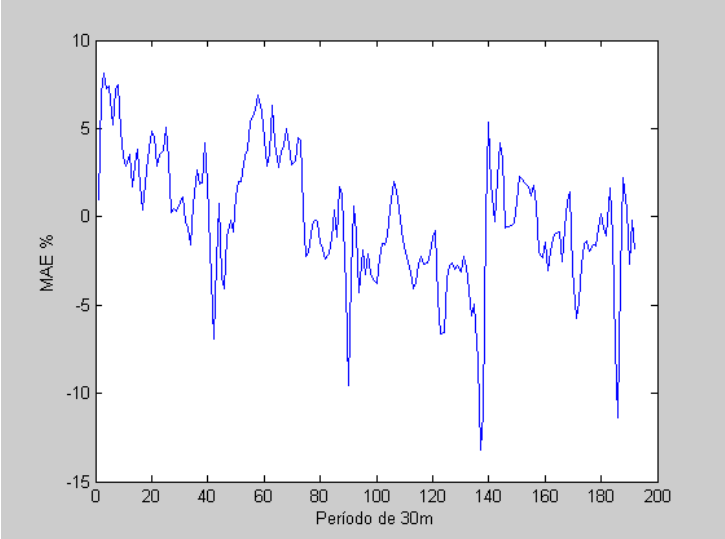


Figura 5.22 – Erro de prognóstico do caso V apresentado na Fig 5.21

O menor valor MAE=2,8523% foi obtido com os parâmetros C=15 e ε=0,0008, para um número de iterações do critério de parada igual a 7500 iterações. O ajuste entre as curvas, real e de prognóstico de demanda, não apresenta uma melhora em relação ao caso anterior em que o conjunto de treinamento foi processado com 2 semanas, pois o acréscimo de mais uma semana no conjunto de treinamento e o deslocamento do conjunto para prognóstico em uma semana, ainda não permitiu incorporar á massa de

dados de treinamento todas as possíveis variações que ocorrem no perfil da curva de demanda. No caso IV, o perfil do erro pontual(Fig 5.18) mostra uma alternância entre valores positivos(valor de prognóstico maior que valor real) e valores negativos(valor de prognóstico menor que valor real), enquanto no caso V, o perfil do erro mostra uma tendência para valores negativos de erro(valor de prognóstico menor que valor real).
Os resultados dos casos IV e V mostram a necessidade de incorporar um maior número de casos de treinamento, para a escolha, pela programação quadrática, daqueles vetores que representarão o subconjunto de características estáveis da série temporal em estudo.

6)Caso VI.

Para o caso VI o conjunto de treinamento está constituído por valores de demanda dos dias situados no meio da semana, como mostrado na tabela da Fig 5.23.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 11.01.2000 a 14.01.2000(terça a sexta-feira), 18.01.2000 a 21.01.2000(terça a sexta-feira), 25.01.2000 a 28.01.2000(terça a sexta) e 01.02.2000 a 04.02.2000(terça a sexta) (16 dias), totalizando 768 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 08.02.2000 a 11.02.2000(terça a sexta) (4 dias), totalizando 192 vetores.
Funções núcleo	$K(x,z) = \langle x \cdot z \rangle$ $K(x,z) = \left(1 + \langle x \cdot z \rangle\right)^2$ $K(x,z) = \exp\left(-\gamma \left\| x - z \right\|^2\right)$

Figura 5.23 – Estrutura do caso VI.

Neste caso o conjunto de treinamento foi acrescido em uma semana, sendo constituído por 4 semanas de valores de terça a sexta, totalizando 768 vetores de entrada. O tempo de processamento, a cada acréscimo de novos vetores, cresce de forma exponencial, dificultando a pesquisa do melhor resultado a partir da variação dos parâmetros da máquina.

A Fig 5.24 apresenta a curva de demanda real e de prognóstico com substituição dos valores, utilizando função núcleo de produto interno, do caso VI.

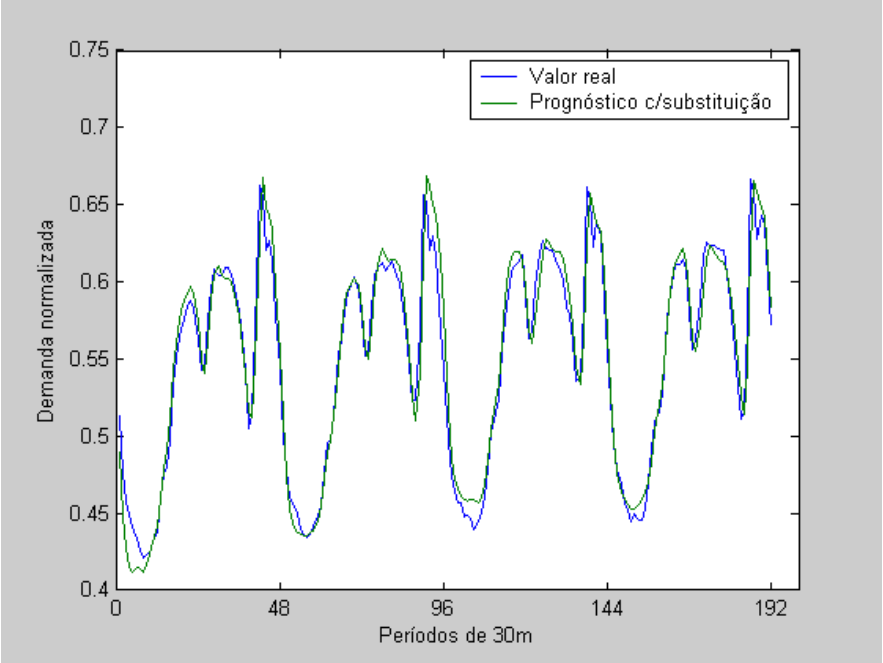


Figura 5.24 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 08.02.2000 a 11.02.2000, do caso VI com função núcleo de produto interno.

A Fig 5.25 mostra a curva do erro pontual do prognóstico apresentado na Fig 5.24, considerando o sinal, ou seja, erros positivos correspondem a valor real maior que valor prognosticado para o período de tempo considerado.

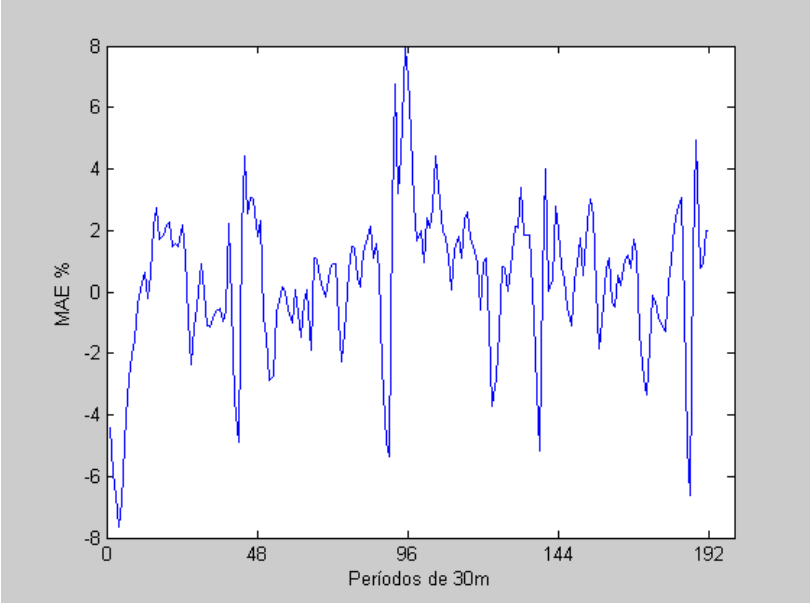


Figura 5.25– Erro de prognóstico do caso VI, apresentado na Fig 5.24.

A Fig 5.26 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.27 a curva do erro de prognóstico com função núcleo polinomial, do caso VI.

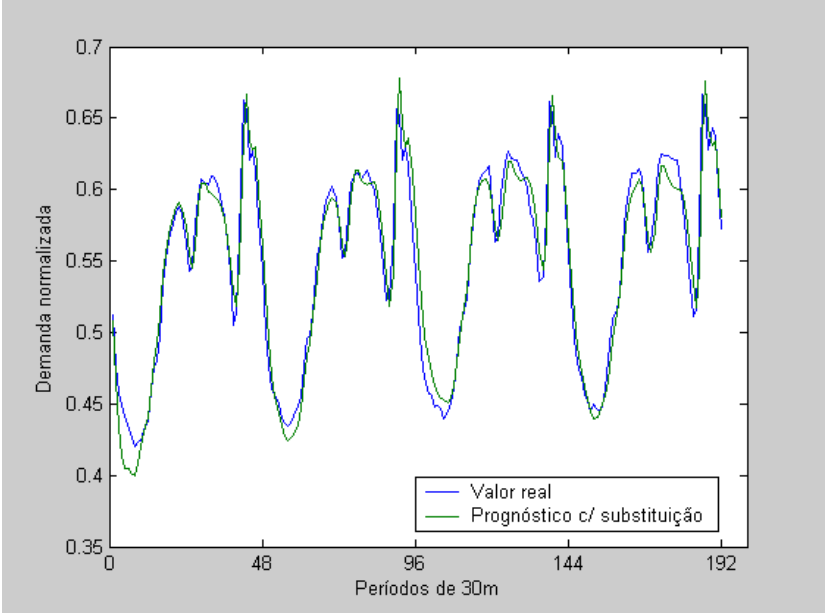


Figura 5.26- Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 08.02.2000 a 11.02.2000, do caso VI com função núcleo polinomial.

