UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

Dissertação de Mestrado

Sistema Inteligente Híbrido Intercomunicativo para Detecção de Perdas Comerciais

Lucas Teles de Faria

Orientador: Prof. Dr. Antonio Padilha Feltrin

Co-Orientador: Prof. Dr. Carlos Roberto Minussi

Ilha Solteira – SP

Março de 2012

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Sistema Inteligente Híbrido Intercomunicativo para Detecção de Perdas Comerciais

Lucas Teles de Faria

Orientador: Prof. Dr. Antonio Padilha Feltrin

Co-Orientador: Prof. Dr. Carlos Roberto Minussi

Dissertação apresentada à Faculdade de Engenharia – UNESP – Campus de Ilha Solteira, para obtenção do título de Mestre em Engenharia Elétrica. Área de Conhecimento: Automação.

Ilha Solteira – SP

Março de 2012

FICHA CATALOGRÁFICA

Elaborada pela Seção Técnica de Aquisição e Tratamento da Informação
Serviço Técnico de Biblioteca e Documentação da UNESP - Ilha Solteira.

Faria, Lucas Teles de.
F224s Sistema inteligente híbrido intercomunicativo para detecção de perdas comerciais / Lucas Teles de Faria. -- Ilha Solteira : [s.n.], 2012
112 f. : il.

Dissertação (mestrado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2012

Orientador: Antonio Padilha Feltrin
Co-orientador: Carlos Roberto Minussi
Inclui bibliografia

1. Energia elétrica. 2. Perdas de energia elétrica. 3. Sistemas de energia elétrica. 4. Energia elétrica - Distribuição. 5. Perdas comerciais. 6. Redes neurais (computação). 7. Lógica difusa. 8. Lógica *Fuzzy*.

unesp UNIVERSIDADE ESTADUAL PAULISTA
CAMPUS DE ILHA SOLTEIRA
FACULDADE DE ENGENHARIA DE ILHA SOLTEIRA

CERTIFICADO DE APROVAÇÃO

TÍTULO: Sistema Inteligente Híbrido Intercomunicativo para Detecção de Perdas Comerciais

AUTOR: LUCAS TELES DE FARIA
ORIENTADOR: Prof. Dr. ANTONIO PADILHA FELTRIN
CO-ORIENTADOR: Prof. Dr. CARLOS ROBERTO MINUSSI

Aprovado como parte das exigências para obtenção do Título de Mestre em Engenharia Elétrica, Área: AUTOMAÇÃO, pela Comissão Examinadora:

Prof. Dr. ANTONIO PADILHA FELTRIN
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Prof. Dr. JOSE ROBERTO SANCHES MANTOVANI
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Prof. Dr. ANDRE NUNES DE SOUZA
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Bauru

Data da realização: 01 de março de 2012.

DEDICATÓRIA

In memoriam de meu pai, José Maria Candido de Faria, que me ensinou o valor do trabalho e da honestidade.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela oportunidade de trabalhar em uma instituição pública de excelência.

Agradeço à minha família por sempre me dar forças e me incentivar nesta etapa da minha vida.

Agradecimentos especiais ao professor Antonio Padilha Feltrin, por ter aceitado orientar-me, pela paciência e pela dedicação a esse trabalho. Também agradeço ao co-orientador professor Carlos Roberto Minussi pelas observações valorosas.

Agradeço aos colegas do Laboratório de Planejamento de Sistemas de Energia Elétrica (LaPSEE) pelo companheirismo. Em especial aos meus grandes amigos: Erick Somocurcio, Marlon Borges, Miguel Paredes, Renan Felix, Joel Trujillo, Rafael Mineiros, Raiane Piacente, João Sousa, Thays Abreu, Ana Paula Sakai, Érica Ribeiro, André Bíscaro, Augusto César Medina, Ana Cláudia Barros e Eliane Souza. A todos os professores e funcionários do Departamento de Engenharia Elétrica da Faculdade de Engenharia de Ilha Solteira.

Agradeço aos meus amigos que me apoiaram em todas as dificuldades que passei durante a construção deste trabalho.

Agradeço a UNESP, ao Departamento de Engenharia Elétrica da FEIS, pela estrutura oferecida para o desenvolvimento deste trabalho e ao CNPq pelo apoio financeiro.

RESUMO

As perdas de energia elétrica por fraudes, ligações clandestinas ou erro na medição são denominadas Perdas Não-Técnicas ou Perdas Comerciais e seu combate tem sido prioridade quer por empresas concessionárias de distribuição de energia elétrica, quer por órgãos reguladores. Nesse contexto, neste trabalho, implementa-se computacionalmente um sistema inteligente híbrido intercomunicativo específico que se baseia no emprego de diferentes técnicas oriundas da área de sistemas inteligentes tais como redes neurais e lógica *fuzzy* em módulos independentes e que se comunicam entre si. O sistema é baseado em três pilares: extração automática de conhecimento a partir da base de dados da concessionária, incorporação na metodologia do conhecimento e experiência de especialistas e, em último, consultas na base de dados por características específicas de cada cliente. A metodologia utiliza simultaneamente inúmeros dados reais de entrada de natureza diversa e combina várias técnicas a fim de verificar o risco percentual de cada cliente de possuir alguma anomalia que implique em perda comercial. Além dos dados cadastrais e do histórico de consumo mensal dos clientes, comumente utilizados pelos trabalhos orientados à detecção de perdas comerciais, a metodologia proposta utilizou também dados adicionais tais como a lista de nomes e de atividades suspeitas. A utilização de dados adicionais possibilitou uma melhoria na detecção de clientes anômalos, grande parte dos quais seriam possivelmente considerados normais pelos trabalhos da literatura avaliada. Neste trabalho, pretende-se detectar as perdas comerciais de maneira mais rápida e precisa possível. Investigações adicionais devem ser feitas posteriormente para encontrar quais são as causas que culminaram em altas perdas comerciais em um alimentador ou em uma região específica. Este trabalho abrange os clientes cadastrados na concessionária, principalmente, clientes residenciais, comerciais e industriais os quais concentram a parcela majoritária das perdas comerciais.

Palavras-chave: Perda de energia elétrica. Sistema de Distribuição. Perdas Comerciais. Perdas Não-Técnicas. Detecção de Perdas Comerciais. Aplicação de redes neurais. Aplicação de lógica *fuzzy*.

ABSTRACT

Electrical energy losses due to theft, fraud or error in the measurement are called Non-Technical Losses and their reduction has been a priority utilities power and by regulators. In this context, this paper presents the computational implementation of an intelligent hybrid system that combines techniques such as neural networks and fuzzy logic. The system is based on three pillars: knowledge extraction from the database utility, incorporating the methodology of knowledge and experience of experts and queries the database for specific features of each client. The methodology uses several simultaneous input of diverse nature and combines several techniques to verify the percentage risk of each customer to have some problem to configure non-technical losses as fraud, defective in the measurement system. In this paper, the main objective is to locate the focus of the problem more quickly. Further investigations should be made later to find what are the causes that resulted in high non-technical losses in a feeder or in a specific region. This work covers the registered customers at utilities power, especially residential, commercial and industrial which concentrate the majority stake of non-technical losses.

Keywords: Loss of electrical power. Distribution System. Commercial Losses. Non-Technical Losses. Detection of commercial losses. Application of neural networks. Application of fuzzy logic.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AGs	Algoritmos Genéticos
ANEEL	Agência Nacional de Energia Elétrica
CI	Consumo Irregular
CV	Coeficiente de Variação
DW	*Data Warehousing*
GSA	Grau de Suspeita do Agrupamento ou *Gravitational Search Algorithm*
HS	*Harmony Search*
IA	Inteligência Artificial
ICMS	Imposto sobre Circulação de Mercadorias e Serviços
IPCA	Índice de Perdas Comerciais do Alimentador
ISA	Índice de Suspeita da Área
ISC	Índice de Suspeita do Cliente
ISE	Índice de Suspeita do Especialista
KDD	*Knowledge Discovery in Database*
KNN	*K-Nearest Neighbor*
LAS	Lista de Atividades Suspeitas
LNS	Lista de Nomes Suspeitos
MCP	McCulloch-Pitts
MD	Mineração de Dados
MLP	*Multilayer Perceptron*
NA	Número de Aumentos em relação ao regime
NQ	Número de Quedas em relação ao regime
NQAR	Número de Queda e Aumento de Regime
NR	Número de Regimes
NRAbFM	Número de Regimes Abaixo da Faixa Média
NRAbFRI	Número de Regimes Abaixo da Faixa do Regime Inicial
NRAcFM	Número de Regimes Acima da Faixa Média
NRAcFRI	Número de Regimes Acima da Faixa do Regime Inicial
NRFM	Número de Regimes na Faixa Média
NRFRI	Número de Regimes na Faixa do Regime Inicial
NTL	*Non-Technical Losses*
NZ	Número de Zeros
PA	Porcentagem de Aumento em relação ao regime
PC(s)	Perda(s) Comercial(is)

PG	Perda Global
PMC	*Perceptron* Multicamadas
PNT(s)	Perda(s) Não-Técnica(s)
PSO	*Particle Swarm Optimization*
PQ	Porcentagem de Quedas em relação ao regime
PQAR	Percentual de Queda e Aumento de Regime
PT	Perda Técnica
PTQA	Percentual de Tempo em regimes de Queda e Aumento
PTRA	Porcentagem de Tempo no Regime Aumento
PTRI	Porcentagem de Tempo no Regime Inicial
PTRQ	Porcentagem de Tempo no Regime de Queda
R	Religamento
RBF	*Radial Basis Function*
RFI	Regimes na Faixa do Regime Inicial
RFM	Regimes na Faixa Média
RNA	Rede Neural Artificial
RPC	Risco de Perda Comercial
SBC	Sistema Baseado em Conhecimento
SDEE	Sistema de Distribuição de Energia Elétrica
SI	Sistemas Inteligentes
SIF	Sistema de Inferência *Fuzzy*
SIH	Sistema Inteligente Híbrido
SOM	*Self Organization Maps*
SVM	*Support Vector Machines*
TC	Transformador de Corrente
TP	Transformador de Potencial
TSK	Takagi-Sugeno-Kang
UCs	Unidades Consumidoras

LISTA DE SÍMBOLOS

η	Taxa de aprendizado da rede neural
α	Constante de *Momentum*
x_i	Entradas da rede neural
y_i	Saídas da rede neural
θ	Limiar ou peso *bias*
$\boldsymbol{w}_i$	Matriz de pesos sinápticos
E	Erro quadrático médio
d_j	Saída desejada em relação ao neurônio *j* da camada de saída da rede neural
n	Número de épocas de treinamento da rede neural
Ω_j	Conjuntos dos neurônios vizinhos do neurônio *j*
$\mu(.)$	Função de pertinência *fuzzy*
c	Confiabilidade Negativa
e	Especificidade
u	Potencial de Ativação ou Campo Local Induzido
$g(.)$	Função de Ativação
$D^{(k)}$	Coeficiente de disparo da regra *fuzzy k*

SUMÁRIO

1 ESCOPO DO TRABALHO

Este capítulo de apresentação visa fornecer uma introdução abrangente às perdas comerciais no Sistema de Distribuição de Energia Elétrica (SDEE), as principais causas do seu surgimento e quais são as inúmeras implicações que elas ocasionam. Ao final do capítulo, tem-se a estruturação do texto.

1.1 INTRODUÇÃO

As perdas comerciais no SDEE, também denominadas perdas não-técnicas, são um dos principais problemas enfrentados pelas empresas concessionárias, principalmente, em países subdesenvolvidos ou emergentes, como o Brasil, por exemplo.

No Brasil (e em outros países em desenvolvimento), as perdas comerciais são altas por inúmeros fatores de cunho socioeconômico e cultural tais como: desemprego, baixa renda, falta de habitação, infraestrutura insuficiente, preço elevado da energia e de acessórios de ligação e impunidade em relação à corrupção e à fraude (BASTOS, 2009).

A minimização das perdas de energia elétrica por furtos e fraudes tem sido prioridade nas empresas concessionárias, bem como nos órgãos reguladores, tanto pelo seu crescimento nos últimos anos quanto pela sua atual dimensão. Fraudes e furtos de energia elétrica realizados por consumidores residenciais, comerciais e até industriais constituem o montante majoritário das perdas comerciais (DANTAS, 2006).

Observa-se que grande parte dessas perdas tem origem em questões de cunho social. O Estado deve e pode estabelecer políticas para resolver ou pelo menos minorar tais questões. Tem-se a lei 10.438 de 2002 que estabelece condicionantes, metas e fontes de recursos para universalização, possibilitando o acesso ao serviço de fornecimento de energia elétrica sem qualquer ônus para todas as UCs (Unidades Consumidoras) com carga instalada de até 50 kW e estabelece critérios para a classificação dos consumidores na subclasse residencial de baixa renda. O objetivo é dotar mecanismos para a manutenção do fornecimento de energia elétrica de forma regular, através de subsídios nas tarifas para as classes menos favorecidas (PENIN, 2008).

Os fatores que geram as perdas comerciais são os mais diversos possíveis e podem ser internos ou externos às concessionárias de energia. Destacam-se:

- aspectos relacionados ao mau gerenciamento dos procedimentos internos das concessionárias como erro do leiturista, erro no cadastramento dos consumidores ou da iluminação pública, defeito em relés fotoelétricos, etc;
- defeito nos medidores de energia causados, por exemplo, por obsolescência ou por surtos na rede de energia;
- fraudes no medidor de energia engendradas pelos clientes ou por ex-funcionários da companhia de energia tais como: auto-religações, alteração das características dos equipamentos de medição e inserção de desvios antes da medição;
- ligações clandestinas;
- aspectos sociais como pobreza, impunidade, corrupção, etc.

A fim de quantificar as perdas comerciais totais, as concessionárias idealmente deveriam realizar inspeções nos medidores de todas as UCs, quantificar todas as ligações clandestinas, além de executar o levantamento da iluminação pública de toda área selecionada. No entanto, devido ao grande número de UCs e ao alto custo das inspeções tal prática é totalmente inviável. Por isso, é preciso que o processo de identificação do perfil do consumidor seja automático (CABRAL; PINTO; GONTIJO, 2004).

Nesse contexto, as concessionárias comumente optam por selecionar de maneira superficial e semi-automática, através de planilhas eletrônicas, alguns clientes cujo histórico de consumo mensal exiba algum comportamento suspeito. Tais clientes serão visitados pelas equipes de inspeção. Os parâmetros geralmente avaliados para cada cliente são a curva de consumo mensal, constituída pela conexão dos pontos de consumo mensais de energia em kWh, adicionado aos dados cadastrais. No entanto, a curva de consumo mensal sofre diversas influências; logo, a análise de clientes a serem inspecionados não deve se basear unicamente pela análise da mesma. Outro parâmetro analisado são os dados cadastrais. Todavia, muitas vezes tais dados contêm informações errôneas ou por estarem desatualizadas ou mesmo por erro no momento da realização do cadastro do cliente.

Portanto, os principais parâmetros analisados para indicar se um determinado cliente é suspeito de ocasionar algum tipo de perda comercial estão longe de serem totalmente confiáveis. Isso confere um alto grau de complexidade ao problema de localização de perdas comerciais no SDEE. Além das informações a respeito de cada consumidor não serem em quantidade suficiente, as existentes não são, na maioria dos casos, totalmente íntegras. Adicionalmente, o perfil dos clientes fraudadores é dinâmico, isto é, os fraudadores buscam mecanismos cada vez mais sofisticados para furtar energia a fim de não serem descobertos.

À medida que se aproxima das cargas aumentam-se as incertezas, porque as informações reduzem em quantidade e em precisão. A única medição que se tem a respeito da maioria dos consumidores – que pertencem ao grupo B[1] – é o consumo mensal de energia em kWh. Além do consumo, têm-se os dados cadastrais, os dados colhidos nas inspeções e a perda comercial por alimentador obtida pela diferença entre a energia distribuída e a energia faturada, descontadas as perdas técnicas.

É difícil identificar precisamente o cliente que possui algum tipo de perda comercial. Mais difícil ainda é identificar o motivo desta perda; se é decorrente de uma fraude ou de um medidor avariado, por exemplo.

1.2 OBJETIVOS

A principal maneira de atuar no combate às perdas comerciais é a realização de inspeções nos pontos de consumo (PERIM; DIAS; VAREJÃO, 2007). Para aperfeiçoar o procedimento das inspeções é preciso identificar previamente os consumidores que apresentam comportamento suspeito. Essa identificação é feita por um Sistema Baseado em Conhecimento (SBC) que automatiza o processo de seleção de clientes a serem inspecionados. Esse sistema engloba técnicas de Inteligência Artificial (IA) tais como Redes Neurais Artificiais (RNAs) e lógica *fuzzy* além de buscas ao banco de dados e o aproveitamento do conhecimento e experiência de especialistas. O objetivo é agilizar o processo de seleção de consumidores a serem inspecionados e aumentar a taxa de acertos de clientes com perdas comerciais de maneira a minorar o custo com inspeções.

Não faz parte do escopo deste trabalho detectar perdas comerciais em áreas em que o fraudador não é cliente da concessionária a exemplo de favelas e invasões. Em tais áreas a fonte de perdas é facilmente detectável; no entanto, o processo de regularização já envolve questões sociais e legais.

1.3 POR QUE REDUZIR AS PERDAS COMERCIAIS?

O desvio de energia elétrica ou fraude é um problema internacional que prejudica a sociedade e acarreta aumento na tarifa de fornecimento e injustiça social. As ligações irregulares na rede de distribuição de energia representam grande risco para a segurança pública, uma vez que modificam as características da rede e podem causar sérios acidentes.

[1] As UCs pertencentes ao Grupo B possuem tensão e potência instalada inferior a 2,3 kV e 75 kW, respectivamente.

Adicionalmente, observa-se que o mercado energético brasileiro tem passado por significativas mudanças no que se refere à liberalização e à competitividade visando incentivar investimentos privados. Em termos gerais, a intenção das mudanças é criar um ambiente competitivo, através de instrumentos de organização tais como a desverticalização, limites de poder de mercado e privatização. Nesse contexto, a localização e o combate as perdas comerciais no SDEE tornam-se cruciais para a sobrevivência das concessionárias em um mercado cada vez mais competitivo e dinâmico.