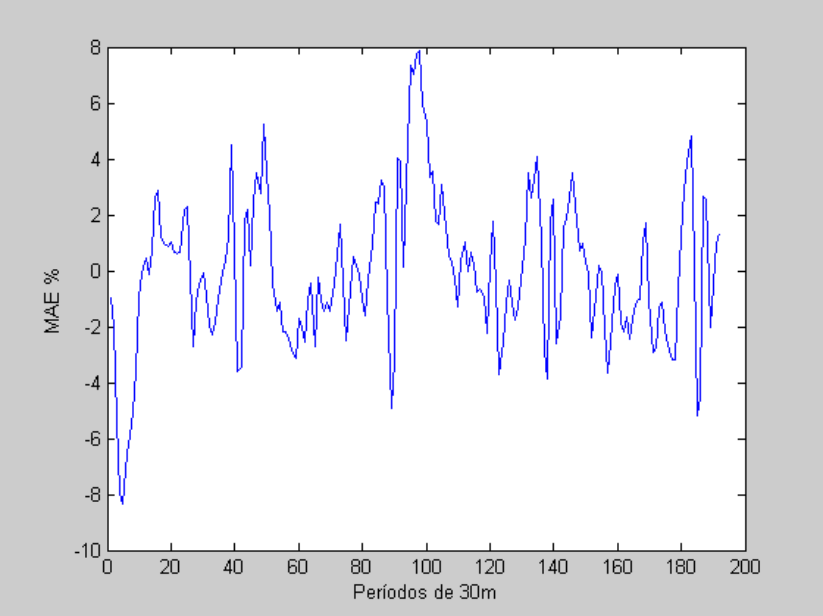


Figura 5.27 – Erro de prognóstico do caso VI, apresentado na Fig 5.26.

A Fig 5.28 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.29 apresenta a curva do erro de prognóstico com função núcleo de base radial, do caso VI.

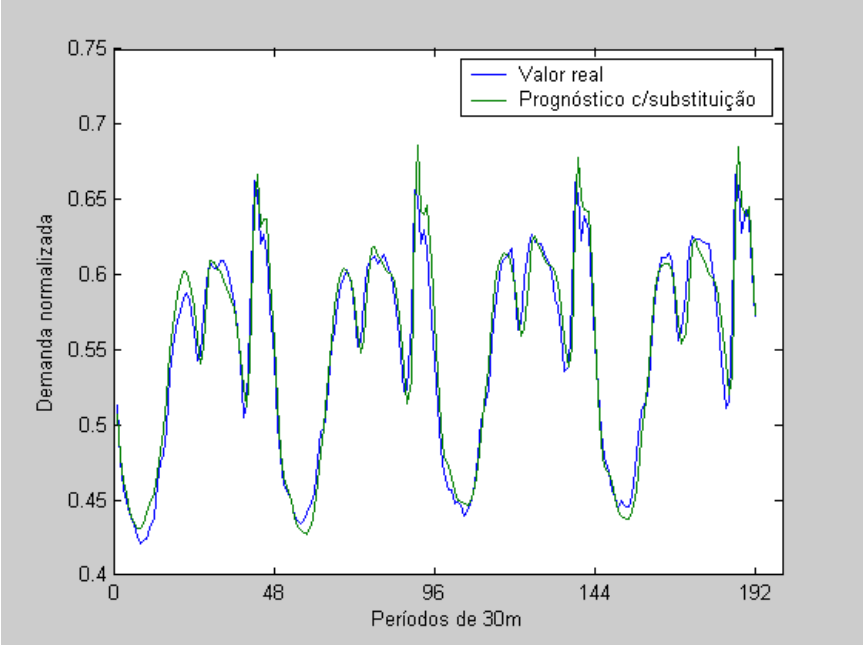


Figura 5.28 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 08.02.2000 a 11.02.2000, do caso VI com função núcleo de base radial.

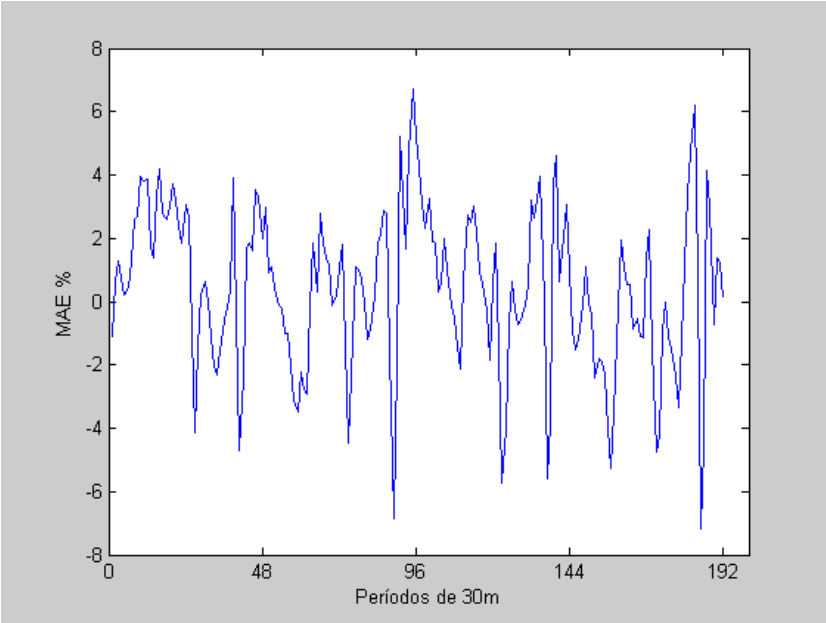


Figura 5.29 – Erro de prognóstico do caso VI, apresentado na Fig 5.28.

O melhor resultado para a função produto interno foi MAE=1,8364% para C=15 e ε=0,0008, obtendo-se uma melhora para o ajuste global das curvas de demanda(Fig 5.24), apresentando erros pontuais positivos(valor de prognóstico maior que valor real) da ordem de 4%(Fig 5.25) com um valor de erro discrepante, da ordem de 8%, e erros pontuais negativos(valor de prognóstico menor do que valor real) da ordem de 6%, representando uma melhora aproximada de 35% no erro MAE para uma semana de prognóstico, e 25% no erro de pico pontual, mostrando que o acréscimo de vetores no conjunto de treinamento, incorpora novas características do perfil de demanda conduzindo a uma melhora do perfil de prognóstico de valores futuros.

Para o caso VI com a utilização de função núcleo polinomial o melhor resultado, dentre os pesquisados, foi MAE=2,0555% para C=15 e ε =0,0015, sendo que a curva dos perfis de demanda(Fig 5.26) mostra um ajuste melhor, com valores de erros pontuais positivos(valor de prognóstico maior que valor real) na ordem de 4%, com um erro discrepante de 8%, e erros pontuais negativos (valor de prognóstico menor que valor real) chegando a 4%, com um erro discrepante de 8%, que, mesmo apresentando um acréscimo de aproximadamente 10% no erro MAE, reduz em 33% (de 6% para 4%) o erro pontual negativo, relativamente ao caso com função núcleo de produto interno.

Para o caso processado com função núcleo de base radial, o melhor valor dentre os pesquisados foi MAE=2,0188% para C=30, ε =0,0015 e γ =2, que apresenta características de ajuste entre perfis de demanda real e valores de prognóstico similares, encontrados para os casos das outras duas funções núcleo, utilizadas na pesquisa.
O erro pontual(Fig 5.29) máximo chegou a aproximadamente 7%, sendo que a curva indicativa mostra uma tendência a valores de erro pontual negativo(valor de prognóstico menor do que valor real).

7)Caso VII.

O caso VII foi pesquisado utilizando uma rede de função de base radial, com os dados do conjunto de treinamento e conjunto para prognóstico do caso VI, conforme apresentado na tabela da Fig 5.30.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 11.01.2000 a 14.01.2000(terça a sexta-feira), 18.01.2000 a 21.01.2000(terça a sexta-feira), 25.01.2000 a 28.01.2000(terça a sexta) e 01.02.2000 a 04.02.2000(terça a sexta) (16 dias), totalizando 768 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 08.02.2000 a 11.02.2000(terça a sexta) (4 dias), totalizando 192 vetores.
Função núcleo	Rede de função de base radial

Figura 5.30 – Estrutura do caso VII.

A Fig 5.31 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.32 apresenta a curva do erro de prognóstico da rede de função de base radial para o caso VII.

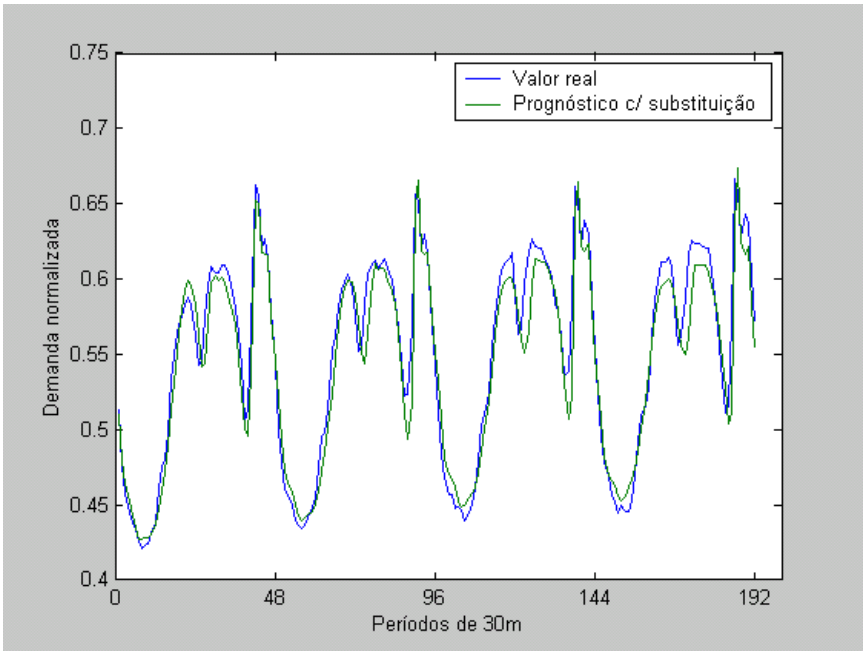


Figura 5.31 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 08.02.2000 a 11.02.2000, do caso VII.

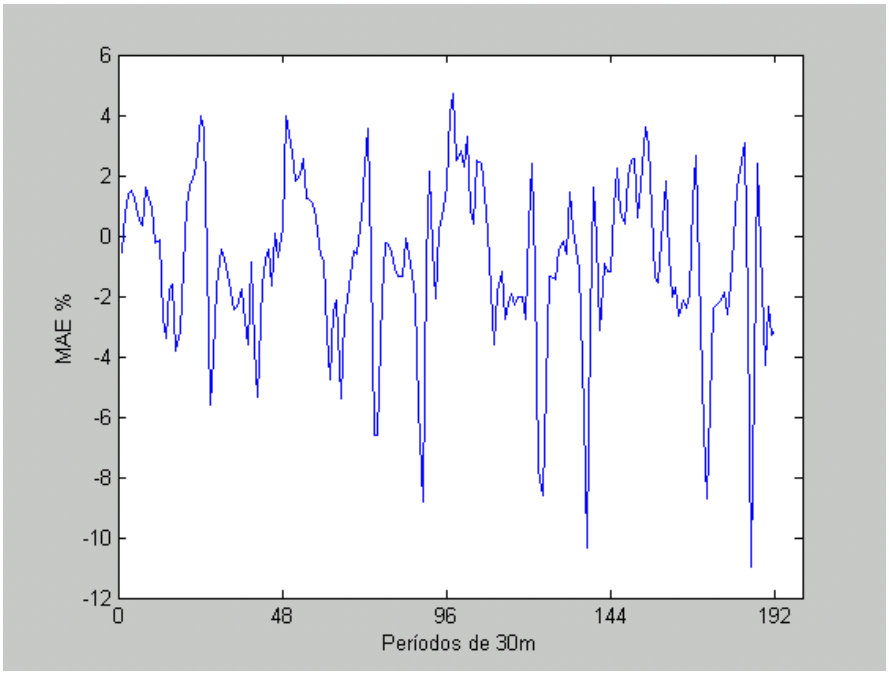


Figura 5.32 - Erro de prognóstico do caso VII, apresentado na Fig 5.31.

O melhor resultado foi MAE=2,1945% para um espalhamento 0,3 e erro de treinamento 0,09%. As curvas dos perfis real e de prognóstico(Fig 5.31), assim como a do erro pontual(Fig 5.32) mostram um crescimento do erro a medida que a curva se afasta do ponto inicial de prognóstico. O erro inicia com picos positivos e negativos na ordem de 4%, mantém os erros pontuais positivos em 4% crescendo os erros pontuais negativos até aproximadamente 11%, para os últimos valores.

8)Caso VIII.

O caso VIII utiliza para conjunto de treinamento valores de sábados, domingos e segundas-feiras, apontados como apresentando características diferentes dos demais dias da semana, conforme mostrado na tabela da Fig 5.33.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo os períodos de 15.01.2000 a 17.01.2000, 22.01.2000 a 24.01.2000, 29.01.2000 a 31.01.2000, 05.02.2000 a 07.02.2000 e 12.02.2000 a 14.02.2000(sábados, domingos e segundas-feira)(15 dias), totalizando 720 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 19.02.2000 a 21.02.2000(sábado, domingo e segunda-feira) (3 dias), totalizando 144 vetores.
Funções núcleo	$K(x,z) = \langle x \cdot z \rangle$ $K(x,z) = \left(1 + \langle x \cdot z \rangle\right)^2$ $K(x,z) = \exp\left(-\gamma \left\| x - z \right\|^2\right)$

Figura 5.33 – Estrutura do caso VIII.

Este caso se diferencia dos demais pelo fato de que as curvas de demanda dos três dias, sábado, domingo e segunda-feira, não apresentarem o mesmo comportamento, diferentemente do que ocorria no caso apresentado com dias da semana de terça a sexta-feira, tornando o processo de ajuste mais difícil de ser obtido.

O caso foi processado com três funções núcleo, diferentes, mantendo-se fixo o valor de ε=0,0025 e assumindo valores para C=[1,4 - 3,0 - 5,0 - 7,5 e 10,0] e um número de iterações do processo de ajuste igual a 7500, com a finalidade de obter o perfil de desempenho geral do ajuste como função das variáveis.

As Figuras 5.34 e 5.35 mostram, respectivamente, a evolução do erro MAE para o prognóstico de valores futuros do período de 19.02.2000 a 21.02.2000, e, o percentual de vetores suporte assumidos pelo processamento do conjunto de treinamento, para a definição da máquina de vetor de suporte, com funções núcleo de produto interno, polinomial e de base radial.

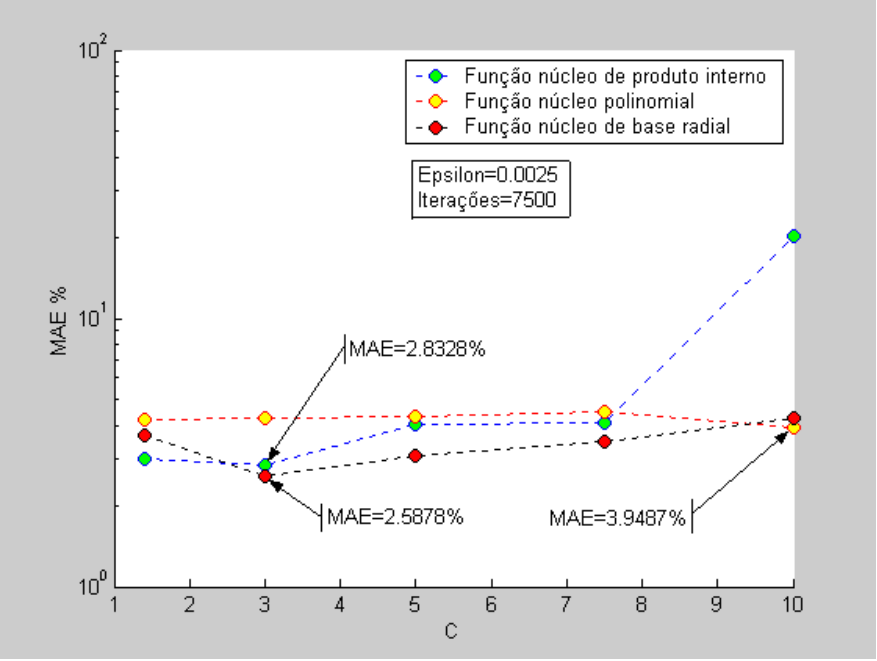


Figura 5.34 – Erro MAE em função de C para três funções núcleo do caso VIII.

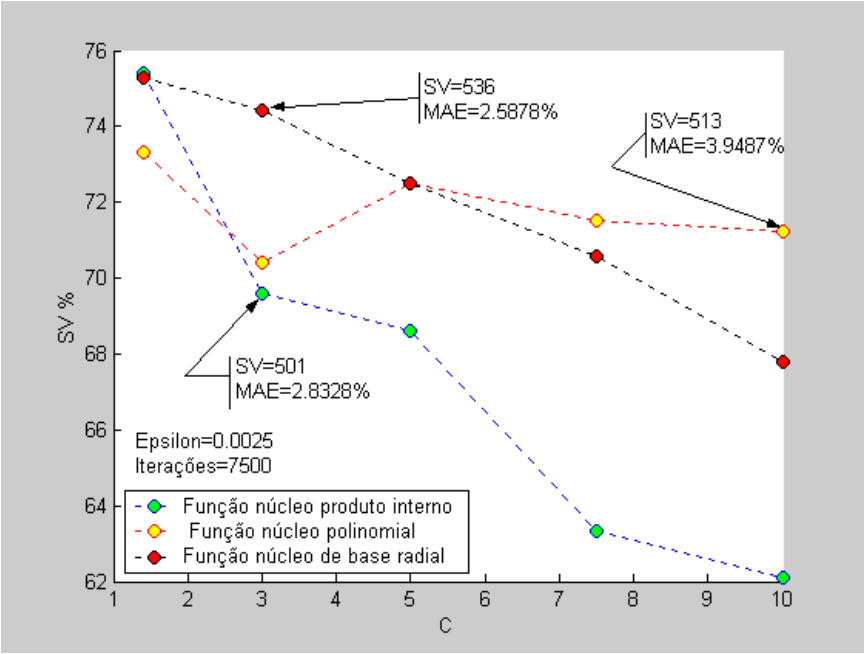


Figura 5.35 – Percentual de vetores suporte em função de C e da função núcleo - caso VIII.

A Fig 5.36 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.37 apresenta a curva do erro de prognóstico, com função núcleo de produto interno, do caso VIII.

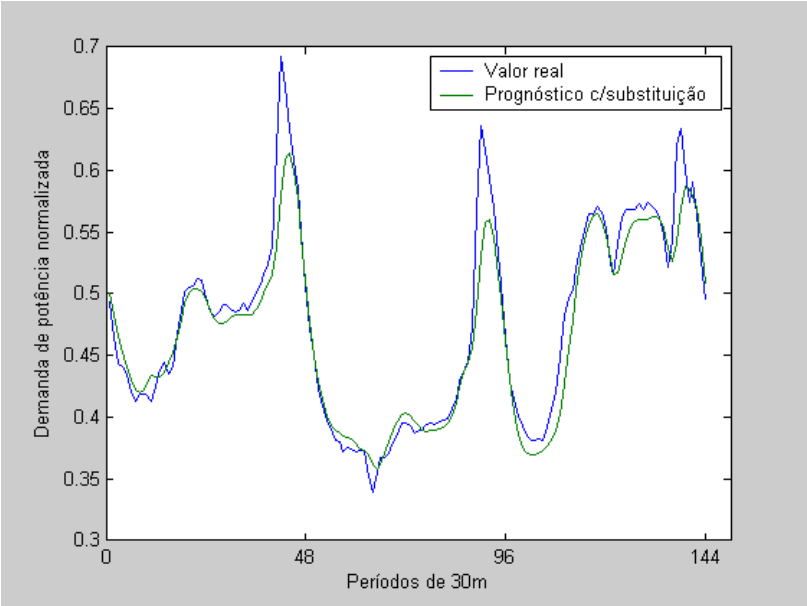


Figura 5.36 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 19.02.2000 a 21.02.2000, do caso VIII com função núcleo de produto interno.