Dentre os principais danos à sociedade, ocasionados pela existência das perdas comerciais destacados em (DANTAS, 2006), têm-se:

- *Insegurança.* Em geral, as ligações clandestinas são realizadas sem rigor técnico e sem um estudo prévio da rede elétrica local. As consequências disso são acidentes graves, redução do nível de tensão local e aumento das interrupções no fornecimento de energia para clientes normais que compartilham a mesma rede.
- *Concorrência desleal.* O furto de energia permite reduzir ilicitamente os custos de atividades comerciais ou industriais, gerando uma concorrência desleal em relação às empresas honestas. Essas empresas são, dessa forma, estimuladas a também aderir a essa prática fraudulenta por uma questão de sobrevivência no mercado.
- *Aumento tarifário.* As empresas concessionárias de energia elétrica são concessões de serviço público regidas por política tarifária. Para determinar o percentual de reajuste, a ANEEL considera a variação de custos que as distribuidoras tiveram nos últimos doze meses. Na conta de consumo de energia elétrica de cada cliente, há uma parcela referente às perdas comerciais que é medida em valores monetários e não em número de clientes fraudulentos. Nessa perspectiva, o consumidor honesto irá pagar pelo consumo fraudado por meio da elevação da tarifa, o que representa uma grande injustiça social.
- *Desperdício de energia.* Consumidores fraudadores ou ligados clandestinamente não pagam a energia elétrica que consomem e, por isso, não têm hábitos de racionalização, o que ocasiona grande desperdício de energia. É comum, nesses casos, lâmpadas acesas durante todo o dia ou aparelhos de ar-condicionado ligados ininterruptamente.
- *Proliferação do roubo de energia elétrica.* A impunidade leva à proliferação de bandidos que oferecem uma forma ilícita de economia através da redução ou mesmo da anulação da tarifa de energia.

- *Não arrecadação de impostos*. A arrecadação de vários impostos é reduzida por fraudes e ligações clandestinas. Dentre esses, destaca-se o ICMS que é proporcional à venda de energia elétrica. Tais recursos não arrecadados pelo Estado deixam de ser aplicados em benefício da própria sociedade.
- *Degradação ambiental*. Grande parcela da energia elétrica consumida é gerada em usinas termoelétricas. Essas utilizam combustíveis fósseis cuja queima libera gases que ocasionam: poluição do ar, chuva ácida, redução da camada de ozônio, aquecimento por efeito estufa, etc. Logo, o aumento da eficiência energética e redução das perdas elétricas contribuem para menor geração de energia e consequente redução da emissão de CO_2.

Em suma, a necessidade de se buscar o maior retorno financeiro ao selecionar os locais a serem inspecionados é de interesse das concessionárias, da ANEEL e de toda a sociedade.

1.4 ORGANIZAÇÃO DO TEXTO

Neste trabalho, apresenta-se a proposta e o desenvolvimento de uma nova metodologia que incorpora outros aspectos à localização de perdas comerciais no SDEE. O texto está dividido em sete capítulos:

- no capítulo 2, abordam-se conceitos teóricos que fundamentam o estudo e cuja revisão visa auxiliar na compreensão do leitor acerca do problema tratado e da solução proposta para o mesmo. Apresenta-se também uma revisão acerca da literatura consultada sobre o tema;
- o capítulo 3 é uma revisão abrangente acerca dos paradigmas das Redes Neurais Artificias (RNAs), com ênfase às redes *Perceptron* Multicamadas (PMC) e ao Mapa Auto-Organizável de Kohonen (SOM, do inglês, *Self Organization Maps*). Tais redes compõem a metodologia proposta neste trabalho;
- o capítulo 4 aborda de maneira abrangente os principais tópicos da Lógica *Fuzzy* e os Sistemas de Inferência Fuzzy (SIFs) Mamdani e Takagi-Sugeno-Kang (TSK);
- o capítulo 5 descreve cada um dos módulos que compõem a metodologia proposta para detecção de perdas comerciais no SDEE. Os capítulos anteriores, principalmente, os capítulos 3 e 4 fornecem o embasamento para o pleno entendimento da metodologia;

- ❖ o capítulo 6 apresenta o emprego do método desenvolvido em clientes residenciais, comerciais e industriais. Os testes foram realizados com dados reais de clientes pertencentes a uma concessionária de uma cidade de porte médio do interior do Estado de São Paulo;
- ❖ o capítulo 7 apresenta as conclusões e as sugestões para trabalhos futuros.

2 CONCEITUAÇÃO E ESTIMATIVAS DAS PERDAS COMERCIAIS NO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Este capítulo objetiva fornecer uma revisão da literatura acerca das principais metodologias para tratar o problema de localização e de estratificação das perdas comerciais em uma área pertencente ao SDEE (Sistema de Distribuição de Energia Elétrica). As perdas elétricas são divididas em perdas técnicas e em perdas comerciais ou perdas não-técnicas. Existe uma vasta literatura que aborda as perdas técnicas. Os primeiros trabalhos a esse respeito datam das primeiras décadas do século XX. Em contrapartida, o tema relativo às perdas comerciais passou a ser pesquisado mais fortemente apenas nos últimos dez anos, sendo a referência Jiang (2002) um dos marcos iniciais.

2.1 NATUREZA DAS PERDAS ELÉTRICAS

No contexto da distribuição de energia, definem-se perdas como a diferença entre a quantidade de energia comprada e distribuída e a quantidade de energia paga pelos consumidores. As perdas são agrupadas conforme sua origem em duas categorias:

- Perdas Técnicas: quantidade de energia consumida por efeito Joule durante o processo de transporte de energia, ocasionado, por exemplo, pelas resistências internas dos condutores e equipamentos de transmissão. Tais perdas podem ser reduzidas através de investimentos na construção de novas redes, da correta manutenção e melhoria dos equipamentos e da melhoria dos processos de distribuição de energia elétrica (COMETTI; VAREJÃO, 2005).
- Perdas Comerciais ou Perdas Não-Técnicas: quantidade de energia comprada pela concessionária e não faturado aos seus consumidores, descontadas as perdas técnicas. As causas mais comuns são: ligações irregulares, erros de medição, defeitos em equipamentos, fraudes, etc.

A detecção exata de todos os pontos em que há perdas comerciais é sobremaneira custoso devido principalmente à grande e crescente quantidade de consumidores, à grande variedade de perfis de consumo de energia elétrica, ao custo elevado das inspeções, à informação insuficiente que se tem disponível principalmente para os clientes do grupo B, a grande diferença na quantidade de clientes normais e de clientes com algum tipo de perda comercial (em média, para cada dez clientes normais há um cliente com alguma anormalidade que eleva as perdas comerciais) e à natureza dinâmica das formas de realizar fraudes que podem ser

feitas desde alterações simples como desvios na instalação elétrica até mecanismos mais complexos como violação de equipamentos de medição, por exemplo. É preciso extrair informações úteis a partir de uma grande base de dados. No entanto, essa é uma tarefa árdua além de exigir muito processamento, frequentemente, os dados estão repetidos, incompletos ou inconsistentes. Como não há uma única metodologia para resolver esse problema, é necessário trabalhar em conjunto com a concessionária para se detalhar melhor os objetivos e entender todas as restrições do problema.

A fraude está diretamente relacionada à renda, aspectos culturais e condições socioeconômicas da população, isto é, variáveis relacionadas à localização geográfica. A tendência de fraudes está também relacionada com a vizinhança, pois um vizinho pode induzir o outro a cometer a fraude. Segundo Dantas (2006), para a maioria das concessionárias, fraudes e defeitos na medição são as principais causas das perdas comerciais de energia elétrica. O Quadro 1 contém as irregularidades mais comuns por fraude e por defeito no medidor em UCs (Unidades Consumidoras) e que por serem de origem externa à concessionária são as mais difíceis de serem localizadas.

Quadro 1: Principais irregularidades das UCs cadastradas.

Irregularidades por Fraude	Irregularidades por Defeito na Medição
Ponte no bloco de terminais	Medidor com disco parado
Ligação direta ou auto-religação	Medidor com defeito
Ligação invertida	Constante de medição errada
Circuito de potencial interrompido	Consumidor não implantado
Desvio aparente antes do medidor	Ligação executada com erro
Desvio embutido na parede	TC (Transformador de Corrente) danificado
Medidor avariado	
Medidor com selo violado	

Fonte: Bastos (2011).

2.1.1 Perdas Técnicas

Os cálculos das perdas técnicas são comumente executados em cada segmento do sistema elétrico de modo a conferir maior exatidão nos resultados. As perdas técnicas em sistemas elétricos são calculadas comumente em três segmentos: alta, média e baixa tensão. As perdas no segmento de alta tensão são encontradas pela diferença de medição nas subestações.

No sistema de distribuição de média tensão, as perdas são calculadas por meio de fluxo de carga. Já a metodologia para o cálculo das perdas técnicas em sistemas de baixa tensão varia entre as concessionárias. No entanto, a maior parte delas agregam os componentes da rede de distribuição por tipo (transformador, rede secundária, ramais de serviço e medidores) e realizam os cálculos baseados em curvas de cargas típicas de consumidores, totalmente independente dos dados de faturamento (DANTAS, 2006).

Mensurar precisamente as perdas técnicas em um SDEE é um problema complexo devido principalmente à grande quantidade de elementos que constituem o sistema, à grande quantidade de dados necessários, ao caráter aleatório das cargas elétricas e ao seu contínuo processo de expansão.

Conhecidas as perdas globais e as perdas técnicas, as perdas comerciais são obtidas pela diferença entre ambas. Logo, erros advindos desses cálculos e estimativas estão incorporados ao valor estimado para a perda comercial. Essa é a prática usual do setor elétrico brasileiro.

2.2 PREVENÇÃO E COMBATE ÀS PERDAS COMERCIAIS

O caminho mais utilizado para quantificação das perdas comerciais ainda tem sido através do cálculo *a priori* das perdas técnicas (OLIVEIRA, 2009). Podem-se relacionar diversas ações que visam à redução das perdas comerciais:

- identificar as localidades com maior risco de fraudes e desenvolver campanhas com o apoio dos líderes locais;
- regularização dos clientes clandestinos;
- adaptação da rede de distribuição de energia elétrica com a instalação de cabos antifurto, redes compactas ou multiplexadas de média tensão;
- implementar políticas de facilitação de quitação de débitos e políticas de cortes;
- ações no sentido de redução do consumo de energia elétrica como: troca de geladeiras; substituição de lâmpadas incandescentes por lâmpadas compactas fluorescentes, instalação gratuita de medidores de energia, instalação de aquecedores solar em substituição aos chuveiros elétricos, etc.

O Quadro 2 é uma compilação das causas principais que influem diretamente no acréscimo das perdas comerciais. Para cada causa têm-se as respectivas medidas preventivas que poderiam ser adotadas pelas concessionárias como forma de minimização das perdas comerciais. Observa-se que as causas estão agrupadas em internas e externas às concessionárias.

A maior parte das causas internas tais como erros do leiturista e no cadastramento de consumidores são mais facilmente combatidas e sua existência está relacionada ao mau gerenciamento dos procedimentos internos da empresa concessionária. Logo, perdas comerciais de origem interna são inaceitáveis e devem ser prontamente combatidas através, por exemplo, das medidas de prevenção e combate do Quadro 2. Em contrapartida, todas as perdas comerciais de origem externa à concessionária tais como os inúmeros tipos de fraudes do Quadro 1 são mais difíceis de serem localizadas e requerem o desenvolvimento de ferramentas sofisticadas.

As informações colhidas durante as inspeções em algumas UCs, em geral previamente selecionadas, são fundamentais para quantificação e qualificação das perdas comerciais em uma dada região. Em geral, as inspeções são motivadas por campanhas de combate às perdas comerciais, suspeitas de fraude, manutenção, etc. Dentre as principais estratégias para localização de UCs com perdas comerciais estão: a varredura, denúncias, análise dos consumidores mais suspeitos e programas computacionais que implementam ferramentas sofisticadas como aquelas pertencentes à área de SI (Sistemas Inteligentes), por exemplo. O sucesso de cada uma das estratégias é mensurado através da taxa de sucesso ou taxa de acerto apresentada na expressão (2.1).

$$Taxa\_Acerto = \frac{\sum Fraudes + \sum Anomalias}{\sum Inspeções} \tag{2.1}$$

2.2.1 Varredura

A estratégia de varredura é uma das mais simples e menos eficientes para a localização de clientes com perdas comerciais. Apesar disso, ainda é largamente empregada por algumas concessionárias. Nessa estratégia, identificam-se quais são os alimentadores da rede de distribuição com as maiores perdas comerciais através do cálculo das perdas globais descontadas as perdas técnicas. Em seguida, inspecionam-se todos os consumidores que são supridos por tal alimentador.

Por ter de inspecionar todos os consumidores de uma região pertencentes a um alimentador com altas perdas comerciais, essa estratégia tem alto custo financeiro e consome muito tempo das equipes de inspeção.

Tais características limitam a execução da operação de varredura a regiões nas quais a quantidade de fraudes é suficientemente grande para compensar os altos custos com a

operação. Ademais, a operação de varredura é facilmente identificada pelos consumidores que podem mascarar suas fraudes momentos antes da realização da inspeção.

Quadro 2: Síntese das principais causas e medidas de prevenção e combate às perdas comerciais no SDEE.

Concessionária	Origem	Medidas de Prevenção e Combate
Clientes	*Interna:* erro do leiturista; defeito ou obsolescência do medidor; erro na ligação; especificação inadequada da medição; engano no cadastro de clientes; funcionários desonestos, etc.	Acompanhamento do processo de ligação e faturamento; atualização cadastral; treinamento do pessoal que especifica o sistema de medição; inspeção dos medidores, etc.
	Externa: fraudes como desvio embutido antes da medição; circuito de potencial interrompido, medidor avariado, ponte nos terminais, etc.	Inspeções; acompanhamento do histórico de consumo; instalação de medição fiscal; desenvolvimento de programas e ferramentas que incorporem IA (Inteligência Artificial) que auxiliem na identificação de desvios de consumo, melhorando a taxa de acerto das inspeções; lacres no sistema de medição; uso de redes antifurto, etc.
Não Clientes	Ex-clientes desligados que se auto-religam.	Adequação tarifária; políticas de subsídios; facilidades para negociação de débitos; inspeção das unidades desligadas; lacres no sistema de medição; etc.
	Conexões ilegais: invasões, loteamentos irregulares, etc.	Mapeamento de zonas de risco; realização de políticas sociais em conjunto com a prefeitura, polícia e justiça; opção pela rede antifurto; etc.

Fonte: Bastos (2011).

A taxa de sucesso da operação de varredura é baixa e oscila entre 10% a 12% (CABRAL; PINTO; GONTIJO, 2004; COMETTI; VAREJÃO, 2005; PERIM; DIAS; VAREJÃO, 2007).

2.2.2 Denúncias

Inspeções são realizadas sempre que ocorrem denúncias. Embora a estratégia de denúncias tenha alto índice de acerto, o número de denúncias é muito pequeno em relação à

quantidade de inspeções que podem ser realizadas ao longo do tempo. A taxa de sucesso das denúncias é de 22% aproximadamente (COMETTI; VAREJÃO, 2005).

2.2.3 Análise dos dados dos consumidores

A estratégia de análise dos dados consiste em aplicar regras heurísticas simples para varrer a base de dados e selecionar clientes com histórico de consumo atípico os quais serão inspecionados. A análise de dados é uma tarefa árdua que requer muito tempo e esforço por parte dos especialistas, os quais precisam analisar grandes bases de dados cadastrais para identificar características suspeitas de cada consumidor.

Uma regra comum é a *Regra do Consumo Zero* que seleciona clientes com consumo inferior a um valor limite durante um determinado período de tempo com exceção de residências desocupadas ou de áreas de veraneio que possuem índice de consumo abaixo do padrão (PERIM; DIAS; VAREJÃO, 2007).

A análise de dados utiliza regras heurísticas baseadas na experiência do especialista. No entanto, uma regra, após ser aplicada sucessivamente ao longo do tempo, tem sua eficácia reduzida. As regras são dinâmicas. É preciso adaptá-las. Além disso, tais regras cobrem um pequeno número de consumidores.

A taxa de sucesso dessa estratégia é moderada. A AES Eletropaulo obteve um índice de acerto de 19,8% nos anos de 2005, 2006 e 2007 (FERREIRA, 2008).

A eficácia dessa estratégia depende exclusivamente do conhecimento e experiência do especialista. No entanto, abordagens exclusivamente dependentes do conhecimento do especialista podem impedir que padrões escondidos nos dados sejam encontrados de forma inteligente, uma vez que o especialista não tem condições de imaginar todas as possíveis relações e associações existentes em um grande volume de dados. Por isso, faz-se necessária a utilização de técnicas de análise dirigidas por computador que possibilitem a extração automática (ou semi-automática) de novos conhecimentos a partir de um grande repositório de dados (REZENDE, 2005).

2.2.4 Programas computacionais que incorporam técnicas de sistemas inteligente s

A seleção dos consumidores a serem inspecionados é realizada comumente através da busca exaustiva (ou varredura) ou por meio da análise da base de dados em planilhas eletrônicas pelos especialistas. No entanto, se a seleção dos consumidores com maiores indícios de irregularidade fosse feita de maneira automática, através de um programa específico,

as inspeções seriam mais baratas e restritas a pequenas regiões ou pontos bem determinados, aumentando a possibilidade de flagrantes de irregularidade (COMETTI; VAREJÃO, 2005).

Os programas computacionais para detecção de perdas comerciais possuem comumente sistemas classificadores que implementam técnicas de extração automática de conhecimento a partir de dados como a mineração de dados, por exemplo. A partir das fontes de dados disponíveis, tais sistemas computacionais classificadores são capazes de decidir com base nas características atuais de um determinado cliente se ele deve ou não ser inspecionado. Fundamental para essa abordagem é a base de dados com o histórico de inspeções que permite gerar exemplos de clientes normais e anormais (fraudadores, por exemplo) para serem usados para treinamento e teste do sistema classificador. Aliado a isso, tem-se a aquisição e incorporação do conhecimento dos especialistas ao sistema desenvolvido. Uma condição fundamental para o desenvolvimento desses sistemas é torná-lo flexível para que os usuários possam facilmente adaptá-lo a fim de aplicar novas regras heurísticas.

O Quadro 3 contém um resumo das principais características das estratégias adotadas pelas empresas concessionárias de energia para localização de UCs com perdas comerciais.