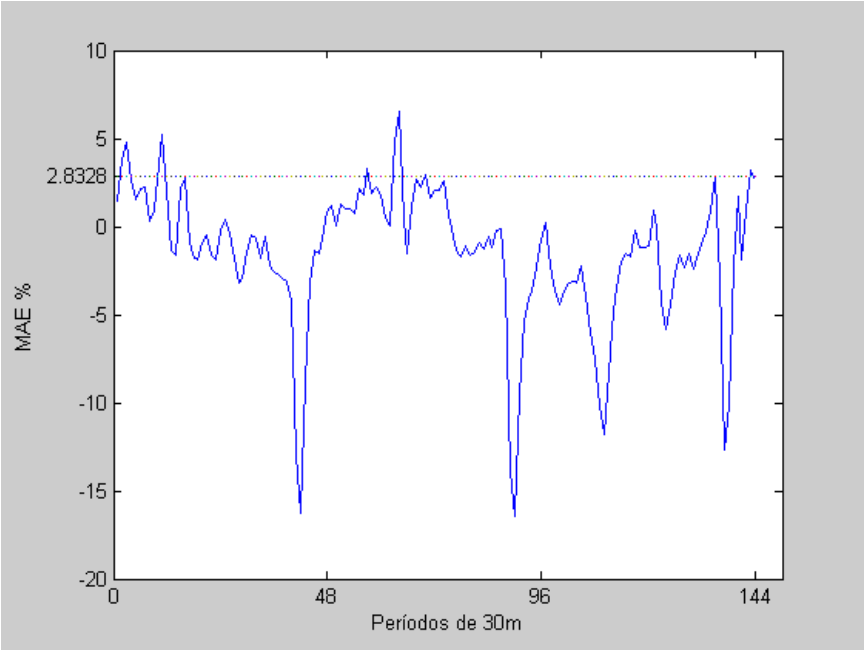


Figura 5.37 – Erro de prognóstico do caso VIII, apresentado na Fig 5.36.

A Fig 5.38 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.39 apresenta a curva do erro MAE com função núcleo polinomial, do caso VIII.

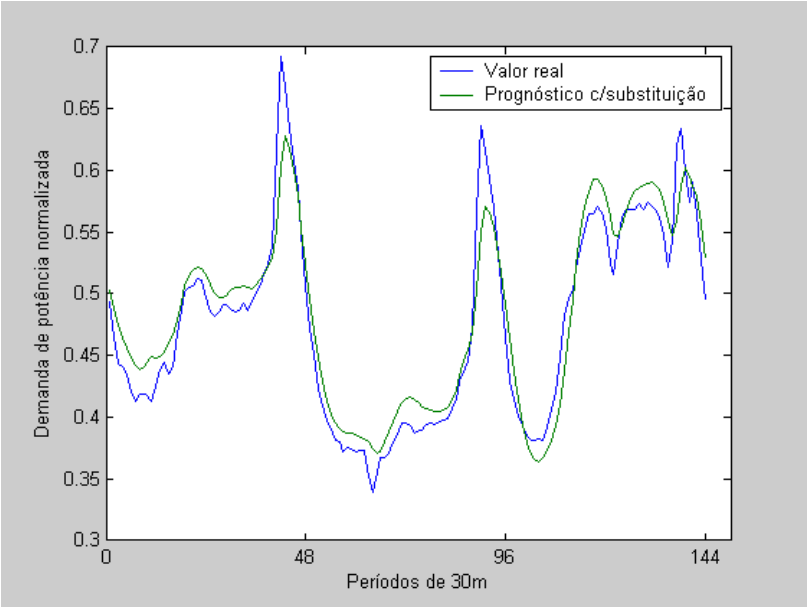


Figura 5.38 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 19.02.2000 a 21.02.2000, do caso VIII com função núcleo polinomial.

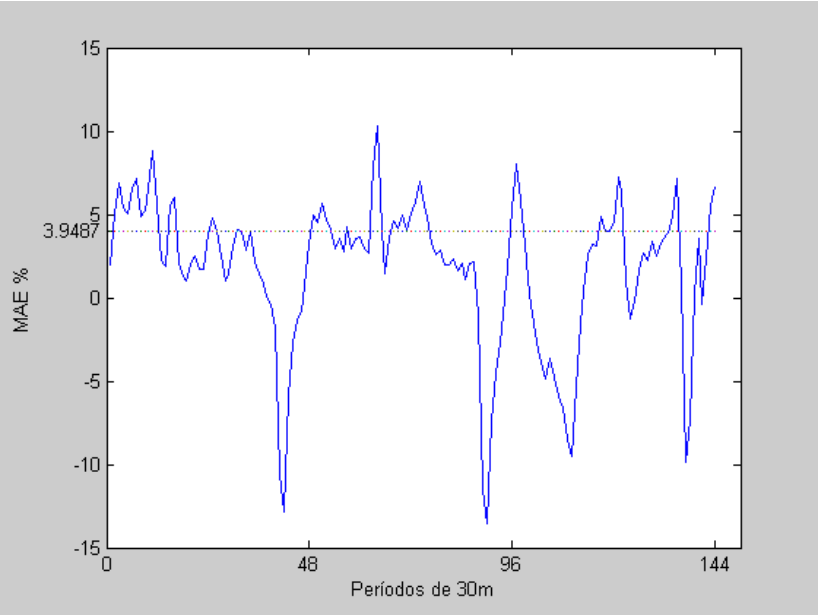


Figura 5.39 – Erro de prognóstico do caso VIII, apresentado na Fig 5.38.

A Fig 5.40 apresenta a curva de demanda real e de prognóstico com substituição dos valores, e a Fig 5.41 apresenta a curva do erro MAE com função núcleo de base radial, do caso VIII.

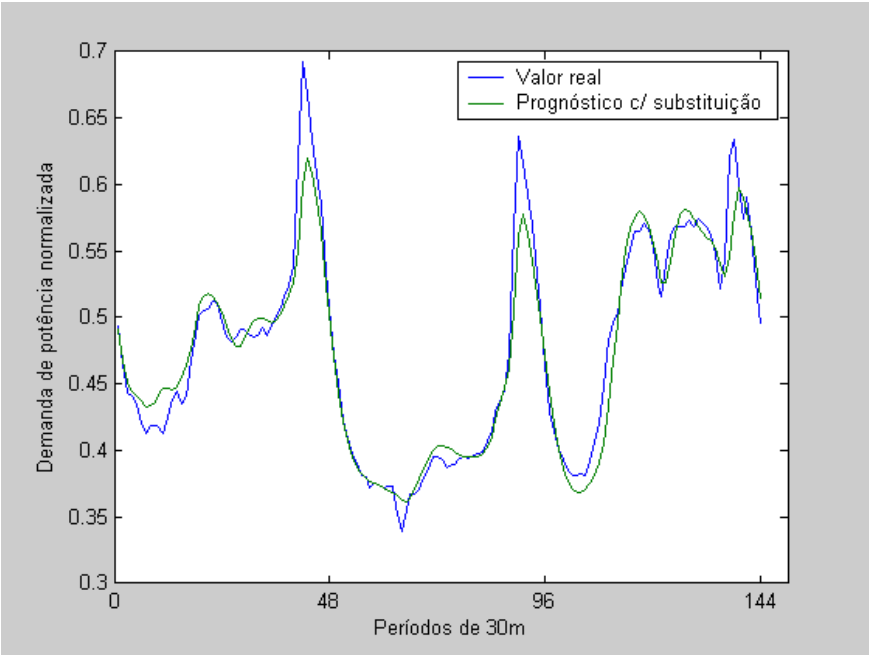


Figura 5.40 - Demanda real e prognóstico com utilização dos novos valores, para determinação de valores futuros do período de 19.02.2000 a 21.02.2000, do caso VIII com função núcleo de base radial.

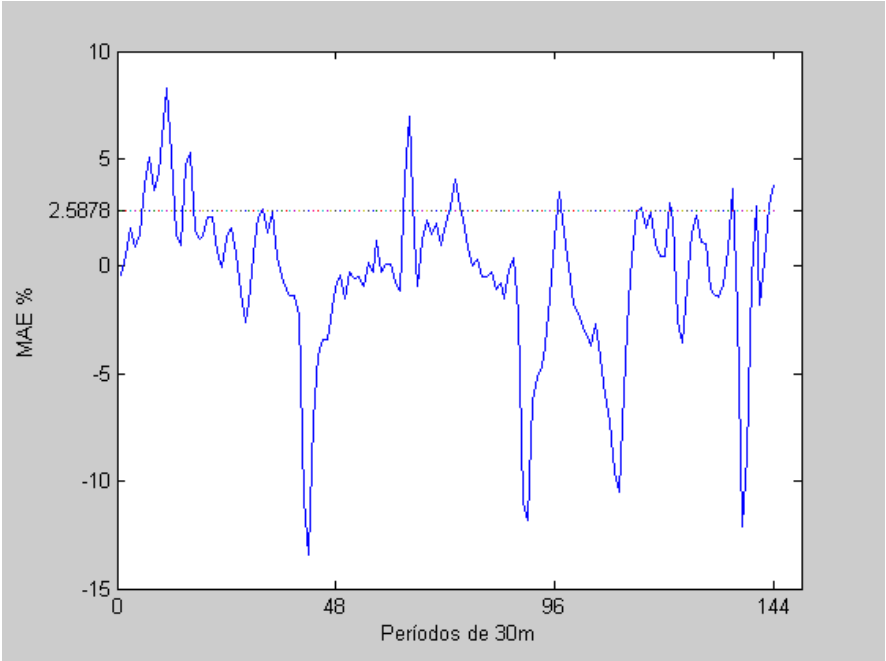


Figura 5.41 – Erro de prognóstico do caso VIII, apresentado na Fig 5.40.

Os menores valores MAE obtidos estão indicados nas Figs 5.34 e 5.35 e correspondem a 2,8328% para C=3,0 e uma função núcleo de produto interno, 3,9487% para C=10,0 e função núcleo polinomial e 2,5878% para C=3,0 e função núcleo de base radial. Para todos os casos foi assumido um valor ε=0,0025.

Os valores obtidos não esgotam as redes que possam apresentar valores melhores de ajuste, uma vez que o tempo de processamento necessário para obter os resultados é extremamente longo, crescendo muito com o acréscimo de novos vetores de entrada ao conjunto de treinamento. A Fig 5.34, que relaciona o erro MAE com a variável C, mostra que não existe uma variação substancial no valor do erro de prognóstico para as redes obtidas com diferentes funções núcleo, a não ser o caso de função núcleo de produto interno que para um valor maior de C o erro também cresce.