2.3 SISTEMAS INTELIGENTES ABORDADOS NO PROBLEMA DE PERDAS COMERCIAIS

Os trabalhos que abordam a detecção de perdas comerciais utilizam inúmeras técnicas de inteligência artificial. Destaque para as técnicas de mineração de dados, técnicas estatísticas e redes neurais como métodos com maior êxito nesse campo de pesquisa (GUERRERO; LEÓN; BISCARRI, 2010).

As técnicas de computação *soft* tentam implementar algumas características humanas como: problemas de decisão, aprendizado e reconhecimento. Conjuntos Rústicos é uma dessas técnicas com recursos para manipular incerteza e imprecisão nos dados (CABRAL; PINTO; GONTIJO, 2004).

As técnicas de aprendizado de máquina tais como Redes Neurais Artificiais (RNAs), Algoritmos Genéticos (AGs), o algoritmo indutor de árvore de decisão C4.5 e o SVM (*Support Vector Machine*) trabalham com problemas mal definidos. Nesses, não é possível analisar todas as alternativas possíveis devido à grande quantidade de possibilidades de encaminhamento até a solução, à imprecisão e à escassez dos dados de entrada.

Quadro 3: Resumo das características das principais estratégias para localização de clientes com perdas comerciais.

Estratégia	$\frac{Retorno\ Financeiro}{Custo\ de\ Execução}$	Tempo de Planejamento	Área de Abrangência	Taxa de Sucesso
Varredura	Baixo	Moderado	Por Área	10% a 12%
Denúncias	Alto	Baixo	Pontual	≈ 22%
Análise dos Dados	Moderado	Alto	Pontual	> 20%
Software Específico	Alto	Alto	Pontual	> 20%

Fonte: Informações da pesquisa do autor.

2.4 LOCALIZAÇÃO E/OU EXPLICAÇÃO DAS CAUSAS DAS PERDAS COMERCIAIS

Os trabalhos da literatura avaliada a respeito das perdas comerciais de maneira geral são orientados à localização ou à estratificação das perdas comerciais em uma região. As metodologias orientadas à localização (na qual se enquadra este trabalho) não se preocupam em explicar o que ocasiona as perdas comerciais, mas sim em localizá-las de maneira mais precisa e ágil possível. Em contrapartida, as metodologias orientadas às causas procuram fornecer um panorama a respeito das causas e quais são os tipos de perdas comerciais que incidem em uma região específica.

Metodologias orientadas à localização são pontuais e objetivam encontrar com a maior taxa de sucesso possível as UCs com algum tipo de irregularidade por fraudes ou por medidor avariado que impliquem em perdas comerciais, por exemplo. Para isso, elas se valem de quaisquer informações que de alguma forma aumente a chance de sucesso. Os dados de entrada mais usuais são: histórico de consumo mensal, dados cadastrais, atividade econômica, comentários dos inspetores e/ou leituristas, perdas globais no alimentador, denúncias, vulnerabilidade da região em que se encontra a UC, potência contratada, etc. Essas metodologias utilizam comumente ferramentas pertencentes a área de Sistemas Inteligentes (SI) tais como: redes neurais, lógica *fuzzy*, aprendizado de máquina, mineração de dados e de texto, etc.

Em contrapartida, o objetivo primeiro das metodologias orientadas às causas é fornecer um panorama geral a respeito das UCs de uma região. Elas efetuam a estratificação e a quantificação das perdas comerciais por classes de consumidores e por tipo de irregularidade. Em geral, utilizam apenas os dados colhidos em inspeções realizadas em UCs previamente selecionadas e a partir dessas amostras efetuam a extrapolação através de técnicas estatísticas

e/ou probabilísticas. As principais ferramentas comumente utilizadas são técnicas estatísticas de amostragem, redes Bayesianas, *Naive Bayes*, etc.

Essas duas abordagens são complementares. As metodologias orientadas à localização possuem caráter determinístico e pontual e as metodologias orientadas às causas têm caráter estatístico e/ou probabilístico.

2.5 ABORDAGENS UTILIZADAS NA DETECÇÃO DE CLIENTES COM PERDAS COMERCIAIS

Em Jiang (2002) tem-se uma abordagem baseada em múltiplos classificadores e coeficientes *wavelets* para identificação de fraudes em clientes cujo consumo de energia é medido em intervalos de quinze minutos.

Rauber, Drago e Varejão (2005) mostra que a extração de novas características a partir da série temporal de consumo mensal melhora o desempenho do classificador que determinará a presença ou a ausência de perdas comerciais. A partir dessa constatação, os autores realizaram a análise comparativa entre três metodologias para extração de atributos: Coeficientes de Fourier, Coeficientes *Wavelets* e Regressão Polinomial. O primeiro método de extração gerou as características mais relevantes, seguido pelos coeficientes *Wavelets* e em último a Regressão Polinomial.

De acordo com Ramos (2012) *et al.*, no entanto, o foco não é na extração de características, mas na seleção daquelas que são mais representativas, que possuem maior poder preditivo e que serão as entradas das técnicas de reconhecimento de padrões também denominados sistemas classificadores. Para tal, os autores efetuam a comparação entre três algoritmos evolutivos: PSO (*Particle Swarm Optimization*), GSA (*Gravitational Seach Algorithm*) e HS (*Harmoy Seach*). Os testes a partir de oito características de clientes comerciais e industriais demonstram a melhor performance do sistema classificador na detecção de clientes suspeitos quando antes é realizada a seleção das características mais relevantes.

Perim, Dias e Varejão (2007), desenvolve um Sistema Baseado em Conhecimento (SBC) composto por simples regras independentes escritas manualmente na linguagem XML baseadas no conhecimento do especialista. Os dados de entrada avaliados são alguns atributos estatísticos como a média e o consumo nulo que são extraídos a partir do histórico de consumo mensal de cada consumidor.

No trabalho de Guerrero, León e Biscarri (2010), objetiva-se explicar o comportamento muitas vezes anômalo e suspeito da curva de consumo mensal como sendo conse-

quência das características de consumo do cliente como, por exemplo, de sua atividade econômica. Os dados de entrada da metodologia são as informações textuais e em linguagem natural contida na base de dados das concessionárias. São informações sobre a documentação, comentários dos inspetores e informações adicionais sobre a instalação elétrica dos consumidores. A metodologia é uma combinação das técnicas de mineração de texto, redes neurais, técnicas estatísticas e conhecimento especialista a fim de extrair conhecimento a partir de informação não estruturada. Dessa forma, pretende-se aumentar a eficiência das campanhas contra perdas comerciais através da redução de falsos positivos.

Cabral, Pinto e Gontijo (2004) utilizam conjuntos rústicos (*Rough Sets*) como técnica para redução do número de atributos usados na indução de um sistema de regras de decisão para detecção de fraudes em consumidores de energia elétrica.

Ferreira (2008) avalia uma combinação entre diferentes ferramentas de classificação e bases de dados para melhorar o índice de acerto das inspeções. As bases de dados são compostas pelo histórico de consumo mensal e por dados cadastrais dos clientes. Para tal, utilizam-se ferramentas de aprendizados de máquina tais como: redes neurais, o algoritmo indutor de árvore de decisão C4.5, aprendizado de máquina e *Naive Bayes*. O trabalho efetua uma análise comparativa entre esses quatro classificadores a fim de detectar aquele que apresenta melhor predição em termos dos critérios de confiabilidade negativa e de especificidade extraídos a partir da matriz de confusão. O algoritmo C4.5 obteve os melhores resultados.

Cometti e Varejão (2005) desenvolvem um sistema para identificação de clientes com perdas comerciais e, na mesma linha que Ferreira (2008), efetua uma análise comparativa entre as seguintes técnicas: *Naive Bayes*, Redes Bayesianas, redes neurais, algoritmo C4.5 e *k* vizinhos mais próximos (KNN, do inglês, *K-Nearest Neighbor*).

Nizar, Dong e Zhang (2008) buscam detectar anormalidades que podem ser devidas às perdas comerciais através da extração de regras em dados cadastrais. Utiliza técnicas de mineração de dados, agrupamento e técnicas de classificação.

Smith (2004) faz uma abordagem qualitativa de grande relevância a qual aborda aspectos sociais em todo o mundo. Essa referência estimou o roubo de eletricidade em 102 países de 1980 até o ano 2000. Ela reconhece a característica complexa e multifacetária do problema de perdas comerciais e o relaciona a aspectos governamentais tais como: instabilidade política, baixa efetividade governamental e alto nível de corrupção.

Dantas (2006) efetua a quantificação e a estratificação das perdas comerciais por classe e por faixa de consumo. Esse trabalho possui embasamento estatístico e baseia-se na

amostragem aleatória de dados levantados em campo. As UCs são agrupadas em quatro categorias: irregular com fraude, irregular com defeito, irregular sem perda e normal.

Em último, Bastos (2009) realiza um diagnóstico ou prospecção das perdas comerciais permitindo identificar suas diversas origens e quantificar as parcelas relacionando-as às suas causas dentre os consumidores e classes de consumo em várias regiões. Utiliza-se para tal as redes Bayesianas[2].

2.6 CONSIDERAÇÕES FINAIS

Neste capítulo fez-se uma explanação abrangente a respeito da natureza do problema de detecção e de estratificação das perdas comerciais no SDEE. Destacaram-se as principais formas de prevenção e combate às perdas comerciais.

Nos dois capítulos subsequentes, abordam-se duas ferramentas computacionais que serão a base da metodologia desenvolvida neste trabalho para detecção das perdas comerciais.

[2] Modelos de gráficos probabilísticos adequados para trabalhar com incertezas.

3 REDES NEURAIS ARTIFICIAIS: *PERCEPTRON* MULTICAMADAS E MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN

Este capítulo visa apresentar uma introdução abrangente e sucinta a respeito das principais características, mecanismo de funcionamento e potencialidades das Redes Neurais Artificiais (RNAs). Ênfase maior é dada à rede *Perceptron* Multicamadas (PMC) e ao Mapa Auto-Organizável de Kohonen (*Self Organization Maps* – SOM). Essas redes neurais compõem a metodologia proposta para prospecção de perdas comerciais e serão utilizadas, respectivamente, para classificação e agrupamento.

3.1 INTRODUÇÃO ÀS REDES NEURAIS ARTIFICIAIS

As redes neurais estão inseridas em uma área conhecida como sistemas inteligentes[3] conexionistas. São modelos matemáticos inspirados nas estruturas neurais biológicas e que possuem capacidade computacional de aprendizagem e de generalização.

As redes neurais são capazes de extrair e de manter o conhecimento (baseado em informações) e podem ser definidas como um conjunto de unidades de processamento elementares que são interligados por um número de interconexões (sinapses artificiais) que são representadas por matrizes de pesos sinápticos.

As características mais atrativas das RNAs englobam a sua elevada capacidade para mapear sistemas não-lineares aprendendo comportamentos envolvidos a partir de informações obtidas.

A aprendizagem da rede está relacionada à sua capacidade de ajustar seus parâmetros internos (pesos sinápticos) como consequência de sua repetida interação com o meio externo. A capacidade de generalização é devido à sua capacidade de responder coerentemente para dados não submetidos a ela durante a fase de treinamento.

Os problemas em que as redes neurais são comumente empregadas envolvem aproximação funcional, predição, classificação, categorização e otimização. Os problemas de aproximação funcional ou regressão são caracterizados por interpolação, isto é, os dados fornecidos estão dentro de determinados limites em que uma função qualquer é definida e o modelo neural é ajustado para fornecer uma boa aproximação aos mesmos. O problema de predição visa à predição de estados seguintes de um determinado sistema baseado nos seus estados

[3] Os sistemas considerados inteligentes podem ser definidos como aqueles que tentam explorar em suas estruturas pelo menos um dos aspectos relacionado ao comportamento dos seres humanos tais como aprendizado, inferência, raciocínio, evolução e adaptação (SILVA, 2010).

anteriores. Em problemas de classificação, o objetivo é identificar cada classe entre um conjunto de classes desconhecidas. Em último, os problemas de categorização envolvem a descoberta de características estatisticamente relevantes de um determinado conjunto de dados e como estes podem ser agrupados em classes ou em *clusters*. Neste trabalho, as redes serão usadas para resolução de problemas de classificação e de categorização.

A seguir, enumeram-se as características mais relevantes das RNAs baseado em Haykin (1999):

- *Adaptação por experiência.* Ajuste dos parâmetros internos da rede (pesos sinápticos) a partir da apresentação sucessivas de exemplos (padrões, amostras, medidas) relacionados ao comportamento do processo, possibilitando aquisição de conhecimento empírico.
- *Capacidade de aprendizado.* A partir de um método de aprendizado, a rede é capaz de extrair o relacionamento que há entre as variáveis envolvidas no sistema em estudo.
- *Habilidade de generalização.* Depois de terminado o processo de treinamento, a rede é capaz de generalizar o conhecimento adquirido, isto é, estimar soluções até o momento desconhecidas.
- *Organização de dados.* Com base em características intrínsecas de determinado conjunto de informações acerca de um processo, a rede é capaz de se organizar internamente para agrupar padrões que possuem particularidades em comum.
- *Tolerância a falhas.* Devido ao elevado nível de interconexões neurais, a rede é capaz de fornecer respostas coerentes mesmo quando parte de sua estrutura interna está sensivelmente corrompida.
- *Armazenamento distribuído.* O conhecimento sobre o comportamento de determinado processo está distribuído no interior da arquitetura neural entre as inúmeras conexões sinápticas artificiais, fato esse que confere maior robustez à arquitetura no caso, por exemplo, de neurônios inoperantes.
- *Facilidade de prototipagem.* A implementação da maioria das arquiteturas neurais pode ser implementada tanto em *hardware* ou em *software*, pois, após o treinamento da rede, os resultados são obtidos a partir de operações matemáticas elementares.

Em resumo, as RNAs oferecem a possibilidade de aprendizado a partir de dados, modelagem empírica do comportamento humano, técnicas de aproximação universal, memórias associativas, paralelismo massivo e robustez. No entanto, apesar de todas essas vanta-

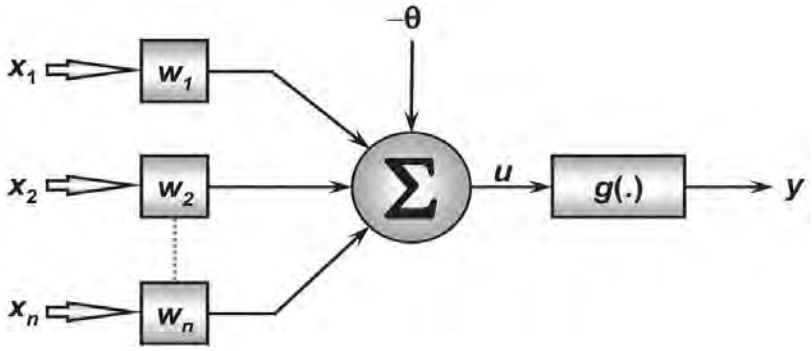
gens, as redes neurais exigem um longo tempo de treinamento, não possuem um mecanismo explicativo, nem um mecanismo automático e eficiente para auxiliar o desenvolvedor na tarefa de projetá-la. O modelo que a rede aprende está implícito, isto é, a rede fornece saídas coerentes, mas não explica as relações entre as variáveis que governam o sistema em estudo.

Nesse contexto, segundo Haykin (1999), as formas possíveis de representação desde as entradas até os parâmetros internos da rede são muito diversificadas, fato esse que torna o desenvolvimento de uma solução satisfatória utilizando uma rede neural um desafio real de projeto.

3.1.1 Modelo de um neurônio artificial

O processamento da informação é feito por estruturas neurais artificiais nas quais o armazenamento e o processamento da informação são executados de maneira paralela e distribuída por elementos processadores simples como o neurônio de MCP (McCulloch-Pitt) esboçado na Figura 1. Cada elemento processador corresponde a um neurônio artificial. Apesar de ser formada por um conjunto de elementos processadores elementares, uma RNA como um todo tem capacidade computacional para resolução de problemas complexos.

Figura 1: Modelo MCP não-linear de um neurônio artificial.



Fonte: McCulloch e Pitts (1943).

As entradas do neurônio artificial da Figura 1 correspondem ao vetor de entrada $\boldsymbol{x} = [x_1\ x_2\ \ldots\ x_n]^T$ de dimensão *n*. Para cada entrada x_i do neurônio há um respectivo peso sináptico w_i. A soma das entradas x_i ponderadas pelos pesos sinápticos w_i adicionado ao *bias* θ é chamado de saída linear, combinação linear, campo local induzido ou potencial de ativação *u*, sendo $u = \sum_{i=1}^{n} w_i x_i - \theta$. θ é o limiar de ativação ou *bias* do neurônio. O limiar de ativação especifica o patamar a partir do qual o resultado produzido pelo combinador linear gera um valor de disparo em direção à saída do neurônio. A saída *y* do neurônio é dita saída de ativação e é obtida pela função *g(.)* aplicada à saída linear *u*, sendo $y = g(u)$. A função

g(.) é denominada função de ativação. Ela pode assumir várias formas e geralmente é não-linear. A função de ativação limita a saída do neurônio.

O número de entradas e de saídas de uma RNA depende da dimensionalidade dos dados, enquanto que o número de neurônios nas camadas intermediárias depende da complexidade do problema.

3.1.2 Funções de ativação típicas

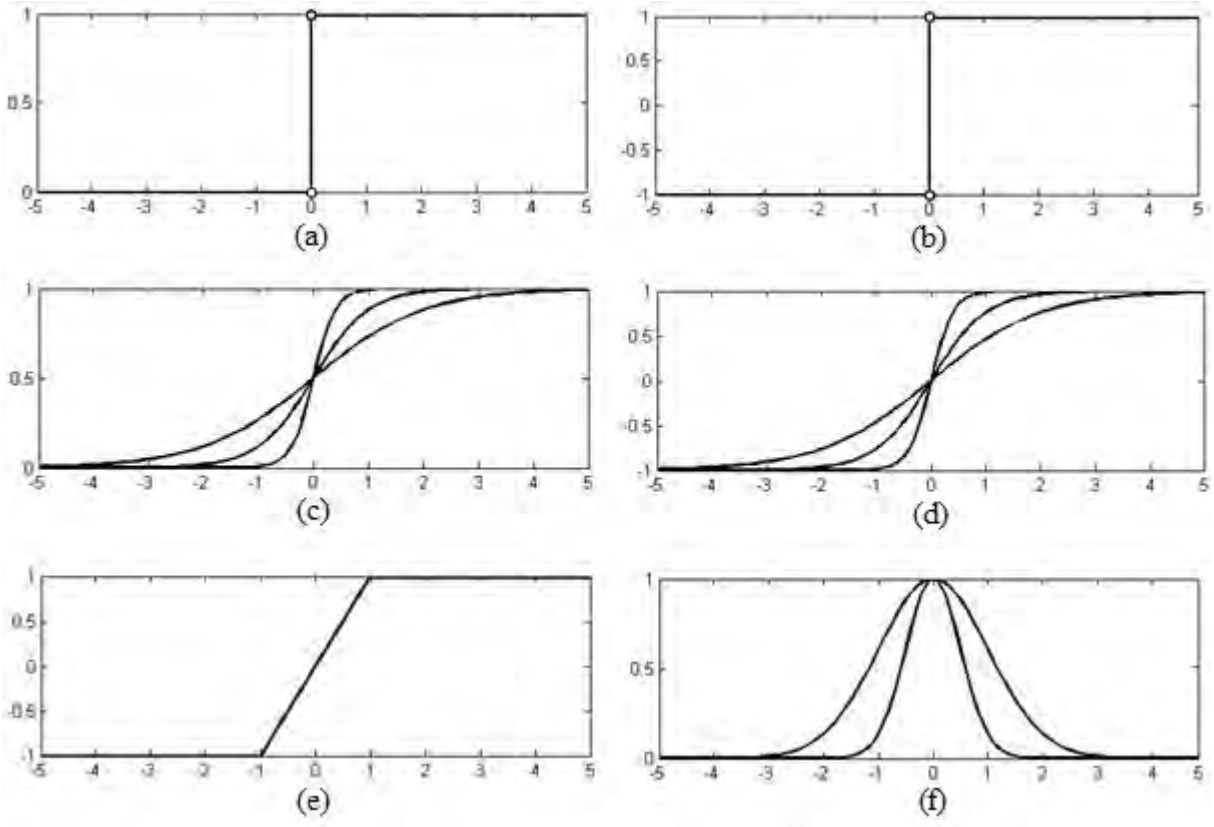
A Figura 2 é um esboço das funções de ativação mais comuns na literatura de redes neurais. As expressões analíticas dessas funções constam no Quadro 4. As funções de ativação degrau e degrau bipolar – Figuras 2 (a) e (b), respectivamente – são não-diferenciáveis, por isso, tais funções tem seu uso limitado. As demais funções da Figura 2 são contínuas e diferenciáveis. As funções sigmoidais – logística e tangente hiperbólica, Figuras 2 (c) e (d), respectivamente – são de longe as mais usadas como funções de ativação em redes neurais. O parâmetro β dessas funções é uma constante real que regula a inclinação das mesmas em relação ao seu ponto de inflexão. Na Figura 2, as funções sigmoidais estão esboçadas para β = 1, 2 e 5. Observa-se que quando $\beta \to \infty$ a função logística será similar ao degrau e a função tangente hiperbólica será similar ao degrau bipolar. Para a rampa simétrica os valores retornados são iguais aos próprios valores dos potenciais de ativação quando estes estão definidos no intervalo $[-a, a]$, limitando-os aos valores limites em caso contrário. No esboço dessa função – Figura 2 (e) – adotou-se $a = 1$. A expressão analítica da função gaussiana – Figura 2 (f) – possui dois parâmetros: sendo *c* um parâmetro que define o centro da função gaussiana e σ define o desvio padrão associado à mesma, ou seja, o quão disperso é a curva em relação ao seu centro. No esboço da Figura 2 (f), adotou-se $c = 1$ para σ = 1 e 2. Quanto maior o valor de σ menos dispersa se torna a curva.

3.1.3 Principais arquiteturas

A arquitetura de uma rede neural define a forma como os seus neurônios estão arranjados ou dispostos uns em relação aos outros. Basicamente, uma RNA pode ser dividida em três tipos de camadas:

- Camada de entrada: camada responsável pelo recebimento das informações ou dados de entrada.

Figura 2: Funções de ativação típicas. (a) Degrau. (b) Degrau bipolar. (c) Logística. (d) Tangente hiperbólica. (e) Rampa simétrica. (f) Gaussiana.



Fonte: Informações da pesquisa do autor.

Quadro 4: Expressões das funções de ativação típicas da Figura 2.

Função de Ativação	Expressão Matemática
Degrau	$g(u) = \begin{cases} 1, se\ u \geq 0 \\ 0, se\ u < 0 \end{cases}$
Degrau Bipolar	$g(u) = \begin{cases} 1, se\ u \geq 0 \\ -1\ se\ u < 0 \end{cases}$
Logística	$g(u) = \dfrac{1}{1 + e^{-\beta u}}$
Tangente Hiperbólica	$g(u) = \dfrac{1 - e^{-\beta u}}{1 + e^{-\beta u}}$
Rampa Simétrica	$g(u) = \begin{cases} a, se\ u > a \\ u, se - a \leq u \leq a \\ -a, se\ u < a \end{cases}$
Gaussiana	$g(u) = e^{-\frac{(u-c)^2}{2\sigma^2}}$

Fonte: Informações da pesquisa do autor.

- Camada escondida, intermediária, oculta ou invisível: camada responsável pela extração das características associadas ao processo ou sistema a ser inferido. Quase todo o processamento interno da rede é realizado nessa camada. O número de camadas escondidas e a quantidade de neurônios nessa camada dependem principalmente da complexidade do problema a ser mapeado pela rede, assim como da quantidade e da qualidade dos dados de entrada disponíveis.
- Camada de saída: responsável pela produção e apresentação dos resultados finais da rede.

As principais arquiteturas de RNAs considerando a disposição dos neurônios, a forma de interligação entre eles e a constituição de suas camadas são:

- Redes *feedforward* (alimentação à frente): nesse tipo de rede, o fluxo de informação segue sempre em uma única direção (rede unidirecional), ou seja, da camada de entrada em direção à camada de saída. Dentre as principais redes cuja arquitetura é do tipo *feedforward* tem-se o *Perceptron* Multicamadas (*Multilayer Perceptron* – MLP) e as redes de base radial (*radial basis function* – RBF).
- Redes Recorrentes: são redes em que as saídas dos neurônios são realimentadas como sinais de entrada para outros neurônios. A realimentação confere a tais redes a capacidade de processamento dinâmico podendo ser utilizadas em sistemas variantes no tempo como previsão de séries temporais, por exemplo. A principal rede que possui realimentação é denominada rede Hopfield. Devido à realimentação, as redes com essa arquitetura produzem saídas levando-se em consideração os valores das saídas anteriores recentes.
- Redes Reticuladas: as redes reticuladas consideram a disposição espacial dos neurônios visando à extração de características, isto é, a localização espacial dos neurônios está relacionada com o processo de ajuste de pesos e limiares. São usadas principalmente em problemas de agrupamento de dados. O mapa auto-organizável de Kohonen é o principal representante desse tipo de arquitetura.

3.1.4 Treinamento e generalização

O treinamento de uma RNA consiste no ajuste dos pesos sinápticos w_i e do *bias* θ através de iterações sucessivas de maneira que a superfície de decisão presente no espaço de dimensão *n* das variáveis de entrada atenda aos requisitos de parada do algoritmo.

Em geral, o conjunto total das amostras disponíveis sobre o comportamento do sistema é dividido em dois subconjuntos: o subconjunto de treinamento e o subconjunto de teste. O subconjunto de treinamento é formado aleatoriamente por 60 a 90% das amostras do conjunto total e será usado na fase de aprendizado da rede. O subconjunto de teste é composto por 10% a 40% das amostras do conjunto total e será usado para verificar se a generalização da rede já atingiu um estágio satisfatório permitindo a validação da topologia arbitrada.

As redes caracterizam-se pelo aprendizado através de exemplos. Para um determinado conjunto de dados, o algoritmo de aprendizado é responsável pela adaptação dos parâmetros livres da rede (pesos sinápticos), de maneira que em um número finito de iterações o algoritmo convirja para uma solução ótima local ou global. O critério de convergência depende do algoritmo adotado e do paradigma de aprendizagem. Alguns exemplos de critérios de convergência são: minimização de uma função objetivo, variação do erro de saída ou mesmo a variação das magnitudes dos vetores de pesos da rede.

Os algoritmos de aprendizado podem ser classificados em três paradigmas distintos: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço.

O aprendizado supervisionado caracteriza-se pela existência de um "professor" externo à rede que tem a função de monitorar a saída y_i da rede para cada vetor de entrada x_i. O conjunto de treinamento é composto por pares de entrada e saída (x_i, y_i^d), onde x_i é o vetor de entrada e y_i^d é a saída desejada para a entrada x_i. O ajuste dos pesos da rede é feito de tal forma que a saída y_i para a entrada x_i se aproxime de y_i^d dentro dos limites de tolerância pré-estabelecidos. Conforme Silva, Spatti e Flauzinho (2010), o treinamento supervisionado é um caso típico de inferência indutiva pura, isto é, parte-se de casos particulares e a partir desses deseja-se generalizar para todos os casos pertencentes ao sistema investigado.

O aprendizado não-supervisionado caracteriza-se pela não existência de saídas desejadas para as entradas x_i. O conjunto de treinamento é formado apenas pelos vetores de entrada x_i. Não há supervisor externo e o ajuste de pesos é obtido apenas com base nos valores dos dados de entrada. O aprendizado não-supervisionado é comumente usado em problemas de categorização.

O aprendizado por reforço é um paradigma intermediário entre o aprendizado supervisionado e o não-supervisionado. O conjunto de treinamento é formado somente por entradas, no entanto, há um elemento externo que, em vez de retornar o erro de saída da rede, retorna um sinal de reforço ou de penalidade associado à última ação da rede. Caso tal ação tenha ocasionado queda no desempenho, ela será penalizada, tendo menor chance de ocorrer

futuramente. Caso contrário, se a ação resultou em uma melhora no desempenho, ela será reforçada aumentando a probabilidade de ocorrência da mesma no futuro.

3.1.5 Tipos de aprendizagem

Existem basicamente dois tipos de aprendizagem: aprendizagem usando lote de padrões (*off-line*) e aprendizagem usando padrão por padrão (*on-line*).

Na aprendizagem *off-line*, os ajustes nos vetores de pesos da rede e em seus limiares são efetivados somente após a apresentação de todo o conjunto de treinamento. Dessa forma, o ajuste leva em consideração o desvio total das amostras de treinamento em relação aos respectivos valores desejados para a saída da rede. Assim, as redes que utilizam aprendizagem usando lote de padrões precisam de uma época de treinamento para realizar um passo de ajuste em seus pesos e limiares.

Em contrapartida, na aprendizagem *on-line*, os ajustes nos pesos e limiares da rede são efetuados após a apresentação de cada amostra de treinamento. Como os padrões são apresentados um por vez, as ações de ajuste dos pesos e limiares são bem localizadas e pontuais nesse tipo de aprendizagem. A rede começará a fornecer respostas mais precisas somente após a apresentação de um número significativo de amostras.

3.2 REDES *PERCEPTRON* MULTICAMADAS

A rede neural *Perceptron* Multicamadas (PMC) é uma das redes mais populares e mais versáteis que existe. Essa rede é capaz de resolver problemas complexos e de natureza diversa. São caracterizadas pela presença de pelo menos uma camada intermediária de neurônios situada entre a camada de entrada e a camada de saída. Ela possui arquitetura do tipo *feedforward* de camadas múltiplas e seu treinamento é efetuado de forma supervisionada.

As redes PMC funcionam da seguinte forma: os sinais são apresentados em sua camada de entrada. Os neurônios pertencentes às camadas intermediárias extraem a maior parte das informações as quais são codificadas em seus pesos sinápticos e limiares. A rede dessa maneira produz uma representação própria e única acerca do ambiente no qual está inserida. Em último, os neurônios da camada de saída recebem os sinais vindos da última camada intermediária e produz uma resposta padrão que será a saída disponibilizada pela rede neural.

3.2.1 Treinamento por retropropagação do erro

A extensão do método do gradiente descendente para redes de múltiplas camadas é conhecida como Regra Delta Generalizada ou *backpropagation*. O treinamento tem duas fases: *forward* e *backward*. Na primeira fase denominada *forward*, o sinal de entrada é propagado para frente, das entradas até a saída da rede. Como o valor da saída desejada para a entrada é conhecido, o erro para a camada de saída pode ser calculado, permitindo o ajuste dos pesos dessa camada. O erro entre as respostas produzidas pelos neurônios de saída da rede – $Y_j^{(3)}(k)$ – em relação aos respectivos valores desejados – $d_j(k)$ – é expresso em (3.1).

$$E(k) = \frac{1}{2}\sum_{j=1}^{n_3}\left[d_j(k) - Y_j^{(3)}(k)\right]^2 \tag{3.1}$$

Assume-se então a função erro quadrático para medir o desempenho local associado aos resultados produzidos pelos n_3 neurônios da camada de saída em relação à *k-ésima* amostra de treinamento. Na expressão (3.1), considera-se, sem perda de generalidade, uma rede composta por três camadas de neurônios com *n* entradas, n_1 neurônios na primeira camada intermediária, n_2 neurônios na segunda camada intermediária e n_3 neurônios na camada de saída. $Y_j^{(3)}(k)$ é o valor produzido pelo *j-ésimo* neurônio da camada de saída da rede para a *k-ésima* amostra de treinamento e $d_j(k)$ é o seu respectivo valor desejado.

O desvio dos neurônios da camada de saída em relação às saídas desejadas da rede são calculados diretamente; no entanto, não existem valores de saída para as camadas neurais intermediárias. Dessa forma, o ajuste dos pesos dessas camadas é feito via propagação para trás (retropropagação) do erro da camada de saída o que caracteriza a fase *backward* do algotimo *backpropagation*.

O algoritmo *backpropagation* ajusta os valores das matrizes de pesos da rede PMC em relação à direção oposta do vetor gradiente da função erro quadrático.

Em resumo, as aplicações sucessivas das fases de *forward* e *backward* fazem com que os pesos sinápticos e limiares dos neurônios se ajustem automaticamente ao final de cada iteração, ocasionando a redução gradativa da soma dos erros produzidos pelas respostas da rede em comparação às respostas desejadas. Assumindo um conjunto de treinamento composto por *N* amostras de treinamento, a medição da evolução do desempenho global do algoritmo *backpropagation* pode ser avaliada através do erro quadrático médio definido em (3.2). Esse erro é função dos parâmetros livres da rede (pesos sinápticos e limiares).

Em suma, o objetivo do processo de aprendizagem da rede é encontrar um conjunto ótimo de parâmetros livres que minimizem o erro quadrático médio expresso em (3.2).

$$E_M = \frac{1}{N}\sum_{k=1}^{N} E(k) \tag{3.2}$$

Recomenda-se a leitura de Haykin (1999) para o leitor que se interesse pelo desenvolvimento detalhado do algoritmo *backpropagation*.

3.2.2 Critério de Parada

Segundo Haykin (1999), não é possível demonstrar que o algoritmo de retropropagação convergiu e não há critérios bem definidos para encerrar a sua operação. Há somente critérios razoáveis para indicar a convergência do algoritmo cada um com o seu mérito particular. A seguir, são apresentados três critérios de parada:

- Considera-se que o algoritmo de retropropagação tenha convergido quando a norma euclidiana do vetor gradiente for suficientemente pequena. O problema desse critério é que, para se obter tentativas bem-sucedidas, os tempos de aprendizagem podem ser longos. Outra desvantagem é que esse método requer o cálculo do vetor gradiente – $\nabla(w)$.
- Considera-se que o algoritmo de retropropagação tenha convergido quando a taxa absoluta de variação do erro médio quadrático por época for suficientemente pequena. Infelizmente, esse critério pode resultar em um encerramento prematuro do processo de aprendizagem. Ademais, ele não avalia o grau de generalização da rede neural.
- Considera-se, em último, que houve convergência quando ficar aparente que o desempenho de generalização da rede atingiu o máximo. Após cada iteração do processo de aprendizagem, a rede é testada a fim de verificar seu desempenho de generalização e é encerrado quando esse desempenho estiver em um patamar adequado.

3.2.3 Dificuldades de Treinamento

A forma irregular da superfície de erro decorrente das não-linearidades das funções de ativação causa dificuldades no treinamento com algoritmos baseados no gradiente a

exemplo do *backpropagation* ou retropropagação. As não-linearidades possibilitam o surgimento de mínimos locais e de regiões planas com gradiente nulo, dificultando o treinamento.

Regiões de gradiente nulo tais como ótimos locais distantes do ótimo global podem ser evitadas por meio do ajuste adaptativo da taxa de aprendizado – η. Pequenos valores de η resultam em maior sensibilidade às variações da superfície e em menor agilidade no treinamento, enquanto que maiores valores de η podem evitar regiões planas e mínimos locais e ainda aumentar a rapidez do treinamento da RNA. No entanto, valores muito elevados para a taxa de treinamento pode conduzir a uma convergência prematura.

Outro desafio durante o treinamento de uma rede PMC é superar o *underfitting* e evitar a condição de *overfitting*. O *overfitting* é a situação em que há sobreparametrização da rede, isto é, quando a rede tem mais parâmetros (pesos sinápticos) do que o necessário para a resolução de determinado um problema. Em contrapartida, o *underfitting* ocorre quando a rede tem menos parâmetros do que o necessário. O objetivo do treinamento é encontrar o melhor ajuste possível na fronteira dessas situações extremas. Modelos sobreparametrizados tendem a ter maior variabilidade nas respostas, enquanto que modelos subparametrizados tendem a ter menos variabilidade e ser polarizados para um determinado tipo de resposta. O desafio em treinar eficientemente uma PMC é caracterizado pela busca do equilíbrio entre a polarização e a variância do modelo. Um modelo com alta variância resulta em *overfitting* enquanto que modelos polarizados resultam em *underfitting*. Tais características são conflitantes, ou seja, uma redução na polarização implica em um acréscimo da variância. Portanto, durante o treinamento da rede neural, é preciso buscar um equilíbrio entre ambas condições.