A Fig 5.35 mostra, claramente, uma redução no número de vetores suporte com o crescimento do valor da variável C, encontrando-se, para o melhor resultado do valor MAE=2,5878%, obtido com C=3,0, γ=2,0 e função núcleo de base radial, a participação de 536 vetores suporte, representando 74,4% do total de vetores do conjunto de treinamento. O menor erro MAE processado com a função núcleo de produto interno foi 2,8328%, sendo que o ajuste, entre a curva real e de prognóstico, conforme mostrado na Fig 5.36, apresenta um erro pontual elevado, da ordem de 15%, conforme mostrado na

Fig 5.37, nos pontos de pico do sábado e domingo, e de 10% no pico de segunda-feira. Para o caso processado com função núcleo polinomial, conforme mostrado nas Figs 5.38 e 5.39, apesar do erro MAE=3,9487%, o ajuste entre curvas real e de prognóstico apresentou redução do erro nos valores de pico, na ordem de 10% e aumento do erro nos demais valores, quando comparado com a função núcleo de produto interno. Para o caso processado com função núcleo de base radial o erro MAE=2,5878% apresentou o menor resultado entre os três casos, no entanto, conforme Figs 5.40 e 5.41 observa-se que, o ajuste entre as curvas de demanda, real e de prognóstico, não apresenta uma melhoria em relação aos demais casos, bem como o erro pontual é superior a 10% para os valores de pico da curva de demanda.

7)Caso IX.

O estudo de caso IX foi processado com o mesmo conjunto de treinamento e prognóstico do caso VIII, conforme apresentado na tabela da Fig 5.42, utilizando uma rede de função de base radial.

Conjunto de treinamento	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, conforme estrutura da Fig 4.11, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo os períodos de 15.01.2000 a 17.01.2000, 22.01.2000 a 24.01.2000, 29.01.2000 a 31.01.2000, 05.02.2000 a 07.02.2000 e 12.02.2000 a 14.02.2000(sábados, domingos e segundas-feiras)(15 dias), totalizando 720 vetores. Vetor de saída, único, correspondendo ao valor que se sucede imediatamente após o último componente, do vetor, de cada entrada.
Conjunto de prognóstico	Vetores de entrada, constituídos por um conjunto de 8 componentes cada um, correspondendo à demandas integralizadas de 30 minutos, que antecedem o valor alvo, abrangendo o período de 19.02.2000 a 21.02.2000(sábado, domingo e segunda-feira) (3 dias), totalizando 144 vetores.
Função núcleo	Rede de função de base radial

Figura 5.42 – Estrutura do caso IX.

O perfil do erro MAE no prognóstico, em função do espalhamento e do erro de treinamento, está apresentado na Fig 5.43.

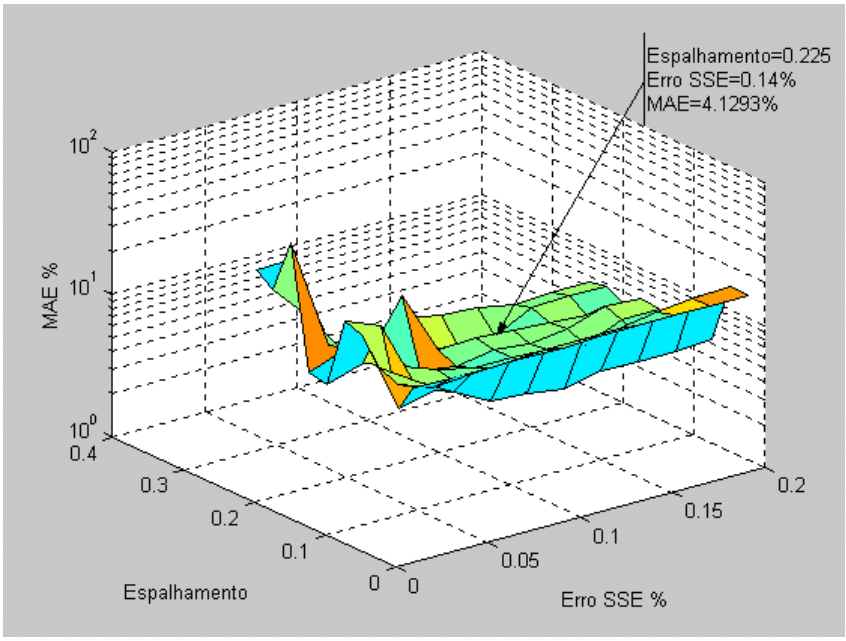


Figura 5.43 – Perfil do erro de prognóstico MAE, como função do erro de treinamento e do grau de espalhamento, do caso IX.

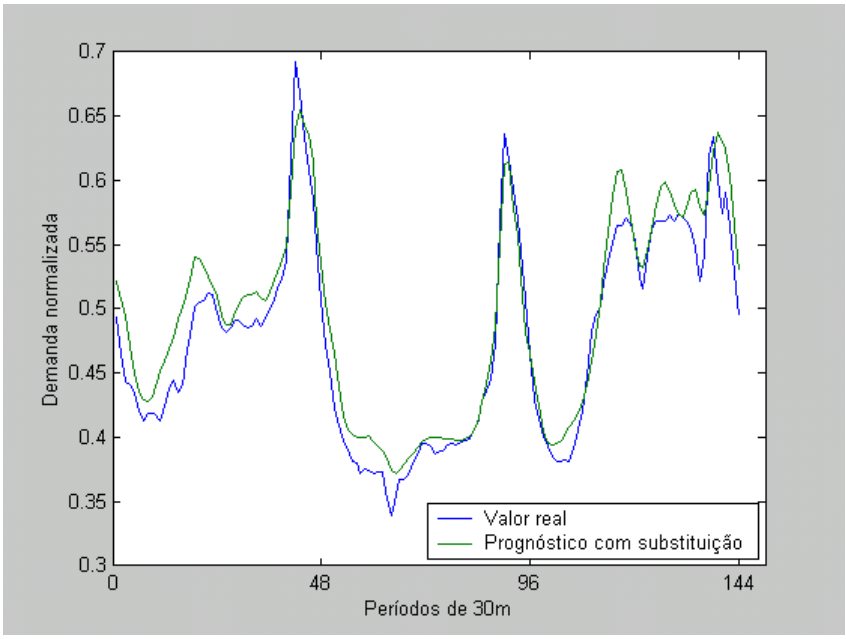


Figura 5.44 - Demanda real e prognóstico com utilização dos novos valores para determinação de valores futuros, do período de 19.02.2000 a 21.02.2000, do caso IX.

As Figs 5.44 e 5.45 apresentam, as curvas de demanda de potência real e prognosticada do período escolhido para simulação de valores futuros e a curva correspondente do erro pontual, respectivamente.

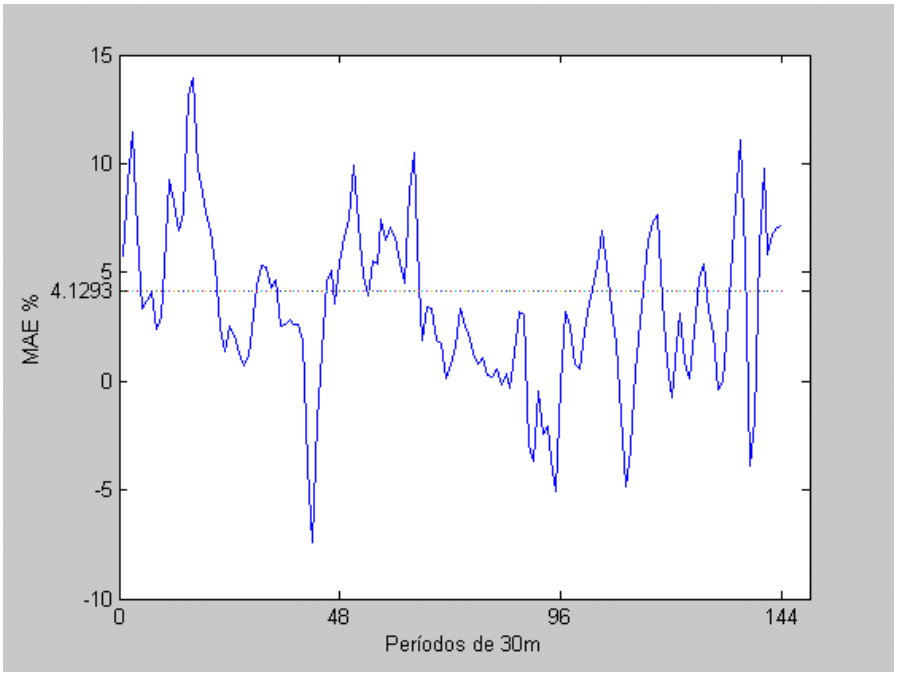


Figura 5.45 - Erro de prognóstico do caso IX, apresentado na Fig 5.44.

Conforme mostra a Fig 5.43, onde está apresentado o conjunto de casos processados, o menor valor encontrado foi MAE= 4,1293%, para um erro de aproximação do processamento de 0,14% e um espalhamento de 0,225. O ajuste da curva de prognóstico com a curva real, conforme mostrado na Fig 5.44, apresenta-se melhor do que nos casos processados com a máquina de vetor de suporte, que se deve a uma quantidade maior de casos estudados, que conduziu à escolha de uma melhor solução.

A curva de prognóstico de demanda apresenta uma maior aproximação com os valores reais nos picos de demanda, e uma maior separação nos pontos da segunda-feira, situados entre os pontos 96 e 144 do eixo de tempos. A Fig 5.45 mostra a curva de erro pontual, onde é observável que o erro varia entre +14% e –7%, apresentando aproximadamente 92% dos erros pontuais fora do intervalo ±2%, que no caso da máquina de vetor de suporte corresponde aproximadamente a 63% dos erros para o estudo de caso com função núcleo de base radial.

5.2 Considerações finais.

A análise dos resultados obtidos nos estudos de caso permite tecer as seguintes considerações a respeito do desempenho da máquina de vetor de suporte para prognóstico da demanda de potência elétrica:

•Os resultados dos prognósticos efetuados são compatíveis com as necessidades de planejamento e operação de sistemas elétricos no que diz respeito à aproximação com o perfil real da demanda de potência e no valor do erro médio absoluto, necessitando melhorar o modelo para o prognóstico de valores de picos de demanda.