3.2.4 Melhoramentos no algoritmo *backpropagation*

Diversas variações do método *backpropagation* original visam tornar o processo de convergência mais eficiente. Dentre as variações mais conhecidas tem-se a inserção do termo *momentum*, o método *resilient-propagation* e o Levenberg-Marquardt.

Uma alternativa para reduzir o tempo de treinamento da rede e para evitar regiões de gradiente nulo é a adição de um termo de *momentum* às equações de ajuste de pesos sinápticos. O termo *momentum* é responsável pelo acúmulo do histórico dos ajustes anteriores, o que resulta em um termo residual quando o ajuste atual é nulo devido ao gradiente nulo. Regiões planas e mínimos locais podem ser evitados para um valor apropriado para a constante de *momentum* – α (REZENDE, 2005). A inserção do termo *momentum* visa ponderar o quanto as matrizes de pesos sinápticas foram alteradas entre duas iterações ou entre duas épocas

consecutivas. No início do treinamento, quando a solução atual encontra-se possivelmente longe da solução final que minimiza a função erro quadrático, a matriz de pesos varia muito entre épocas consecutivas. No entanto, com o transcorrer das épocas, o algoritmo tende a convergir para um ponto ótimo (local ou global) e as variações na matriz de peso tornam-se mínimas. Portanto, na etapa inicial da fase de treinamento, o termo *momentum* é significativo e contribui para acelerar a convergência, mas após algumas épocas, as variações na matriz de peso tornam-se ínfimas e o termo *momentum* torna-se irrelevante. Na etapa final da fase de treinamento retorna-se ao *backpropagation* convencional.

O método *resilient-propagation*, em vez de considerar as variações das magnitudes do gradiente da função de erro, considera apenas a variação de seu sinal. Isso porque variações muito pequenas do gradiente da função de erro – expressão (3.1) – combinado com as faixas de saturação das funções de ativação sigmoidais fazem com que o processo de convergência do algoritmo *backpropagation* se torne lento. Dessa forma, há um gasto de esforço computacional adicional para conduzir os valores das matrizes de pesos do PMC para regiões de variação dinâmica das funções de ativação.

No *resilient-propagation*, quando os sinais do gradiente forem os mesmos entre duas épocas consecutivas significa que é possível aumentar a taxa de aprendizado, porque o ponto de mínimo encontra-se relativamente distante. Em contrapartida, se os sinais do gradiente forem diferentes, significa que o ponto de mínimo foi ultrapassado. Nesse momento, deve-se reduzir a taxa de aprendizagem a fim de se convergir suavemente para o mesmo.

Em último, tem-se o método de Levenberg-Marquardt que é uma aproximação do método de Newton. O algoritmo *backpropagation* é um método de descida no gradiente da função erro quadrático a fim de minimizá-la. Já o algoritmo de Levenberg-Marquardt é um método do gradiente de segunda ordem baseado no método dos mínimos quadrados para modelos não-lineares o qual pode ser incorporado ao *backpropagation* a fim de aumentar a eficiência do treinamento.

3.2.5 Especificação da configuração topológica da rede PMC

Os fatores mais relevantes para definição da topologia da rede neural (número de camadas intermediárias e de neurônios que estão contidos nelas) são: a classe do problema, a disposição espacial das amostras de treinamento, a complexidade do problema a ser mapeado, o algoritmo de aprendizado usado, o valor inicial dos pesos sinápticos, dos limiares e dos parâmetros de treinamento.

Uma rede PMC com um neurônio em sua camada de saída é capaz de distinguir apenas duas classes; se tiver dois neurônios, pode representar, no máximo, quatro classes. Em termos gerais, uma rede PMC com *m* neurônios em sua camada de saída seria capaz de classificar, em termos teóricos, até 2^m classes.

No entanto, em termos práticos, devido à complexidade do problema abordado, tal codificação sequencial de conjuntos pode tornar o treinamento da rede sobremaneira custoso, pois as classes estariam sendo representadas por pontos que estão espacialmente bem próximos entre si (SILVA; SPATTI; FLAUZINO, 2010).

Um dos métodos utilizados para reduzir a complexidade topológica da rede é o *one of c-classes*. Tal método consiste em associar a saída de cada neurônio diretamente à classe correspondente; nesse caso, a quantidade de neurônios na camada de saída é igual ao número de classes do problema.

Dessa forma, a quantidade de entradas e de saídas do PMC é conhecida e depende basicamente das características do problema no qual se emprega a rede neural. No entanto, a quantidade de camadas intermediárias e o número de neurônios que há na(s) camada(s) intermediária(s) não é conhecido. Ademais, não há um método analítico definitivo que determina a melhor configuração topológica da rede PMC. Os métodos mais usuais para definição da arquitetura neural como a validação cruzada (*cross-validation*) são fundamentalmente empíricos e se baseiam no teste exaustivo de arquiteturas candidatas. Há três variantes para a validação cruzada:

- Validação cruzada por amostragem aleatória (*random subsampling cross-validation*): o conjunto total de amostras disponíveis é aleatoriamente dividido nos subconjuntos de treinamento e de teste. O subconjunto de treinamento será utilizado para treinar todas as topologias candidatas e o subconjunto de teste é aplicado para selecionar aquela que apresentar melhor resultado de generalização. Cerca de 60% a 90% das amostras pertencem ao subconjunto de treinamento e as demais; ao subconjunto de teste. Esse procedimento de partição deve ser repetido várias vezes durante o processo de aprendizado das topologias candidatas de maneira que diferentes amostras pertençam aos subconjuntos de treinamento e de teste. O desempenho global de cada topologia candidata será a média dos desempenhos em cada experimento.
- Validação cruzada *k*-partições (*k-fold cross-validation*): realiza-se a divisão do conjunto total de amostras em *k* partições sendo que *(k–1)* delas comporá o subconjunto de treinamento e as demais; o subconjunto de teste. O processo se repete

k vezes até que todas as partições tenham sido utilizadas como subconjunto de teste. Em geral, o número de partições está compreendido entre cinco e dez.

- Validação cruzada por unidade (*leave-one-out cross-validation*): consiste na utilização de uma única amostra para o subconjunto de teste e as demais para o subconjunto de treinamento. O processo é repetido até que todas as amostras tenham pertencido ao subconjunto de teste. Essa técnica é um caso particular do método de *k*-partições em que o parâmetro *k* é igual ao número de amostras disponíveis.

Um procedimento simples utilizado em conjunto com a técnica *cross-validation* é a parada antecipada (*early stopping*) em que o processo de aprendizagem para uma topologia candidata é checado após cada iteração pela aplicação das amostras do subconjunto de teste e é finalizado quando houver elevação do erro quadrático desse subconjunto entre épocas sucessivas.

3.3 MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN

Os Mapas Auto-Organizáveis de Kohonen ou SOM (*Self-Organization Maps*) é um tipo de rede neural que possui arquitetura reticulada, aprendizado não-supervisionado competitivo e que permite a identificação de padrões em vetores de dados multivariáveis (SPERANDIO; COELHO; QUEIROZ, 2003).

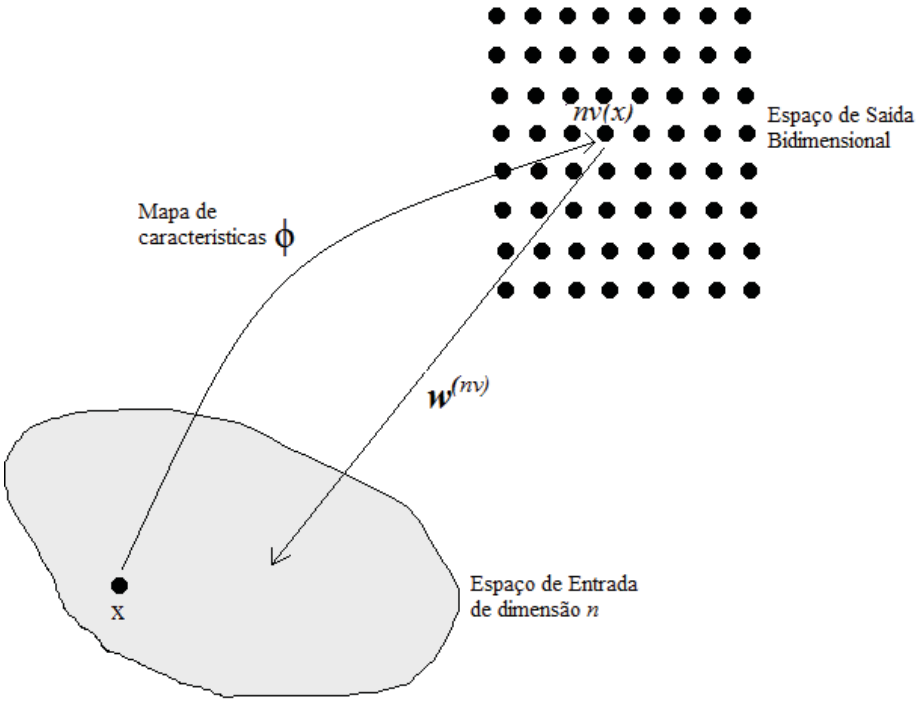
É possível extrair relações abstratas entre as variáveis do vetor de dados através de sua posição no mapa. As dependências estatísticas mútuas entre os elementos dos dados geralmente são não-lineares e dinâmicas. O SOM é uma ferramenta largamente utilizada para reconhecimento de padrões nas mais diversas áreas e se enquadra na função de realizar agrupamentos além de outros benefícios como facilitar a visualização de relações entre as variáveis (SPERANDIO; COELHO; QUEIROZ, 2003).

O mapa auto-organizável proposto por Teuvo Kohonen é uma rede neural construída em torno de uma grade unidimensional ou bidimensional de neurônios para capturar similaridades, regularidades e correlações importantes do espaço de entrada agrupando-os em classes ou em *clusters*. Tal mapeamento fornece uma representação estrutural dos dados de entradas por vetores de pesos dos neurônios. O algoritmo SOM é inspirado na neurobiologia incorporando mecanismos básicos da auto-organização cerebral como competição, cooperação e auto-amplificação.

Um mapa auto-organizável é caracterizado pela formação de um mapa topográfico dos padrões de entrada no qual as localizações espaciais dos neurônios na grade são indica-

tivas das características estatísticas intrínsecas contidas nos padrões de entrada (HAYKIN 1999), conforme está ilustrado na Figura 3. O padrão de entrada *x* pertencente ao espaço de entrada de dimensão *n* é representado no espaço de saída bidimensional pelo neurônio vencedor – $nv(x)$ – cuja localização está em função de *x*. $\boldsymbol{w}(nv)$ é a matriz de pesos do neurônio vencedor.

Figura 3: Mapeamento $\mathbb{R}^n \Rightarrow \mathbb{R}^2$.

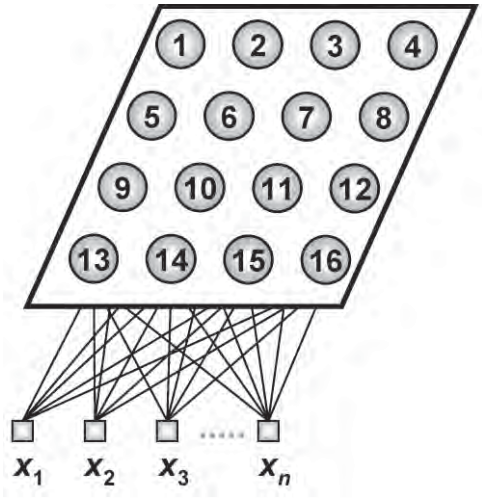


Fonte: Haykin (1999).

Esboça-se na Figura 4 uma grade neural bidimensional particular composta por dezesseis neurônios e *n* entradas. Pretende-se realizar um mapeamento $\mathbb{R}^n \Rightarrow \mathbb{R}^2$. Na expressão (3.3), tem-se a convenção para representar os vetores de pesos de cada um dos dezesseis neurônios que compõem a grade neural da Figura 4.

$$\begin{aligned} \boldsymbol{w}^{(1)} &= [w_{1,1} \;\; w_{1,2} \;\; \ldots \;\; w_{1,n}]^T \\ \boldsymbol{w}^{(2)} &= [w_{2,1} \;\; w_{2,2} \;\; \ldots \;\; w_{2,n}]^T \\ &(\ldots) \\ \boldsymbol{w}^{(16)} &= [w_{16,1} \;\; w_{16,2} \;\; \ldots \;\; w_{16,n}]^T \end{aligned} \tag{3.3}$$

Figura 4: Mapa topológico bidimensional 4 x 4.



Fonte: Simpson (1989).

3.3.1 Aprendizagem competitiva

O princípio básico do aprendizado não-supervisionado competitivo é a concorrência entre os neurônios. As amostras dos dados de entrada para treinamento da rede são apresentadas sequencialmente. Para cada amostra há um único neurônio vencedor.

Uma das principais estratégias para selecionar o neurônio vencedor consiste em determinar a proximidade entre o vetor de pesos de cada neurônio em relação ao vetor de entrada contendo os elementos da *k-ésima* amostra $\boldsymbol{x}^{(k)}$, conforme expressão (3.4).

$$dist_j^{(k)} = \sqrt{\sum_{i=1}^{n} \left(x_i^{(k)} - w_i^{(j)}\right)^2}, onde\ j = 1, \dots, n_1 \tag{3.4}$$

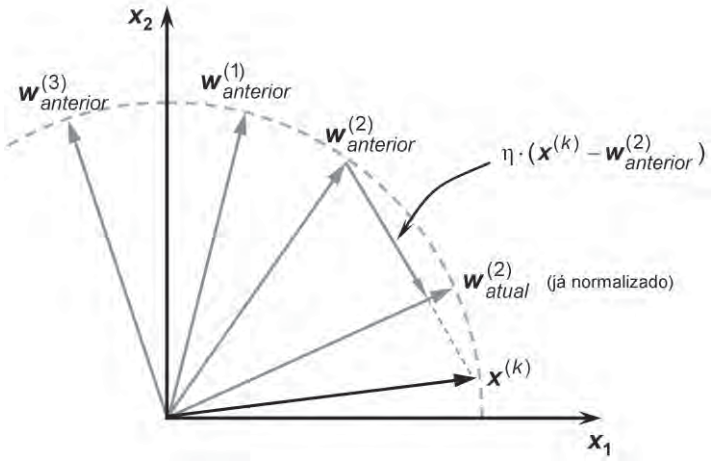
Sendo $dist_j^{(k)}$ a norma euclidiana entre o vetor de entrada $\boldsymbol{x}^{(k)}$ e o vetor de pesos $\boldsymbol{w}^{(j)}$ referentes à *k-ésima* amostra e ao vetor de pesos do *j-ésimo* neurônio, respectivamente. Na expressão (3.4), existem *n* amostras de entrada e n_1 neurônios na grade neural.

O neurônio *j* cujo vetor de pesos minimiza a expressão (3.4) é dito neurônio vencedor da competição e é considerado o neurônio mais ativo em relação à amostra *k*. O vetor de pesos do neurônio vencedor é ajustado de maneira a aproximá-lo ainda mais da amostra recém-apresentada. O processo de ajuste é conforme a expressão (3.5) e é ilustrado pela Figura 5. $\boldsymbol{w}^{(venc)}$ é o vetor de pesos do neurônio vencedor a ser ajustado e o parâmetro $\eta(n)$ define a taxa de aprendizagem e é função do número de épocas – *n*.

$$\boldsymbol{w}_{atual}^{(venc)} = \boldsymbol{w}_{anterior}^{(venc)} + \eta(n)\left(\boldsymbol{x}^{(k)} - \boldsymbol{w}_{anterior}^{(venc)}\right) \tag{3.5}$$

O processo de ajuste de pesos do neurônio vencedor é esquematizado na Figura 5. Três neurônios da rede de Kohonen representados pelos vetores de peso $\boldsymbol{w}^{(1)}$, $\boldsymbol{w}^{(2)}$ e $\boldsymbol{w}^{(3)}$ competem entre si em relação à amostra $\boldsymbol{x}^{(k)}$ pertencente ao espaço bidimensional composto pelas amostras de entrada x_1 e x_2. Todos os vetores de pesos e o vetor de entrada estão normalizados e encerrados em um círculo de raio unitário. O vetor $\boldsymbol{w}^{(2)}$ é o vencedor da competição, pois é aquele que mais se aproxima do vetor $\boldsymbol{x}^{(k)}$. Observa-se que o ajuste representado pela expressão (3.5) consistiu somente em rotacionar o vetor de pesos do neurônio vencedor em direção ao vetor representando a amostra.

Figura 5: Visão geométrica do ajuste do vetor de pesos do neurônio vencedor.



Fonte: Silva, Spatti e Flauzino (2010).

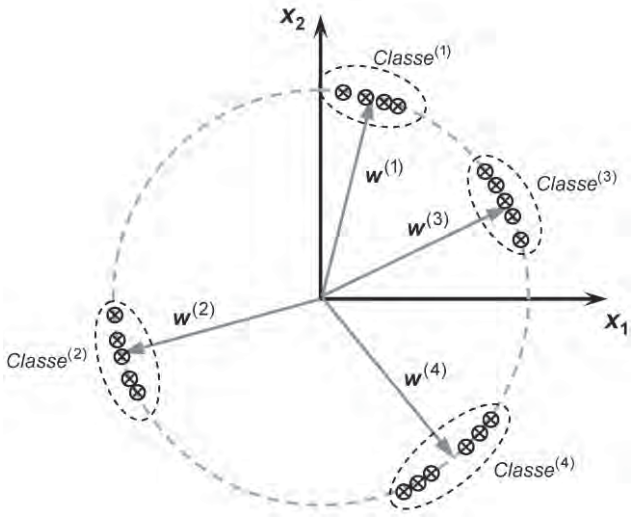
Na Figura 6 tem-se a distribuição espacial bidimensional de quatro vetores de pesos representando cada um dos quatro neurônios de uma rede neural de Kohonen após convergência do algoritmo competitivo. Cada um dos vetores de peso dos neurônios está posicionado no centro dos quatro aglomerados ou *clusters.* Os aglomerados representam as classes associadas ao problema. A quantidade de classes é vinculada ao número de neurônios utilizados na estrutura neural.

Observa-se que a quantidade inicial de neurônios a ser utilizada, visando representar as classes com características em comum é desconhecida *a priori*, porque o aprendizado é não-supervisionado.

A partir das Figuras 5 e 6 verifica-se que os padrões a serem classificados estão projetados no espaço em que estão definidos os vetores de pesos e cuja dimensão é igual àquela na qual as amostras de entradas estão definidas. Um dos requisitos para melhor efici-

ência do processo de convergência é que as variáveis de entrada sejam independentes para que a projeção ocorra em uma hiperesfera.

Figura 6: Distribuição dos vetores de pesos após treinamento do SOM.



Fonte: Silva, Spatti e Flauzino (2010).

Sperandio, Coelho e Queiroz (2003) resumem o processo de aprendizagem do SOM em três fases:

- Competição: cada amostra de dados é apresentada à rede consecutivamente. Procura-se o neurônio da grade bidimensional que melhor represente cada amostra de dado. O neurônio cujo vetor de pesos tiver maior similaridade com o vetor que representa a amostra atual é declarado vencedor da competição e nele é centrada uma função de vizinhança. Pode ser utilizada a menor distância euclidiana para encontrar o neurônio vencedor, por exemplo.
- Cooperação: o neurônio vencedor influenciará seus neurônios vizinhos a se aproximarem dele. Isso é feito através de uma função simétrica que tem amplitude máxima no centro e que decai à medida que a distância lateral aumenta. Tipicamente usa-se uma curva gaussiana[4]. A abrangência dessa função deve diminuir com o aumento das épocas de treinamento, de maneira que o algoritmo interfira cada vez menos na estrutura topológica do mapa bidimensional e convirja.

[4] Em sentido qualitativo, a vizinhança topológica gaussiana é biologicamente mais apropriada do que uma vizinhança retangular. Seu uso faz com que o algoritmo SOM convirja mais rapidamente do que com uma vizinhança topológica retangular (HAYKIN, 1999).

- Adaptação: o vetor de pesos do neurônio vencedor e de seus vizinhos são adaptados conforme as expressões (3.5) e (3.7), respectivamente. Objetiva-se aproximar esses vetores de pesos ao vetor que representa os dados de entrada.

Após a convergência do SOM, quando uma nova amostra for apresentada, basta verificar qual será o neurônio vencedor, pois este estará indicando a classe para a qual a presente amostra possui maior afinidade.

3.3.2 Conjunto de vizinhança e mapas de contexto

Os mapas topológicos informam como estão organizados espacialmente os neurônios da rede sendo normalmente unidimensionais ou bidimensionais.

Definido o arranjo espacial dos neurônios, basta especificar o critério de vizinhança interneurônios. Um dos critérios de vizinhança mais comuns consiste em adotar um raio *R* de abrangência que será utilizado pelos neurônios da rede a fim de definir seus respectivos vizinhos. Considere, por exemplo, que a distância entre os neurônios da grade neural da Figura 4 seja unitária. Define-se o conjunto de vizinhanças do neurônio *j* do mapa topológico pelo parâmetro $\Omega_j^{(R)}$ que indica os neurônios vizinhos do neurônio *j* delimitados por um círculo de raio *R* e cujo centro é o próprio neurônio *j*. Os neurônios vizinhos de cada um dos dezesseis neurônios da grade neural da Figura 4 estão indicados na expressão (3.6).

Considerando que um neurônio *j* venceu a competição em relação à apresentação de uma amostra, tanto o seu vetor de pesos como o de seus vizinhos serão ajustados. No entanto, os neurônios que estão na vizinhança do neurônio vencedor serão ajustados com taxas inferiores em relação ao neurônio vencedor. Os vetores de pesos do neurônio vencedor e de seus respectivos vizinhos são ajustados pelas expressões (3.5) e (3.7), respectivamente. $\eta(n)$ é a taxa de treinamento e é função do número de épocas – *n*.

A fim de aumentar a eficiência do treinamento, a taxa de aprendizagem deve reduzir gradativamente no decorrer do treinamento. Haykin (1999) utiliza a expressão (3.8) na qual a taxa de treinamento é função exponencial decrescente do número de épocas. Segundo essa mesma referência, a fase de treinamento é subdividida em duas outras fases: a fase de ordenação e a fase de refinamento. Conforme expressão (3.9), a fase de ordenação e a fase de refino duram, respectivamente, τ_2 e $(500 \times n_1)$ épocas.

$$\Omega_1^{(1)} = \{2, 5\}$$

$$\Omega_2^{(1)} = \{1, 3, 6\}$$

$$(...) \tag{3.6}$$

$$\Omega_{16}^{(1)} = \{12, 15\}$$

A referência adota $\tau_2 = 1.000$. n_1 é o número de neurônios na grade neural. Para treinar a grade neural da Figura 4, por exemplo, com $n_1 = 16$ duraria, aproximadamente, 9.000 épocas de treinamento.

$$\boldsymbol{w}_{atual}^{(viz)} = \boldsymbol{w}_{anterior}^{(viz)} + \eta(n)\, \alpha^{(\Omega)}(n) \left(\boldsymbol{x}^{(k)} - \boldsymbol{w}_{anterior}^{(viz)}\right) \tag{3.7}$$

$$\eta(n) = \eta_0 e^{\left(-\frac{n}{\tau_2}\right)} \tag{3.8}$$

$$n \cong \tau_2 + 500 \times n_1 \tag{3.9}$$

A taxa de ajuste dos vetores de peso dos vizinhos do neurônio vencedor – expressão (3.7) – é ponderada por uma função gaussiana conforme expressão (3.10). Quanto maiores forem as distâncias dos vetores de peso dos neurônios vizinhos em relação ao neurônio vencedor, menores serão os ajustes. O parâmetro $\alpha^{(\Omega)}$ é função de outro parâmetro $\sigma(n)$ que é função da quantidade de épocas – expressão (3.11). O parâmetro $\sigma(n)$ é função exponencial decrescente do número de épocas. O parâmetro τ_1 que consta na expressão (3.11) é avaliado em (3.12) e depende de duas outras constantes τ_2 e σ_0. À medida que o número de épocas aumenta, o vetor de pesos do neurônio vencedor e de seus respectivos vizinhos são ajustados com taxas menores de maneira que o algoritmo convirja.

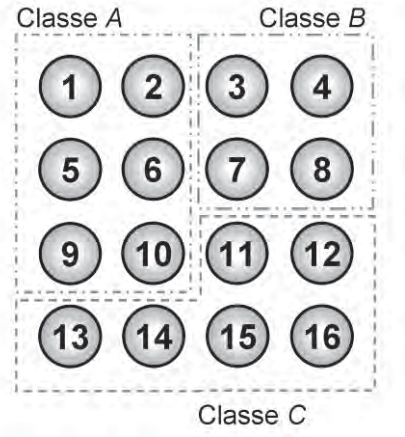
$$\alpha^{(\Omega)}(n) = e^{-\frac{\left\| w^{(venc)} - w^{(\Omega)} \right\|^2}{2\sigma(n)^2}} \tag{3.10}$$

$$\sigma(n) = \sigma_0\, e^{-\frac{n}{\tau_1}} \tag{3.11}$$

$$\tau_1 = \frac{\tau_2}{\log(\sigma_0)} \tag{3.12}$$

Após treinamento do SOM, podem-se utilizar ferramentas estatísticas e conhecimento especialista a fim de identificar possíveis classes que agrupam dados com características em comum. O mapa topológico da Figura 4 após treinamento e posterior análise pode ser dividido em regiões ou em grupos de neurônios que definem classes representativas do problema. Esse novo arranjo é denominado mapa de contexto. O mapa de contexto da Figura 7 possui três classes: a classe A representada pelos neurônios 1, 2, 5, 6, 9 e 10; a classe B representada pelos neurônios 3, 4, 7, 8 e a classe C representada pelos neurônios 11, 12, 13, 14, 15 e 16.

Figura 7: Mapa de contexto após treinamento do SOM da Figura 4.



Fonte: Silva, Spatti e Flauzino (2010).

A estrutura de configuração simples, bem como a dinâmica de treinamento diferenciada faz dos mapas auto-organizáveis de Kohonen uma ferramenta sofisticada para aplicações em problemas que envolvem classificação de padrões e identificação de agrupamentos com características próprias.

3.4 MÉTRICAS DERIVADAS DA MATRIZ DE CONFUSÃO

Neste trabalho, emprega-se a rede neural PMC em um Módulo de Classificação. Há várias outras técnicas que são usualmente utilizadas para tal finalidade tais como: aprendizado de máquina, *Naive Bayes*, etc. A métrica comumente utilizada para avaliar o desempenho de sistemas classificadores é a taxa de acerto apresentada na expressão (3.13). No entanto, essa métrica é inadequada para o problema de perdas comerciais, pois as duas classes de clientes compostas por clientes normais e anômalos são muito desiguais. Em média, nas ins-

peções em campo, para cada cliente anômalo encontrado, há nove clientes normais. Por exemplo, se um classificador considerar todos os clientes como normais, ele poderia ter uma alta taxa de acerto sem acertar um único caso de fraude.

$$Taxa\_Acerto = \frac{Quantidade\ de\ Acertos}{Quantidade\ Total} \tag{3.13}$$

No Quadro 5, tem-se a matriz de confusão para o problema de perdas comerciais com duas classes: clientes normais (N) e clientes fraudadores (F).

Quadro 5: Matriz de confusão para classificação com duas classes.

Matriz de Confusão		Classe Predita pelo Classificador	
		N	F
Classe Real	N	Q_{NN}	Q_{NF}
	F	Q_{FN}	Q_{FF}

Fonte: Ferreira (2008).

Sendo:

Q_{NN}: quantidade de clientes normais que foram corretamente classificados, ou seja, são os verdadeiros positivos;

Q_{NF}: quantidade de clientes que são normais, mas que foram incorretamente classificados como anormais, ou seja, são os falsos negativos;

Q_{FN}: quantidade de clientes que são anômalos, mas que foram incorretamente classificados como normais, ou seja, são os falsos positivos;

Q_{FF}: quantidade de clientes anômalos e que foram corretamente classificados, ou seja, são os verdadeiros negativos.

As métricas de confiabilidade negativa e especificidade são mais adequadas ao problema de perdas comerciais. Elas são obtidas a partir dos dados da matriz de confusão através das expressões (3.14) e (3.15), respectivamente. A confiabilidade negativa é a razão entre o número total de fraudadores corretamente classificados pelo número de casos classificados como suspeitos. A especificidade é a razão entre o número de fraudadores corretamente classificados pelo número total de fraudes que há.

$$c = \frac{Q_{FF}}{Q_{FF} + Q_{NF}} \tag{3.14}$$

$$e = \frac{Q_{FF}}{Q_{FF} + Q_{FN}} \tag{3.15}$$

3.5 CONSIDERAÇÕES FINAIS

Neste capítulo, abordaram-se as principais características das redes neurais com ênfase às redes neurais PMC e SOM.

No próximo capítulo, aborda-se a lógica *fuzzy* que, juntamente com as redes neurais, são parte fundamental da metodologia para detecção de clientes com perdas comerciais a ser detalhada no capítulo 5.

4 LÓGICA *FUZZY*

Neste capítulo são abordados os conceitos básicos a respeito da lógica *fuzzy* e o funcionamento das principais ferramentas que utilizam tal paradigma, isto é, os SIFs (Sistemas de Inferência *Fuzzy*) do tipo Mamdani e TSK (Takagi-Sugeno-Kang).

4.1 INTRODUÇÃO À LÓGICA *FUZZY*

A lógica *fuzzy* também denominada lógica nebulosa ou lógica difusa é uma teoria que incorpora a experiência, a intuição, o conhecimento especialista e a natureza imprecisa do processo decisório humano através de um conjunto de regras ou heurísticas simples.

A teoria dos conjuntos clássica e a teoria das probabilidades, embora úteis, nem sempre conseguem captar a riqueza da informação fornecida por seres humanos. A teoria dos conjuntos clássica não é capaz de tratar o aspecto vago da informação e a teoria das probabilidades (na qual a probabilidade de um evento determina completamente a probabilidade do evento contrário) é mais adaptada para tratar de informações frequencistas do que aquelas fornecidas por seres humanos (SANDRI; CORREA, 1999).

Em 1965, Lotfi A. Zadeh publicou o primeiro artigo sobre a teoria dos conjuntos nebulosos para tratar as incertezas não-probabilísticas, o que foi denominado *Fuzzy Sets* (ZADEH, 1965). A partir de 1978, Zadeh desenvolveu a teoria de possibilidades que trata a incerteza da informação, podendo ser comparada com a teoria de probabilidades. Esta teoria, por ser menos restritiva, pode ser considerada mais adequada para o tratamento de informações fornecidas por seres humanos do que a teoria de probabilidades. Tais teorias têm sido cada vez mais usadas em sistemas que utilizam informações fornecidas por seres humanos para automatizar procedimentos quaisquer.

As aplicações da lógica *fuzzy* são as mais diversas possíveis. Destacam-se o controle de processos, a aproximação funcional, o apoio à decisão, etc.

Dentre as vantagens da utilização da lógica *fuzzy* destacam-se: mecanismo de raciocínio similar ao do ser humano, por meio do uso de termos linguísticos; modelagem de conhecimento de senso comum; conhecimento ambíguo e conhecimento impreciso, mas racional; técnica de aproximação universal; robustez e tolerância à falha; além do baixo custo de desenvolvimento e de manutenção. As limitações estão relacionadas à geração das regras *fuzzy* e à definição das funções de pertinência; ambas, baseadas em uma avaliação subjetiva do conhecimento do especialista. Adiciona-se a isso a inexistência de técnicas de aprendizado.

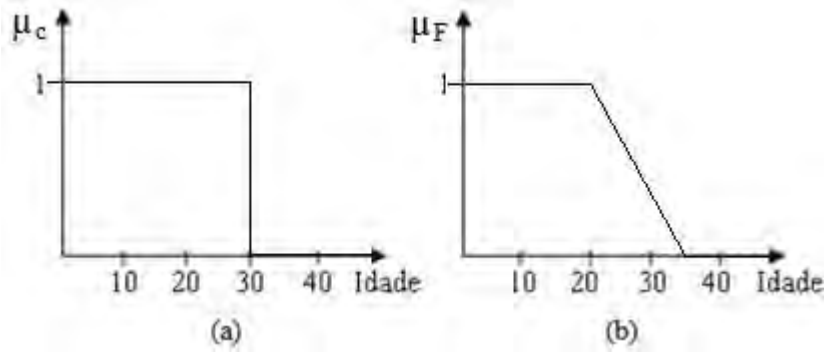
Em aplicações complexas, a tradução de informação imprecisa utilizando a teoria tradicional é inviabilizada em razão da complexidade matemática que poderia resultar. Entretanto, a teoria dos conjuntos *fuzzy* proporciona grande facilidade para descrever e processar esse tipo de informação através de variáveis linguísticas e de uma base de conhecimento *fuzzy* representada por um conjunto de regras (REZENDE, 2005).

4.2 CONJUNTOS *FUZZY*

Um conjunto *fuzzy F* de um universo de discurso *U* é representado comumente por uma função de pertinência real mapeada por $\mu_F: U \rightarrow [0,1]$, a qual associa cada elemento $x \in U$ a um número real $\mu_F(x)$ pertencente ao intervalo fechado [0,1], o qual representa o grau de pertinência de *x* em *F* (REZENDE, 2005).

Na Figura 8, tem-se o esboço de duas funções de pertinência representando o conjunto de indivíduos jovens. Na Figura 8 (a) está representada a função de pertinência para o conjunto clássico também denominado conjunto *crisp* e na Figura 8 (b) representa-se a função de pertinência para o conjunto *fuzzy*. Essas funções são definidas analiticamente pelas expressões (4.1) e (4.2), respectivamente. Observa-se que no conjunto *fuzzy* a transição entre o indivíduo totalmente jovem ($\mu_F = 1$) e o indivíduo não jovem ($\mu_F = 0$) ocorre de forma suave; ao contrário do conjunto crisp no qual tal transição ocorre de forma abrupta, descontínua.

Figura 8: Funções de pertinência para indivíduos jovens. (a) Conjunto *crisp*. (b) Conjunto *fuzzy*.



Fonte: Domingos (1998).

$$\mu_C = \begin{cases} 1, & se\ Idade \leq 30 \\ 0, & se\ Idade > 30 \end{cases} \tag{4.1}$$

$$\mu_F = \begin{cases} 1, & se\ Idade < 20 \\ \dfrac{35 - Idade}{15}, & se\ 20 \leq Idade \leq 35 \\ 0, & se\ Idade > 35 \end{cases} \tag{4.2}$$

No mundo real (e em grande parte das aplicações de interesse em engenharia) existem propriedades que são vagas, incertas ou imprecisas e, por isso, não é possível modelá-las satisfatoriamente através da lógica clássica bivalente. A teoria dos conjuntos *fuzzy* pode ser considerada como uma extensão da teoria clássica dos conjuntos e foi criada para tratar graus de pertinência intermediários entre a pertinência total e a não-pertinência. Na lógica *fuzzy*, o grau de pertinência de um elemento em relação a um conjunto é definido por uma função de pertinência que assume qualquer valor pertencente ao intervalo real fechado [0,1].

4.2.1 Operações Básicas entre conjuntos *fuzzy*

Nesta seção pretende-se apresentar as operações básicas de complemento, união e interseção entre conjuntos *fuzzy*.

- Complemento: o complemento de um conjunto *fuzzy A* possui função de pertinência conforme expressão (4.3). A operação complemento corresponde ao conectivo "NÃO".

$$\mu_{\neg A} = 1 - \mu_A(x) \tag{4.3}$$

- União: a união de dois conjuntos *fuzzy A* e *B* pode ser representada por $A \cup B$ ou por $A + B$. A união entre esses conjuntos *fuzzy* possui função de pertinência definida pela expressão (4.4). A operação união corresponde ao conectivo "OU".

$$\mu_{A \cup B} = max[\, \mu_A(x_i),\, \mu_B(x_i)\,] \tag{4.4}$$

- Interseção: a interseção entre os conjuntos *fuzzy A* e *B* pode ser representada por $A \cap B$ ou por $A \cdot B$. Tal operação resulta na função de pertinência expressa em (4.5). A interseção corresponde ao conectivo "E".

$$\mu_{A \cap B} = min[\, \mu_A(x_i), \mu_B(x_i)\,] \tag{4.5}$$

As definições apresentadas para a união e para a interseção de conjuntos *fuzzy* são particulares e foram propostas por Zadeh. Uma forma mais geral para definir a operação de união e de interseção entre conjuntos *fuzzy* é através das normas *S* e das normas *T*, respectivamente. Outros esclarecimentos constam em Rezende (2005) e em Domingos (1998).