• Não é observável uma possível relação entre os parâmetros C e e da máquina de vetor de suporte, de forma a permitir o estabelecimento de relações que facilitem a escolha dos valores ótimos para o processamento.

• Para o caso brasileiro, deve-se observar que os conjuntos utilizados para treinamento e prognóstico de demanda de potência estejam situados dentro de um mesmo período de perfil de consumo(horário de verão ou horário normal).

•Os valores do erro médio absoluto, encontrados nos estudos de caso são superiores àqueles dos trabalhos apresentados no texto para outros métodos. Isso se deve ao fato de que a região de ocorrência da demanda é, geograficamente, bastante extensa (Estado do Paraná) e contempla todas as possíveis contingências do sistema elétrico, introduzindo alterações no perfil da curva de demanda.

•A separação dos dados de entrada entre valores de meio de semana(terça a sexta) e finais de semana(sábado, domingo e segunda)permitiu o aumento do conjunto de treinamento e conseqüentemente uma melhoria na performance da máquina de vetor de suporte

•Observou-se que o crescimento do parâmetro C ocasionou uma diminuição do número de vetores suporte e um acréscimo no erro de prognóstico. Isso se deve ao fato de que o valor C penaliza os pontos situados fora do tubo ε, e, o seu crescimento, faz com que um número maior de vetores seja eliminado da base que dará suporte à máquina, e como conseqüência um número menor de características estáveis da série em estudo estarão presentes, aumentando o erro de prognóstico de valores futuros. As variações dos

valores de C e ε e exprimem o compromisso entre resultados com excesso de ajuste(*overfitting*) e falta de ajuste(*underfitting*).

•A função núcleo de produto interno apresentou desempenho melhor, na máquina de vetor de suporte, quando aplicada para o caso em que os perfis de demanda se apresentavam com curvas diárias semelhantes(caso do meio de semana). Para o caso de perfis de demanda de finais de semana o melhor desempenho ocorreu com função núcleo de base radial. Esta constatação mostra que para perfis de demanda de potência com maiores variações nos seus valores, é necessário utilizar uma função núcleo com grau maior de não-linearidade.

•Nos casos estudados, o erro médio absoluto do prognóstico, com RBF foi superior ao da aplicação de SVM, mostrando que valores de prognóstico melhores são obtidos a partir das características estáveis da série em estudo obtidos com os vetores suporte, e não da utilização de toda a massa de dados disponível para o estudo.

5.3 Síntese.

Neste capítulo foram apresentados os resultados do estudo de caso de utilização da máquina de vetor de suporte a uma série temporal, com valores discretos de demanda total de potência elétrica, integralizados a cada 30 minutos, circunscritos à área de concessão da COPEL - Companhia Paranaense de Energia.

Os resultados apresentados podem ser divididos em algumas categorias, de acordo com algumas características próprias de cada estudo efetuado. As principais variações estudadas estão a seguir apresentadas:

- Conjunto de treinamento com 48 valores que antecedem o valor alvo e com 8 componentes, sendo 6 que antecedem, um valor a 24h e um valor a 1 semana do valor alvo.
- Utilização de funções núcleo de produto interno, polinomial e de base radial e rede de função de base radial.
- Separação dos conjuntos de treinamento em valores similares de meio de semana(terça a sexta-feira) e finais de semana(sábado, domingo e segunda-feira).

- Variações no tamanho do conjunto de treinamento com a finalidade de aumentar o desempenho da máquina de vetor de suporte com o acréscimo de características estáveis da curva de demanda.

Com base nos estudos de caso efetuados, serão apresentados no capítulo que segue, as conclusões obtidas no estudo e de acordo com os objetivos propostos no início do trabalho.

CAPÍTULO VI – CONCLUSÕES E TRABALHOS FUTUROS.

Os objetivos apresentados para o tema do trabalho, foram atingidos, abordando-se os principais métodos de análise de séries temporais e analisando a teoria da máquina de vetor de suporte e os algoritmos de solução de problemas de regressão com a utilização de funções núcleo, podendo-se concluir sobre o estudo de caso:

1- A máquina de vetor de suporte, aplicada ao caso em estudo, foi capaz de prognosticar demandas de potência, compatíveis, com as necessidades do planejamento e operação de sistemas elétricos.
2- Os melhores resultados encontrados, na determinação do prognóstico de demanda de potência elétrica de curto prazo, foram obtidos com a separação dos valores da série temporal em dois conjuntos para processamento, sendo: O primeiro, com valores de terça a sexta-feira, conjunto de treinamento com quatro semanas, função núcleo de produto interno, apresentando erro médio de prognóstico da semana seguinte de 1,8364%, e o segundo, com valores de sábado a segunda-feira, conjunto de treinamento com cinco semanas, função núcleo de base radial, apresentando erro médio de prognóstico da semana seguinte de 2,5878%.
3- A escolha das características estáveis da série em estudo, com máquina de vetor de suporte, apresentou resultados de prognóstico, e, portanto, capacidade de generalização, melhor do que a RBF(rede de função de base radial).
4- O estudo de caso efetuado não permitiu relacionar variáveis do modelo com os dados de entrada.

As dificuldades encontradas no processamento de regressão com a máquina de vetor de suporte sugerem, a continuidade dos trabalhos de pesquisa, voltados principalmente para o estudo de relações entre os parâmetros C e ε e os vetores do conjunto de treinamento, utilizando para o processamento dos dados, ao invés da programação quadrática, o método SMO *(Sequential Minimal Optimization)*, de forma a reduzir o tempo de resposta do treinamento.

A aplicação de diversas funções núcleo, para a regressão da série de demanda de potência, permitirá verificar o comportamento comparativo entre as mesmas.

A adequação da função núcleo a ser adotada para o processamento da máquina de vetor de suporte, para o caso específico de série temporal de demanda de potência, constitui-se em um vasto campo a ser pesquisado.

O desempenho da máquina de vetor de suporte poderá ser melhorado com a incorporação de conhecimentos prévios a respeito do comportamento da série em estudo. Assim, a introdução de mecanismos capazes de efetuar o desenvolvimento da máquina de vetor de suporte, com a introdução de conhecimentos prévios, podem melhorar consideravelmente o desempenho, e é um assunto que poderá ser explorado em temas de trabalho de pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

I. Aleksander, H. Morton. *"An introduction to neural computing"*. London: Chapman and Hall, 1990.

N. Amjady. "Short-term hourly load forecasting using time-series modeling with peak load estimation capability". *IEEE Transactions on Power Systems*, vol.16, N°3, pg 498-505, 2001.

M. Aoki. *"State Space Modeling of Time Series"*. Berlin: Springer-Verlag, 1990.

D. D. Bedworth, J.E. Bailey. "*Integrated Production Control Systems"*. John Wiley and Sons, Inc., 1986.

R. Bellman. *"Dynamic programming"*. Princenton NJ: Princenton University Press, 1957.

K. P. Bennett, C. Campbell. " Support vector machines: Hype or Hallelujah?". *SIGKDD Explorations*, vol 2, issue 2, pg 1-13, 2000.

M. J. A. Berry, G. Linoff. *"Data mining techniques for marketing, sales and customer support"*. John Wiley & Sons, Inc., 1997.

T. Bollerslev. "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics.* vol. 31, pg 307-328, 1986.

B. E. Boser, I. M. Guyon, V. N. Vapnik. "A training algorithm for optimal margin classifiers". In D. Haussler, edition, *Procedings of the Fifth Annual Wokshop on Computational Learning Theory, Pittsburgh,* pg.144-152, 1992. Acesso em 20.04.04. http://citeseer.ist.psu.edu/boser92training.html.

G. E. P. Box, G. M. Jenkins. *"Time-Series Analysis, Forecasting and Control.* San Francisco, CA: Holden-Day, 1976.

L. Breiman, J. Friedman, R. Olshen, C. Stone. *"Classification and regression trees"*. New York: Chapman and Hall, 1984.

M. Brooks.. *"The Matrix Reference Manual"*. Imperial College UK, 2005. Acesso em 20/04/05. http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/.

R. G. Brown. *"Smoothing, Forecasting and Prediction of Discrete Time Series"*. Englewood Cliffs, NJ: Prentice-Hall, 1963.

J. A. Bullinaria. *"Introduction to neural networks - Radial basis function networks: Introduction"*. Course Lectures, University of Birmingham, 2003. Acesso em 20/01/04. http://www.cs.bham.ac.uk/~jxb/inn.html

D. W. Bunn. "Forecasting loads and prices in competitive power markets". *Proceedings of the IEEE*, vol. 88, N° 2, pg 163-169, 2000.

D. W. Bunn, E. D. Farmer. *"Comparative models for Electrical Load Forecasting"*. John Wiley & Sons, 1985.

C. J. C. Burges. "A tutorial on Support Vector Machines for pattern Recognition". *Data Mining and Knowledge Discovery,* vol.2, pg 121-167, 1998. Acesso em 25.03.04. http://aya.technion.ac.il/karniel/CMCC/SVM-tutorial.pdf.

W. Charytoniuk, M. –S. Chen "Very Short-Term Load Forecasting Using Artificial Neural Network". *IEEE Transactions on Power Systems,* vol.15, N°1, pg 263-268, 2000.

w. Charytoniuk, M. S. Chen, P. Van Olinda. "Nonparametric Regression Based Short-Term Load Forecasting". *IEEE Transactions on Power Systems*, vol.13, N°3, pg 725-730, 1998.

C. Chatfield. *"The analysis of time series. An introduction"*. 6˚Edition. Chapman & Hall /CRC Press, 2003.

B. –J. Chen, M. –W. Chang, C. –J. Lin. *"Load forecasting using support vector machines: A study on EUNITE(European Network on Intelligent Technologies for Smart Adaptive Systems) competition 2001"*. National Taiwan University, Acesso em 20/04/04. http://neuron.tuke.sk/competition/

P. H. Chen, C. -J. Lin, B. Schölkopf. "*A tutorial on ν-Support Vector Machines"*. Applied Stochastic Models in Bussines and Industry. 2004. Acesso em 20.10.04. www.csie.ntu.edu.tw/~cjlin/papers/nusvmtutorial.pdf.