4.3 REPRESENTAÇÃO DO CONHECIMENTO

4.3.1 Funções de Pertinência

As funções de pertinência são funções contínuas. Podem ter diferentes formas, no entanto, as mais comuns são as funções triangulares, trapezoidais, gaussianas e exponenciais.

Nos conjuntos *fuzzy*, a pertinência é gradual ao contrário do que ocorrem com os conjuntos da teoria de conjuntos clássica denominados conjuntos *crisp*. Nesses, a pertinência assume apenas dois estados: compatível ou incompatível.

4.3.2 Regras *Fuzzy*

As regras *fuzzy* fornecem uma descrição qualitativa do sistema em estudo. O conhecimento em um SIF é armazenado em geral na forma de regras composta pelos termos antecedente e o consequente conforme expressão (4.6).

SE <antecedente> **ENTÃO** <consequente> (4.6)

O antecedente é composto por um conjunto de condições envolvendo variáveis *fuzzy* e expressões linguísticas que, quando satisfeitas, mesmo que parcialmente, determinam o processamento da parte consequente da regra através de um mecanismo de inferência *fuzzy*. Tal processo é chamado disparo da regra.

Já o consequente é composto por um conjunto de ações que são geradas com o disparo da regra. Os consequentes de todas as regras disparadas são processados em conjunto para gerar uma resposta determinística para cada variável de saída do SIF.

Conforme Sandri e Correa (1999), em um SIF, é importante que existam tantas regras quantas forem necessárias para mapear totalmente as combinações dos termos das variáveis, isto é, que a base de regras seja completa, garantindo que exista sempre pelo menos uma regra a ser disparada para qualquer entrada.

4.3.3 Variáveis Linguísticas

As funções de pertinência para os conjuntos *fuzzy* devem ser definidas na base de dados que é uma parte do sistema (usualmente na forma paramétrica ou como tabela). Normalmente, vários termos linguísticos A_i são definidos no domínio de uma variável, e a coleção destes conjuntos *fuzzy* $[A_1, A_2, \ldots, A_M]$ são chamados de partição *fuzzy*. O número de termos

linguísticos na partição, chamado de cardinalidade do modelo, está relacionado com o nível de precisão com que o processo é descrito (DOMINGOS, 1998).

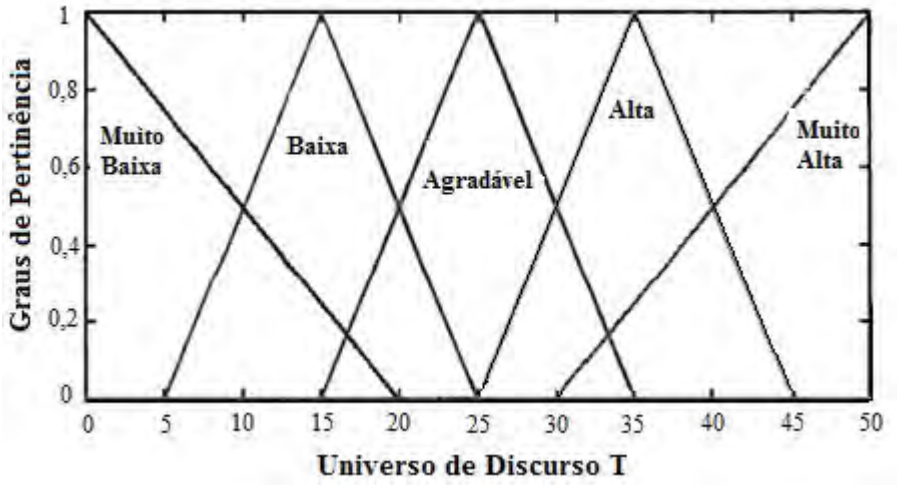
A base de regras junto com a base de dados constituem a base de conhecimento do sistema *fuzzy*. Os conjuntos *fuzzy* podem ser usados para construir conjuntos de termos linguísticos. Os conjuntos de termos representam abstrações dos valores da variável, isto é, são uma partição *fuzzy* de seus possíveis valores.

Em geral, uma variável linguística é associada a um conjunto de termos, no qual cada termo é definido no mesmo universo de discurso. A partição *fuzzy* determina quantos termos existirão no conjunto (DOMINGOS, 1998).

A capacidade de qualificar as variáveis de um problema de modo impreciso, em termos qualitativos em vez de quantitativos, fornece uma ideia do que é uma variável linguística. Define-se uma variável linguística como uma entidade utilizada para representar de modo impreciso, através da linguagem cotidiana, um conceito ou uma variável de um dado problema (REZENDE, 2005). Em geral, a cada variável linguística, estão associadas expressões linguísticas também denominadas termos primários. Tais expressões linguísticas qualificam as variáveis em termos linguísticos, portanto, de forma imprecisa.

Na Figura 9, por exemplo, tem-se uma partição *fuzzy* representando a variável linguística temperatura – T. Nessa partição existem cinco conjuntos *fuzzy* definidos no universo de discurso – intervalo fechado [0, 50] – dessa variável. Aos conjuntos *fuzzy* são associadas expressões linguísticas que qualificam de maneira imprecisa a variável temperatura. São elas: Muito Baixa, Baixa, Agradável, Alta e Muito Alta. Cada conjunto *fuzzy* é representado por sua respectiva função de pertinência que nesse exemplo é do tipo triangular.

Figura 9: Conjuntos fuzzy associados às expressões linguísticas da variável *fuzzy* temperatura – T.



Fonte: Rezende (2005).

Em suma, para um dado elemento *x* pertencente ao universo de discurso de uma variável *fuzzy* qualquer, o valor da função de pertinência $\mu_A(x)$ indica o quanto tal elemento satisfaz o conceito representado pelo conjunto *fuzzy* A.

Nas aplicações da teoria de conjuntos *fuzzy*, as funções de pertinência estão associadas a termos linguísticos, visando dar noções linguísticas às variáveis envolvidas.

As proposições *fuzzy* podem ser formadas atribuindo um valor *fuzzy*, ou seja, um conjunto *fuzzy*, à variável linguística *temperatura*. A partir da Figura 9 obtêm-se cinco proposições nebulosas:

- temperatura é Muito Baixa;
- temperatura é Baixa;
- temperatura é Agradável;
- temperatura é Alta;
- temperatura é Muito Alta.

4.4 MODELOS DE INFERÊNCIA *FUZZY*

Nesta seção discutem-se os dois modelos de inferência *fuzzy* mais usuais na literatura. A saber: o Modelo de Mamdani e o Modelo TSK (Takagi-Sugeno-Kang). Neste trabalho, os SIFs (Sistemas de Inferência *Fuzzy*) implementados são do tipo Mamdani. Devido a isso, mais detalhes serão fornecidos ao modelo Mamdani do que ao modelo TSK.

4.4.1 Modelo de Mamdani

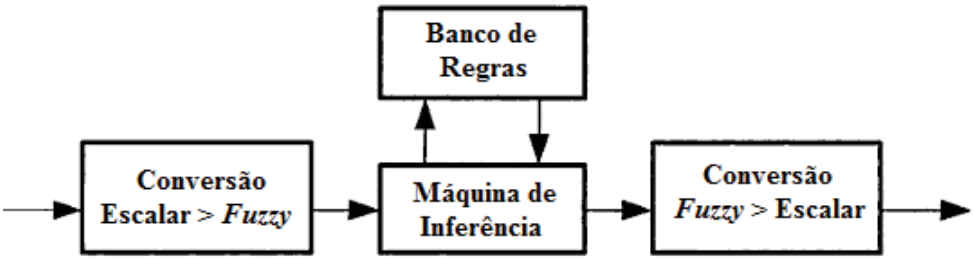
O modelo de Mamdani também conhecido como Modelo Nebuloso Linguístico é uma síntese do conhecimento. Para implementá-lo, é imprescindível o conhecimento de um especialista no momento da criação dos conjuntos *fuzzy* representativos das expressões linguísticas para cada variável linguística e, principalmente, para criação da base de regras que modela completamente o sistema em estudo.

A base de regras em um modelo Mamdani possui operações entre conjuntos *fuzzy* nos antecedentes e nos consequentes das regras. Na Figura 10, tem-se um diagrama simplificado relativo ao modelo de Mamdani. No módulo de entrada, as variáveis *fuzzy* são convertidas em conjuntos *fuzzy* equivalentes. Esse processo é denominado fuzificação. A máquina de inferência recebe valores *fuzzy* advindos do módulo de entrada, processa as regras existentes no banco de regras e gera um conjunto *fuzzy* de saída que é uma composição de todas as regras disparadas.

A regra semântica utilizada para o processamento de inferências com o modelo de Mamdani é a inferência *Max-Min*. Ela utiliza as operações de união e de interseção entre conjunto por meio dos operadores máximo e mínimo, respectivamente.

Desde que a saída do SIF seja um conjunto *fuzzy*, será necessário utilizar um processo de defuzificação para converter o valor inferido para um valor numérico a ser aplicado. Este valor, que deve ser aplicado ao sistema, é comumente citado na literatura como valor *crisp*, ou seja, um valor bem delimitado. Um dos métodos de defuzificação mais utilizado é o método do centróide. Outras estratégias de defuzificação são empregadas para aplicações específicas. De um modo geral, os métodos de defuzificação são cálculos intensos e não há um caminho rigoroso para analisá-los, a não ser através de estudos empíricos.

Figura 10: Diagrama simplificado de um modelo de inferência Mamdani.



Fonte: Rezende (2005).

A seguir, será explicitado o método de inferência *Max-Min* proposto por Zadeh (1965). A expressão (4.7) exibe uma regra *fuzzy* genérica; x_i são as entradas do sistema, $A_1, \dots, A_j$ são as expressões linguísticas definidas nas partições *fuzzy* de cada variável de entrada, y_1 e y_2 são as variáveis de saída e $B_1, \dots, B_m$ são expressões linguísticas definidas em suas respectivas partições *fuzzy*.

$$\boldsymbol{SE}\ x_1 = A_i\ \boldsymbol{E}\ x_2 = A_j\ \boldsymbol{E}\ \cdots \boldsymbol{E}\ x_p = A_j\ \boldsymbol{ENTÃO}\ y_1 = B_i\ \boldsymbol{E}\ y_2 = B_m \tag{4.7}$$

Durante o processo de conversão de escalar para *fuzzy*, os antecedentes de cada regra são avaliados através da interseção *fuzzy* entre os graus de pertinência das entradas atuais nas expressões linguísticas definidas para cada uma, conforme expressão (4.7). Tal ação gera um grau de pertinência de disparo para cada regra *fuzzy*. Conforme expressão (4.8), calcula-se para a *k-ésima* regra da base de conhecimento um coeficiente de disparo $D^{(k)}$ no qual os índices *k* nas expressões linguísticas representam os conjuntos *fuzzy* que compõem a regra *k* na base de conhecimento. Tal processo é dito fuzificação e transforma informações quantitativas em informações qualitativas.

Os conjuntos nebulosos *A* e *B* na regra (4.7) são geralmente conjuntos *fuzzy* multidimensionais. No entanto, as condições do antecedente são usualmente definidas na forma decomposta, utilizando simples proposições para os componentes individuais do vetor de entrada *x*. As proposições empregam conjuntos *fuzzy* constantes e são combinadas através de conectivos lógicos que representam as operações entre os conjuntos *fuzzy*. Normalmente, são utilizados os conectivos "E" (*AND* – conjunção) e "OU" (*OR* – disjunção). A forma conjuntiva é geralmente muito usual.

$$D^{(k)} = min\left[\mu_{A_1^k}(x_1),\ \mu_{A_2^k}(x_2), \cdots, \mu_{A_p^k}(x_p)\right] \tag{4.8}$$

Uma operação global de união produzirá um conjunto *fuzzy* para cada variável de saída, contendo informações sobre todas as regras disparadas para as entradas atuais. Na expressão (4.9) é mostrada a composição deste conjunto para a saída y_2 da regra genérica expressa em (4.7).

$$\mu_{B_i'}(y) = \max_{k=1,\cdots,n}\left[min\left(D^{(k)}, \mu_{B_i}(y)\right)\right], \forall\ y \in U_{y2} \tag{4.9}$$

O processo de inferência descrito na expressão (4.9) transforma uma informação qualitativa em outra informação qualitativa. O conjunto *fuzzy* gerado durante o processo de inferência pode ser utilizado diretamente em um diagnóstico qualitativo ou pode ser convertido em um escalar proporcional para atuação externa.

A conversão de um conjunto *fuzzy* para um valor escalar transforma informações qualitativas em informação quantitativas. Tal processo é dito defuzificação. Os métodos de defuzificação mais comuns são o centróide e o método da média dos máximos.

O método do centróide expresso em (4.10) calcula, para um conjunto *fuzzy* de saída, a abscissa (no universo de discurso definido para a variável de saída em questão) do ponto do centro de massa correspondente.

$$\hat{y}2 = \frac{\sum_{y \in U_{y2}}\ y.\mu_{B_i'}(y)}{\sum_{y \in U_{y2}}\ \mu_{B_i'}(y)} \tag{4.10}$$

Já no método da média dos máximos, conforme expressão (4.11), o valor numérico da saída é o ponto do universo de discurso que corresponde à média dos pontos de máximo locais da função de pertinência do conjunto de saída produzida pelo processo de inferência.

$$\bar{y}2 = \frac{\sum_{\hat{y} \in U_{y2}} \hat{y}_k . \mu_{B_i'}(\hat{y}_k)}{n_{\hat{y}}}; \ \ onde \ \hat{y} = \max_{y \in U', U' \subset U_{y2}} \left[\mu_{B_i'}(y)\right] \tag{4.11}$$

Existe um grau de liberdade para a criação de sistemas *fuzzy*, onde se define a estrutura das regras, o número e a definição dos conjuntos *fuzzy* de referência e a escolha do mecanismo de inferência.

Observa-se que não há um procedimento rigoroso para a síntese e para o desenvolvimento de modelo de inferência *fuzzy*. Uma base de regras de produção deve ser criada a partir do conhecimento do sistema (pela experiência de um especialista na operação ou por leis físicas). No entanto, há abordagens que proporcionam a geração automática de regras e partições *fuzzy* baseadas nos dados de entrada e de saída do sistema como os sistemas *neuro-fuzzy*, por exemplo.

4.4.2 Modelo de TSK (Takagi-Sugeno-Kang)

O SIF do tipo TSK tem por objetivo desenvolver uma aproximação sistemática, gerando regras *fuzzy* a partir de um conjunto de dados de entrada-saída disponíveis (DOMINGOS, 1998).

O modelo TSK, também denominado de sistema de inferência *fuzzy* paramétrico, é um aproximador universal, porque ele é capaz de aproximar qualquer função ou relação entre variáveis com uma dada precisão (REZENDE, 2005).

Similarmente ao modelo de Mamdani, os modelos *fuzzy* TSK são baseados na utilização de uma base de regras condicionais de inferência. No entanto, no modelo TSK, nos consequentes das regras, em vez de relações *fuzzy*, tem-se equações paramétricas que relacionam as entradas e as saídas do processo. A expressão (4.12) mostra uma regra genérica da base de conhecimento de um modelo TSK.

O modelo TSK tem sido comumente utilizado para a modelagem de sistemas não-lineares complexos a partir de dados principalmente como aproximador de funções e no controle de processos.

$$\boldsymbol{SE}\ x_1 = A_i\ \boldsymbol{E}\ x_2 = A_j\ \boldsymbol{E}\ \cdots \boldsymbol{E}\ x_p = A_m\ \boldsymbol{ENTÃO}\ y = \phi\left(x_1, x_2, \ldots, x_p\right) \tag{4.12}$$

O processamento de conhecimento em um modelo TSK é análogo ao dos modelos Mamdani. A etapa de fuzificação, isto é, conversão de escalar para conjunto *fuzzy* é idêntica conforme expressão (4.13). No entanto, a norma *T* é utilizada em vez da função *min()*.

$$D^{(k)} = T\left[\mu_{A_1^k}(x_1), \mu_{A_2^k}(x_2), \dots, \mu_{A_p^k}(x_p)\right] \tag{4.13}$$

A saída numérica é calculada diretamente pela soma das saídas das regras ponderadas pelos valores de ativação $D^{(k)}$ de cada uma delas conforme expressão (4.14).

$$\hat{y} = \frac{\sum_{i=1,\dots,k} D^{(i)}.\, \phi_i(x_1, \dots, x_p)}{\sum_{i=1,\dots,k} D^{(i)}} \tag{4.14}$$

Os modelos TSK são mais utilizados na substituição de modelos matemáticos convencionais em um esquema de controle ou na modelagem de sistemas reais (REZENDE, 2005). Nesse caso, é preciso que o modelo seja ajustado para se comportar como o sistema real que está representando. A partir dos conjuntos de dados de entrada e saída do sistema a ser modelado, os parâmetros *P*, dos consequentes das regras são estimados segundo algum índice de desempenho. A minimização do erro quadrático entre a saída do modelo TSK e os dados de saída disponíveis é normalmente utilizado como medida de desempenho.

Uma vez que as partições *fuzzy* e os parâmetros de saída estejam otimizados, o modelo está pronto para substituir o modelo convencional. A expressão (4.15) relaciona as entradas e saídas de um modelo TSK de primeira ordem, utilizando um conjunto de *k* regras *fuzzy* e *p* entradas.

$$y(x_1, x_2, \dots, x_p) = \sum_{i=1}^{k} \beta_i (P_{i0} + P_{i1}x_1 + \cdots + P_{ip}x_p) \tag{4.15}$$

Sendo:

$$\beta_i = \frac{T\left[\mu_{A_{i1}}(x_1), \dots, \mu_{A_{ip}}(x_p)\right]}{\sum_{j=1}^{k} T\left[\mu_{A_{j1}}(x_1), \dots, \mu_{A_{jp}}(x_p)\right]} \tag{4.16}$$

Uma aplicação comum dos modelos de inferência *fuzzy* é sua utilização para aproximação de funções não-lineares. Os modelos de inferência TSK são os mais adequados para esse fim. A existência de funções paramétricas nos consequentes de suas regras e a facilidade de se ajustarem a partir de um conjunto de dados de entrada e de saída faz com que eles sejam estritamente relacionados à tarefa de aproximador geral de funções (REZENDE, 2005).