S. Chen. "Nonlinear time series modeling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning". *Electronics Letter,* vol.32, N°2,pg 117-118, 1995.

C. Chinrungrueng, C. H. Séquin. "Optimal adaptive k-means algorithm with dynamic adjustment of learning rate". *IEEE Transactions on Neural Networks*, vol.6, pg 157-169, 1994.

C. Cortes, V. N. Vapnik. "Support-Vector Networks*"*. *Machine Learning*, vol.20, pg 273-297,1995. Acesso em 20.03.04.
http://citeseer.ist.psu.edu/cortes95supportvector.html.

R. Courant, D. Hilbert. *"Methods of mathematical physics"*. Wiley Interscience New York, 1970.

T. M. Cover. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". *IEEE Transactions on Electronic Computers,* vol. EC-14, pg.326-334, 1965.

T. M. Cover, P. E. Hart. " Nearest neighbor pattern classification". *IEEE Transactions on Information Theory,* vol. IT-13, N° 1, pg 21-27, 1967.

N. Cristianini, J. S-Taylor. *"An Introduction to Support Vector Machines and other kernel-based learning methods".* Cambridge University Press, 2002.

J. G. De Gooijer, K. Kumar. "Some recent developments in non-linear time series modeling, testing, and forecasting". *International Journal of Forecasting*, N° 8, pg 135-156, 1992.

I. Drezga, S. Rahman. "Input variable selection for ANN-based short-term load forecasting". *IEEE Transactions on Power Systems*, vol.13, N°4, pg 1238-1244, 1998.

I. Drezga, S. Rahman. "Short-term load forecasting with local ANN predictors". *IEEE Transactions on Power System,* vol.14, N°3, pg 844-850, 1999.

R. F. Engle. "Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation". *Econometrica*, vol.50, pg. 987-1008, 1982.

T. Evgeniou, M. Pontil, T. Poggio. "Regularization networks and support vector machines". Edited by Smola A.J.; Bartlett P.L.; Schölkopf B.; Schuurmans D. in *Advances in large margin classifiers.* The MIT Press, 2000. Acesso em 20/04/04. http://www.kernel-machines.org/nips98/book.html.

L. V. Fausset. *"Fundamentals of neural networks".* Prentice-Hall, Inc, 1994.

R. A. Fischer. "Theory of statistical estimation". *Proceedings of the Cambridge Philosophical Society,* vol.22, pg.700-725, 1925.

R. Fletcher. *"Practical Methods of Optimization".* John Wiley & Sons, New York, 1987.

D. M. Georgoff, R. G. Murdick. "Manager´s Guide to Forecasting". *Harvard Business Review,*pg 110-120, January-February 1986.

C. W. J. Granjer, A. P. Andersen. *"An Introduction to Bilinear Time Series Models"*. Vandenhoeck & Ruprecht, Göttingen, 1978.

S. R. Gunn. "*Support Vector Machines for Classification and Regression"*. Technical Report. University of Southampton, 1998. Acesso em 20.04.04. www.ecs.soton.ac.uk/~srg/publications/pdf/SVM.pdf.

T. Haida, S. Muto. "Regression based peak load forecasting using a transformation technique". *IEEE Transactions on Power Systems*, vol.9, N°4, pg 1788-1794, 1994.

M. E. Harmon, S. S. Harmon. *"Reinforcement learning: A tutorial"*. Wright-Patterson AFB and Wright State University., 1996. Acesso em 20/04/04. http://citeseer.nj.nec.com/harmon96reinforcement.html

S. Haykin. *"Redes neurais: Princípios e prática"*. 2˚edição Bookman Companhia Editora. Porto Alegre - Brasil , 2001.

H. S. Hippert, C. E. Pedreira, R. C. Souza. "Neural Networks for Short-Term Load Forecasting: A review and evaluation". *IEEE Transactions on Power Systems,* vol.16, N°1, pg 44-55, 2001.

B. F. Hobbs, S. Jitprapaikulsarn, S. Konda, V. Chankong, K. A. Loparo, D. J. Maratukulam. "Analysis of the Value for Unit Commitment of Improved Load Forecasts". *IEEE Transactions on Power Systems*, vol. 14, N° 4, pg 1342-1348, 1999.

J. J. Hopfield. "Neural netwoks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences USA*, vol.79, pg. 2554-2558, 1982.

C. –M. Huang, H. –T. Yang. "Evolving wavelet-based networks for short-term load forecasting". *IEE Proceedings – Generation, Transmission, Distribution*, vol. 148, N° 3, pg 222-228, 2001.

P. J. Huber. *"Robust statistics: a review"*. New York: Wiley, 1981.

T. Iizaka, T. Matsui, Y. Fukuyama. *"A Novel Daily Peak Load Forecasting Method using Analyzable Structured Neural Network"*. IEEE Transmission and Distribution Asia, 2002. Acesso em 20.03.04. http://homepage2.nifty.com/fukuyama-yoshikazu/T&D2002%20(Peak%20Load%20Forecasting).pdf.

A. K. Jain, J. Mao, K. Mohiuddin. "Artificial neural networks: A tutorial". *IEEE Computer Issue on Neural Computing*, 1996. Acesso em 20/01/04. http://citeseer.nj.nec.com/jain96artificial.html

A. K. Jain, M. N. Murty, P. J. Flynn. "Data clustering: A review". *ACM Computing Surveys*, vol.31, N°3, pg. 264-323, 1999.

T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". *Proceedings of the European Conference on Machine Learning,* pg. 137-142, 1998.

L. P. Kaelbling, M. L. Littman, A. W. Moore. *"Reinforcement learning: A survey"*. Brown and Carnegie Mello University's, 1996. Acesso em 20/04/04. http://www-2.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling.pdf

S. V. Kartalopoulos. *"Understanding neural networks and fuzzy logic"*. IEEE Press, 1996.

W. Karush. "Minima of functions of several variables with inequalities as side constraints". Master's thesis, Department of Mathematics, Chicago University, 1939.

A. Khotanzad, R. Afkhami-Rohani, T. –L. Lu, A. Abaye, M. Davis, D. J. Maratukulam. "ANNSTLF – A Neural-Network-Based-Electric-Load-Forecasting System". *IEEE Transactions on Neural Networks,* vol.8, N°4, pg 835-846, 1997.

A. Khotanzad, R. Afkhami-Rohani, D. J. Maratukulam. " ANNSTLF – Artificial Neural Network Short-Term Load Forecaster – Generation Three". *IEEE Transactions on Power Systems,* vol.13, N°4, pg 1413-1422, 1998.

A. Khotanzad, R. C. Hwang, A. Abaye, D. J. Maratukulam. "An adaptive modular artificial neural-network hourly load forecaster and its implementation at electric utilites". *IEEE Transactions on Power Systems*, vol.10, N°3, pg. 1716-1722, 1995.

V. S. Kodogiannis, E. M. Anagnostakis. "A study of advanced learning algorithms for short-term load forecasting". *Engineering Applications of Artificial Intelligence*, N°12, pg 159-173, 1999.

T. Kohonen. The self-organizing map. *Proceedings of IEEE*, volume 78, N° 9, pg 1464-1480, 1990.

B. Kröse, P. van der Smagt. *"An introduction to neural networks"*. University of Amsterdam, 1996. Acesso em 20/04/04. http://neuron.tuke.sk/math.chtf.stuba.sk/pub/vlado/NN_books_texts?Krose_Smagt_neuro-intro.pdf

H. W. Kuhn, A. W. Tucker. "Nonlinear programming". *Procedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* California University Press, pg. 481-492, 1961.

R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, V. O. Cencelli. "A neural network based technique for short-term forecasting on anomalous periods". *IEEE Transactions on Power Systems,* vol.11, N°4, pg 1749-1756, 1996.

S. Lawrence, C. L. Giles, A. C. Tsoi. *"What size neural network gives optimal generalization? Convergence properties of backpropagation"*. Technical report University of Maryland, 1996

S. Lawrence, C. L. Giles, A. C. Tsoi. Lessons in neural network training: Overfitting may be harder than expected. *Proceedings of the Fourteenth National Conference on Artificial Intelligence,* AAAI Press, pg 540-545, 1997.

S. H. Ling, H. F. Leung, H. K. Lam, Y. –S. Lee, P. K. S. Tam. "A novel Genetic-Algorithm-Based Neural Network for short-term load forecasting". I*EEE Transactions on Industrial Electronics,* vol.50, N°4, pg 793-799, 2003.

D. Lowe. "Adaptive radial basis function nonlinearities, and the problem of generalization". *First IEE International Conference on Artificial Neural Networks*, pg. 171-175, 1989.

C. N. Lu, H. T. Wu, S. Vemuri. "Neural network based short term load forecasting". *IEEE Transactions on Power Systems,* vol.8, N°1, pg 336-342, 1993.

D. G. Luenberger *" Introduction to Linear and Nonlinear Programming"*. Addison-Wesley Publishing Company, 1973.

S. Makridakis, R. M. Hogarth. "Forecasting and Planning: An evaluation". *Management Science,* vol.27, N°2, pg.115-138, 1981.

F. J. Marin, F. Garcia-Lagos, G. Joya, F. Sandoval. "Global model for short-term load forecasting using artificial neural network". *IEE Procedings*, vol. 149, N°2, pg 121-125, 2002.

F. Markowetz. "*Support Vector Machines in Bioinformatics"*. Diplomarbeit Universität Heidelberg. 2003. Acesso em 20.04.04. http://www.molgen.mpg.de/~markowet/docs/diplom.pdf.

W. S. McCulloch, W. Pitts. “A logical calculus of the ideas imminent in nervous activity”. *Bulletin of Mathematical Biophysics*, vol.5, pg.115-133, 1943.

J. M. Mendel, R. W. McLaren. “Reinforcement-learning control and pattern recognition systems”. *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*. New York: Academic Press, vol.66, pg. 287-318, 1970.

J. Mercer. “Functions of positive and negative type, and their connection with the theory of integral equations”. *Transactions of the London Philosophical Society,* 209, pg.415-446, 1909.

I. Moghram, S. Rahman. “Analysis and evaluation of five short-term load forecasting techniques*”*. *IEEE Transactions on Power Systems*, vol.4, N°4, pg 1484-1491, 1989.

H. Mori, A. Yuihara. “Deterministic Annealing Clustering for ANN-Based Short-Term Load Forecasting”. *IEEE Transactions on Power Systems*, vol.16, N°3, pg 545-551, 2001.

K. Morik. *“The representation race – preprocessing for handling time phenomena”*. Department of Computer Science, University of Dortmund, 2000. Acesso em 20/01/06. http://citeseer.ist.psu.edu/morik00representation.html.

K. –R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf. “An Introduction to Kernel-Based Learning Algorithms”. *IEEE Transactions on Neural Networks*, vol.12, N°2,pg 181-202, 2001.

K. –R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, V. Vapnik. “Predicting time series with support vector machines”. *Procedings of ICANN 97*, pg 999-1004, 1997. Acesso em 20/01/06. http://citeseer.ist.psu.edu/uller97predicting.html.

P. Murto. " Neural network models for short-term load forecasting". Master's thesis. Department of Engineering Physics and Mathematics, Helsinki University of Technology, 1998. Acesso em 20.03.04. http://www.sal.hut.fi/Publications/pdf-files/tmur98.pdf

D. F. Nicholls, A. R. Pagan. "Varying coefficient regression". In *Handbook of statistics*, vol. 5 ed. E.J. Hannan, P.R. Krishnaiah, M.M. Rao Amsterdam, 1985.

D. F. Nicholls, B. G. Quinn. *"Random Coefficient Autoregressive Models: An Introduction"*. Springer-Verlag, New York, 1982.

N. Nikolaev. *"Neural Networks - Radial basis function networks"*. Tutorial Course University of London, 2003. Acesso em 20/01/04. http://homepages.gold.ac.uk/nikolaev/311rbf.html

F. J. Nogales, J. Contreras, A. J. Conejo, R. Espínola. "Forecasting next-day electricity prices by time series models". *IEEE Transactions on Power Systems,* vol. 17, N°2, pg 342-348, 2002.

E. E. Osuna, R. Freund, F. Girosi. "*Support Vector Machines: Training and Applications"*. MIT, 1997. Acesso em 20.04.04. http://citeseer.ist.psu.edu/osuna97support.html.

J. H. Park, Y. M. Park,K. Y. Lee. "Composite modeling for adaptive short-term load forecasting". *IEEE Transactions on Power Systems*, vol. 6, N° 2, pg 450-457, 1991.

D. B. Parker. "Learning logic". Technical Report, Center of Computational Research in Economics and Management Science, MIT, 1982.

M. P. Perrone. "General averaging results for convex optimization". *Proceedings of the Connectionist Models Summer School,* pg. 364-371, 1994.

R. Pfeifer. *"Neural Networks, SS2000"*. Notas de aulas, University of Zürich, 2000. Acesso em 20/01/04. http://www.ifi.unizh.ch/groups/ailab/teaching/NN2000/

A. Piras, A. Germond, B. Buchenel, K. Imhof, Y. Jaccard. "Heterogeneous artificial neural network for short load forecasting". *IEEE Transactions on Power Systems*, vol.11, N°1, pg 397-402, 1996.

T. Poggio, F. Girosi. "Networks for approximation and learning". *Procedings of the IEEE,* vol.78, pg. 1481-1497, 1990.

M.; J. D. Powell. "Radial basis functions for multivariable interpolation: A review". *IMA Conference on Algorithms for the Approximation of Functions and Data*, pg.143-167, England, 1985.

M. J. D. Powell. "Radial basis function approximations to polynomials". *Numerical analysis Proceedings*. pg. 223-241, Dundee, UK, 1988.

M. B. Priestley. "State-dependent models: a general approach to non-linear time series analysis". *Journal of Time Series Analysis*, vol.1, pg 47-71, 1980.
J. R. Quinlan. "Induction of decision trees". *Machine Learning*, vol.1, pg. 81-106, 1986.

J. R. Quinlan. "Improved use of continuous attributes in C4.5". *Journal of Artificial Intelligence Research,* N° 4, pg 77-90, 1996.

F. Rosenblatt. "The Perceptron: A probabilistic model for information storage and organization in the brain". *Psychological Review*, vol.65, pg.386-408, 1958.

D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning representation of back-propagation errors". *Nature*, vol.323, pg.533-536, London, 1986a.

D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning internal representations by error propagation". *In D. E. Rumelhart, J. L. McCleland, edition, vol.1, Chapter 8, Cambridge, MA: MIT Press*, 1986b.

T. P. Runarsson, S. Sigurdsson. *"Chapter 3 - Kernel – induced feature space"*. Notas de aulas da Sigillum Universitatis Islandiae, 2003. http://cerium.raunvis.hi.is/~tpr/courseware/svm/notes.

T.P.Runarsson, S. Sigurdsson. *" Chapter 6 – Support vector machines"*. Notas de aulas da Sigillum Universitatis Islandiae, 2003a. http://cerium.raunvis.hi.is/~tpr/courseware/svm/notes.

S. Rüping. *"SVM kernels for time series analysis"*. Department of Computer Science, University of Dortmund, 2001. Acesso em 25/07/05. http://citeseer.ist.psu.edu/483548.html.

S. Rüping, K. Morik *"Support vector machines and learning about time"*. Department of Computer Science, University of Dortmund, 2003. Acesso em 20/01/06. http://citeseer.ist.psu.edu/566731.html.

L. M. Saini, M. K. Soni. "Artificial neural network based peak load forecasting using Levenberg-Marquardt and quasi-Newton methods". *IEE Procedings*, vol.149, N°5, pg 578-584, 2002.

L. M. Saini, M. K. Soni. "Artificial Neural Network-Based Peak Load Forecasting Using Conjugate Gradient Methods". *IEEE Transactions on Power Systems*, vol.17, N°3, pg 907-912, 2002.

B. Schölkopf. *"Support Vector Learning"*. Master's thesis, Technische Universität Berlin, 1997. Acesso em 20.04.04 http://www.svms.org/learnability/Scho97.pdf.

B. Schölkopf, A. J. Smola, K. –R. Müller. "Nonlinear component analysis as a kernel eigenvalue problem". *Neural computation*, vol.10, pg.1299-1319, 1998.

T. Senjyu, H. Takara, K. Uezato, T. Funabashi. "One-hour-ahead load forecasting using neural network". *IEEE Transactions on Power Systems*, vol. 17, N° 1,pg 113-118, 2002.

A. Sfetsos. "Short-term load forecasting with a hybrid clustering algorithm". *IEE Procedings,* vol.150, N°3, pg 257-262, 2003.

P. K. Simpson. "*Artificial Neural Networks – Paradigms, applications and implementations"* Pergamon Press, 1990.

L. Sjoberg. "Aided and unaided decision making: Improved Intuitive Judgment". *Journal of Forecasting*, vol.1, pg.349-363, 1982.

A. J. Smola. *"Regression estimation with Support Vector Learning Machines"*. Master 's thesis. Physik Department, Technische Universität München, 1996. Acesso em 20.01.06. http://citeseer.ist.psu.edu/smola96regression.html.

A. J. Smola. *"Learning with Kernels"*. PhD thesis, Technische Universität Berlin, 1998. GMD Research Series. Acesso em 20.04.04. http://www.bi.fraunhofer.de/publications/research/1998/025/Text.pdf.

A. J. Smola, P.L. Bartlett, B. Schölkopf, D. Schuurmans. " Introduction to large margin classifiers". In A. J. Smola, P.L. Bartlett, B. Schölkopf, D. Schuurmans, editors, *Advances in large margin classifiers.* pp 1-29, Cambridge, 2000, MIT Press. http://www.kernel-machines.org/nips98/book.html.

A. J. Smola, B. Schölkopf. "*A tutorial on Support Vector Regression"*. NeuroCOLT2 Technical Report Series, 1998. Acesso em 15.05.04. http://citeseer.ist.psu.edu/smola98tutorial.html.

A. J. Smola, B. Schölkopf, K. –R. Müller. *"General cost functions for support vector regression"*. Acesso em 20.01.06. http://citeseer.ist.psu.edu/smola98general.html.

J. S. Taylor, N. Cristianini *"Kernel Methods for Pattern Analysis"*. Cambridge University Press, 2004.

A. N. Tikhonov. "On solving incorrectly posed problems and method of regularization". *Doklady Akademii Nauk,* vol.151, pg.501-504, USSR, 1963.

H. Tong, K. S. Lim. "Threshold autoregression, limit cycles and cyclical data". *Journal of the Royal Statistical Society,* vol.B42, pg 245-292, 1980.

M. Unser. "Splines: A perfect fit for signal/image processing". I*EEE Signal Processing Magazine*, vol.16, N°6,pg.22-38,1999.

V. N. Vapnik *" The nature of statistical learning theory"*. 2°edition, Springer - Verlag New York, 2000.

C. J. C. H. Watkins. "*Learning from delayed rewards"*. PhD. Thesis, University of Cambridge, England, 1989.

E. W. Weisstein. "*Mathworld"*.. Virginia University, 2004. Acesso em 20.04.04. http://mathworld.wolfram.com/

P. J. Werbos. "*Beyond regression: New tools for prediction and analysis in behavioral sciences"*. Ph.D. thesis, Harvard University, Cambridge, MA, 1974.

D. Wettschereck, T. Dietterich. "Improving the performance of radial basis function networks by learning center locations". *Advances in Neural Information Processing Systems,* vol.4, pg. 1133-1140, Morgan Kaufmann, 1992.

D. J. Willshaw, C. von der Malsburg. “How patterned neural connections can be set up by self-organization”. *Proceedings of the Royal Society of London*, vol.194, pg.431-445, 1976.

Wolfram *Research. ”Neural networks documentation-A short tutorial”*. 2004. Acesso 05/01/04. http://documents.wolfram.com/applications/neuralnetworks/NeuralNetworkTheory/2.1.0.html

H. Yang. “Margin variations in support vector regression for the stock market prediction”. Master’s thesis, Department of Computer Science and Enginnering, Chinese University of Hong Kong, 2003. Acesso em 20.04.04. http://www.cse.cuhk.edu.hk/~hqyang/papers/thesis.pdf