Como sistemas capazes de processar eficientemente informações imprecisas e qualitativas de forma geral, os modelos de inferência *fuzzy* são adequados em processos que exigem tomadas de decisão por parte de operadores e gerentes de operação. Tais aplicações representam o conhecimento e a experiência existentes sobre um determinado estado do pro-

cesso e a partir da entrada de dados sobre seus estados atuais, podem-se inferir sua evolução temporal, as variações importantes que ocorreram ou mesmo gerar sugestões sobre as próximas ações a serem tomadas (REZENDE, 2005).

A tradução de informação imprecisa utilizando a teoria de controle convencional é inviabilizada em razão da complexidade matemática que poderia resultar. Todavia, a teoria de conjuntos *fuzzy* proporciona grande facilidade para descrever e processar esse tipo de informação, por meio de variáveis linguísticas e de regras *fuzzy*.

A seguir, os passo para implementação de um SIF:

- Definição dos universos de discurso das variáveis de entrada e saída.
- Partição dos universos de discursos definidos, isto é, criação das expressões ou termos linguísticos e graus de pertinência dos conjuntos *fuzzy* que representam cada termo.
- Determinação das regras que compõem a base de conhecimento.
- Definição de parâmetros semânticos tais como: escolha de operações *fuzzy* adequadas, formas de conversão das variáveis de entrada e de saída, etc.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo, apresentaram-se as estruturas principais da lógica *fuzzy*, conjuntos *fuzzy* e os SIFs do tipo Mamdani e TSK. Há cada vez maior abertura para se trabalhar com a lógica *fuzzy* em processos complexos que normalmente são difíceis de serem tratados com a teoria de conjuntos clássicos.

Fica implícito que a lógica *fuzzy* é baseada na grande variedade de manipulações dos conjuntos *fuzzy*, através dos quais o raciocínio humano é mais convenientemente modelado por meio da linguagem natural. De fato, o nível de precisão para determinadas aplicações pode ser ajustado para adaptar às necessidades e à exatidão dos dados disponíveis. Esta flexibilidade constitui um dos principais aspectos dessa teoria. Os modelos *fuzzy* baseados em regras permitem fazer uma interpretação do modelo de modo similar à forma humana de descrever a realidade do processo.

No capítulo 5, são abordados os detalhes dos SIFs do tipo Mamdani desenvolvidos para detecção de perdas comerciais no sistema de distribuição de energia.

5 METODOLOGIA PARA DETECÇÃO DE PERDAS COMERCIAIS NO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Este capítulo visa descrever em detalhes a metodologia proposta para detecção de perdas comerciais no SDEE (Sistema de Distribuição de Energia Elétrica) em cada um dos clientes das concessionárias de energia; principalmente, clientes residenciais, comerciais e industriais os quais consomem a maior parte da energia distribuída e são responsáveis pela parte majoritária das perdas comerciais. A metodologia emprega técnicas oriundas da área de sistemas inteligentes tais como: lógica *fuzzy* e as redes neurais PMC (*Perceptron* Multicamadas) e os Mapas Auto-Organizáveis de Kohonen (*Self Organization Maps* – SOM).

5.1 INTRODUÇÃO

A detecção precisa de quais clientes ocasionam perdas comerciais é uma tarefa complexa devido principalmente à grande quantidade de clientes, grande variedade de perfis de consumo, grande diversidade de fraudes e a informação insuficiente e muitas vezes desatualizada que se tem disponível a respeito de cada consumidor.

Adicionalmente, quanto mais próximo da carga, isto é, dos consumidores, maiores são as incertezas. As únicas informações que as concessionárias possuem a respeito da maioria de seus clientes são os dados cadastrais e o histórico de consumo mensal em kWh.

Com vistas à redução das perdas comerciais, as concessionárias inspecionam esporadicamente alguns de seus clientes a fim de detectar aqueles que possuem perdas comerciais. As informações obtidas nessas inspeções constituem o histórico de inspeções que contêm informações importantes a respeito dos clientes com esse tipo de perda elétrica. O histórico de inspeções e de consumo mensal junto com os dados cadastrais de cada cliente constitui a parte principal dos dados de entrada da metodologia proposta.

Neste trabalho, desenvolveu-se um sistema inteligente híbrido intercomunicativo que combina duas técnicas de sistemas inteligentes: redes neurais e lógica *fuzzy*. As redes neurais utilizadas são os Mapas Auto-Organizáveis de Kohonen e a rede *Perceptron* Multicamadas.

As redes neurais são fortemente baseadas em dados. São capazes de extrair conhecimento de uma base de dados; no entanto, esse conhecimento fica retido na própria rede neural. Em contrapartida, a lógica *fuzzy* utiliza o conhecimento fornecido a ela por um especialista, por exemplo. O conhecimento de um especialista é inserido em um SIF (Sistema de

Inferência *Fuzzy*) por meio de uma base de conhecimento composta por regras simples. Além disso, esse conhecimento é facilmente interpretado por um especialista. Logo, nesse aspecto, as redes neurais e a lógica *fuzzy* são técnicas complementares.

5.2 CLASSIFICAÇÃO DOS SISTEMAS INTELIGENTES HÍBRIDOS

Em Rezende (2005), apresenta-se um esquema de classificação para Sistemas Inteligentes Híbridos (SIHs) que utilizam mais de uma técnica de sistemas inteligentes para resolver problemas complexos. Tal classificação considera fatores como: funcionalidade, arquitetura de processamento e requerimentos de comunicação. Esse esquema é composto por três classes:

- Classe 1: Substituição de Função. Nesta categoria, utiliza-se uma técnica para implementar uma função de outra técnica. Essa forma de hibridismo não acrescenta nenhuma funcionalidade ao sistema inteligente, apenas tenta superar alguma limitação da técnica principal ou otimizá-la. Um exemplo é o uso de redes neurais para calibrar os parâmetros das funções de pertinência de um SIF.
- Classe 2: Híbridos Intercomunicativos. Categoria de sistemas híbridos usada para resolver problemas complexos que possam ser divididos em vários módulos independentes e intercomunicativos. Cada módulo implementa uma técnica inteligente a qual resolve uma parte do problema principal.
- Classe 3: Híbridos Polimórficos. Nesta categoria, uma única técnica é adaptada para realizar uma tarefa inerente a outra técnica. Visa-se dessa forma testar novas funcionalidade de uma técnica e entender como diferentes técnicas podem se relacionar.

De acordo com tal classificação, neste trabalho desenvolveu-se um SIH intercomunicativo (Classe 2) que combina as técnicas de sistemas inteligentes de redes neurais e lógica *fuzzy* em módulos independentes e que se comunicam.

Nesse contexto, segundo Haykin (1999), as redes neurais não podem fornecer uma solução trabalhando individualmente. Em vez disso, elas precisam ser integradas em uma abordagem consistente de engenharia de sistemas. Dessa forma, um problema complexo é decomposto em um número de tarefas relativamente simples e atribui-se às redes neurais um subconjunto dessas tarefas que coincidem com suas capacidades inerentes.

5.3 DESCRIÇÃO DO SISTEMA INTELIGENTE HÍBRIDO INTERCOMUNICATIVO

Esta seção visa descrever o SIH intercomunicativo proposto para detecção de clientes com perdas comerciais no sistema de distribuição de energia. A Figura 11 apresenta o fluxograma geral da metodologia. Ela é composta pelas seguintes entradas: dados dos clientes, dados de inspeções e alguns dados adicionais. Os sete módulos independentes e intercomunicativos que integram a metodologia são:

- ❖ módulo de pré-processamento;
- ❖ módulo de extração de atributos estatísticos;
- ❖ módulo de agrupamento;
- ❖ módulo de classificação;
- ❖ módulo de busca;
- ❖ módulo especialista;
- ❖ módulo principal.

As entradas do SIH bem como seus respectivos módulos são detalhados nas próximas seções.

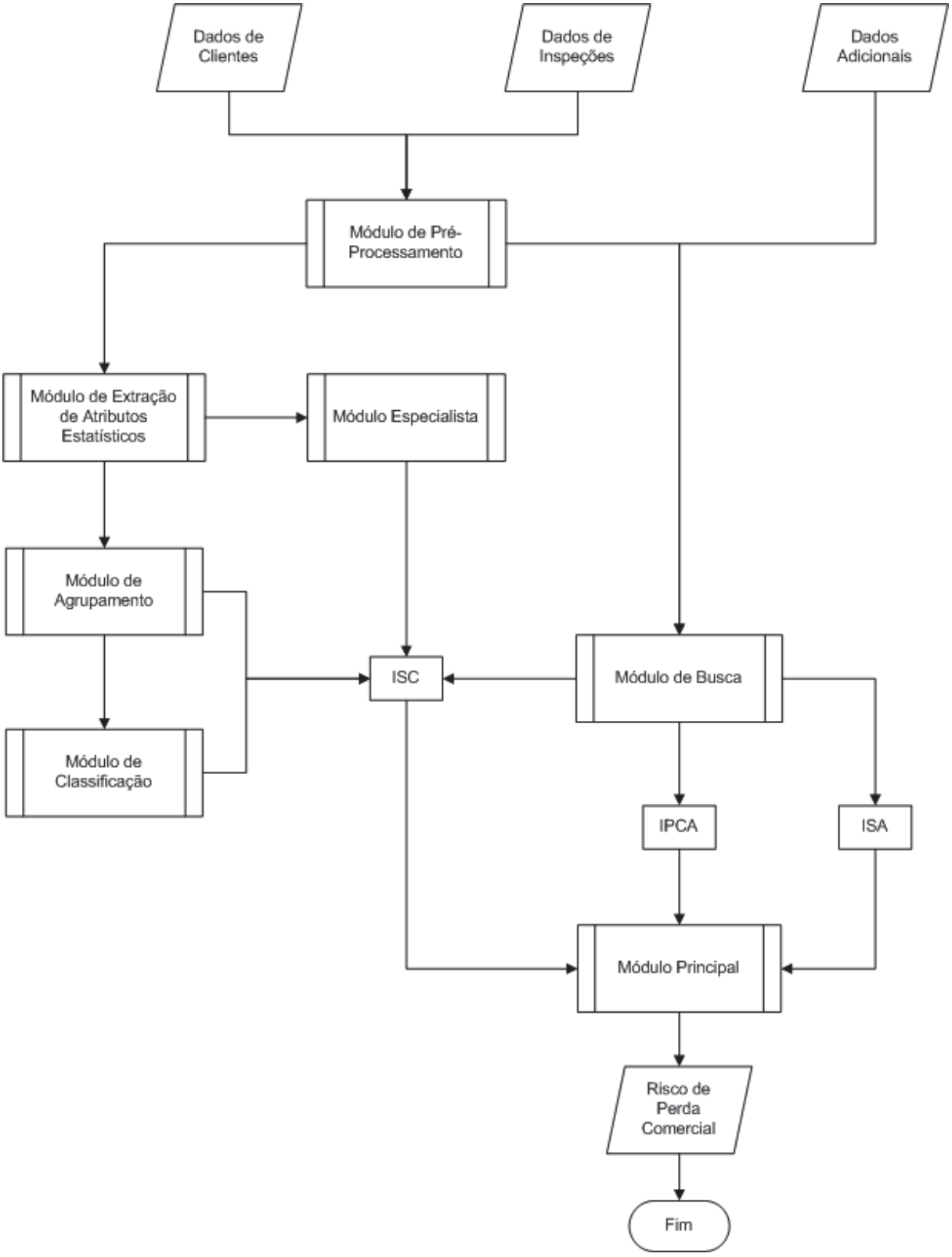
5.3.1 Dados de entrada do SIH Intercomunicativo

Os dados disponíveis a respeito de cada cliente são insuficientes e muitas vezes imprecisos, errôneos ou desatualizados. Devido a isso, a fim de detectar os clientes com perdas comerciais de forma mais precisa possível, é preciso utilizar todas as fontes de informações que se tem disponível e ponderá-las conforme a confiabilidade de cada uma.

Os dados cadastrais utilizados são:

- ❖ Classe do cliente: residencial, comercial, industrial, rural, poder público, iluminação pública, serviço público e consumo próprio.
- ❖ Nome completo.
- ❖ Localidade: nesta metodologia, a localidade considerada foram os bairros.
- ❖ Endereço completo.
- ❖ Faturamento: monofásico, bifásico ou trifásico.
- ❖ Carga declarada: demanda que o cliente indica no momento da ligação que está relacionada com a quantidade de aparelhos ou máquinas que ele possui.

Figura 11: Fluxograma geral do sistema inteligente híbrido para detecção de perdas comerciais no SDEE.



Fonte: Informações da pesquisa do autor.

- Atividade: varia conforme a classe do consumidor. Se a classe for comercial a atividade poderia ser farmácia, açougue ou padaria, por exemplo; se for residencial, a atividade poderia ser república de estudantes, casa de família, etc.
- Consumo irregular (*CI*): atributo binário que indica se o consumidor possui histórico de fraude.
- Religamentos (*R*): indica se o consumidor já se auto-religou após ter sua ligação cortada pela companhia por inadimplência, por exemplo.

Em último, têm-se os dados adicionais. São compostos pelo alimentador no qual o cliente está conectado, pelo *IPCA* (Índice de Perda Comercial do Alimentador) e pelo *ISA* (Índice de Suspeita da Área).

O *IPCA*, conforme expressão (5.1), corresponde a um percentual em relação às Perdas Globais (PGs) de um dado alimentador devido exclusivamente às Perdas Comerciais (PCs). A perda global de cada alimentador é conhecida e é calculada pela diferença entre a energia distribuída e a energia faturada aos consumidores. As perdas comerciais do alimentador são estimadas pela diferença entre as perdas globais e as perdas técnicas do alimentador previamente calculadas[5].

$$IPCA = 100 \times \frac{PC}{PG} \tag{5.1}$$

O *ISA* representa a vulnerabilidade de uma determinada área às perdas comerciais. Essa área pode ser um bairro, um grupo de bairros, uma área específica e corresponde a uma região que contém os clientes de um alimentador específico em análise. Múltiplos fatores de natureza diversa e subjetiva influem no *ISA* como pobreza, impunidade, corrupção, falta de infraestrutura, etc. Por isso, o desenvolvimento de uma metodologia que forneça o valor do *ISA* é algo muito complexo e impreciso e, por isso, neste trabalho, seu valor é baseado na experiência de especialistas. A partir da experiência, sabe-se, por exemplo, que uma região na qual a maioria das UCs (Unidades Consumidoras) possuem ar-condicionado ou na qual a maior parte dos clientes são estudantes universitários, possui maior *ISA*. Neste trabalho, o *ISA* pode assumir três valores: baixo, médio e alto. Seu valor é baseado unicamente no conhecimento do especialista acerca das características técnicas, sociais, políticas e históricas da região em análise.

[5] Para informações detalhadas sobre o calculo das Perdas Técnicas (PTs) em alimentadores da rede de distribuição consultar Oliveira (2009).

5.3.2 Dados de pré-processamento

Os dados de entrada são avaliados a fim de verificar a sua integridade e confiabilidade, pois, para que a metodologia funcione eficientemente é preciso ter uma base de dados consistente, atualizada e íntegra que represente os consumidores de uma dada região. Em termos práticos, é feita uma limpeza dos dados e uma formatação para adequá-los ao formato estabelecido *a priori* da base de dados desenvolvida. Em último, é feita uma verificação a fim de eliminar registros duplicados, nulos ou inconsistentes.

5.3.3 Módulo de Extração de Atributos Estatísticos

Este módulo objetiva extrair atributos estatísticos baseados na detecção de regimes a partir da curva de consumo mensal dos clientes. Essa curva é formada a partir da conexão dos pontos discretos que representam o consumo de cada mês.

A extração de atributos visa obter dados mais representativos e com menor dimensionalidade do que os originais. Adicionalmente, reduzem-se as redundâncias dos dados.

Segundo Cometti e Varejão (2005), a importância dos atributos extraídos da curva de consumo varia conforme a região e conforme a base de dados. Então, devem-se selecionar os atributos mais significativos a fim de melhorar a qualidade da predição.

No entanto, existem incontáveis fatores que interferem no perfil da curva de consumo mensal como: condições climáticas, aquisição de um novo eletrodoméstico, período de férias (no qual há acréscimo do consumo porque a família recebeu visitas ou decréscimo porque a família costuma viajar nesse período), etc. Tais fatores são difíceis de serem monitorados. Todavia, mesmo diante da grande imprecisão da curva de consumo mensal, diversos trabalhos utilizam ferramentas para extração de características dessa curva tais como: coeficientes de Fourier, coeficientes *Wavelets*, coeficientes de polinômios ortogonais, etc.

Este módulo visa extrair dezesseis atributos estatísticos da curva de consumo mensal a fim de detectar fraudes e/ou anomalias a exemplo de trabalhos anteriores. Um diferencial dos atributos estatísticos em relação aos demais, é que eles são mais intuitivos e, portanto, mais fáceis de serem analisados por um especialista.

Os atributos estatísticos baseados na detecção de regimes da curva de consumo mensal foram adaptados a partir do trabalho de Ferreira (2008) e são listados a seguir:

- Número de Regimes (*NR*): afere o número de regimes ou patamares da curva de consumo mensal do cliente.

- Coeficiente de Variação (*CV*): representa a variabilidade do histórico de consumo mensal em relação à sua média. É uma medida estatística adimensional e é obtida conforme expressão (5.2).

$$CV = 100 \times \frac{Desvio\ Padrão}{Média} \tag{5.2}$$

- Porcentagem de Quedas em relação ao regime (*PQ*): porcentagem de meses que estão abaixo de seu respectivo regime.
- Porcentagem de Aumento em relação ao regime (*PA*): porcentagem de meses que estão acima de seu respectivo regime.
- Número de Quedas em relação ao regime (*NQ*): afere a quantidade de regimes de queda.
- Número de Aumentos em relação ao regime (*NA*): afere a quantidade de regimes de aumento.
- Porcentagem de Tempo no Regime Inicial (*PTRI*): afere a porcentagem de tempo, isto é, o percentual de meses, em relação ao período total avaliado, em que o consumo esteve no regime inicial.
- Porcentagem de Tempo no Regime de Queda (*PTRQ*): percentual de meses em que o consumo esteve em um regime de queda.
- Porcentagem de Tempo no Regime Aumento (*PTRA*): percentual de meses em que o consumo esteve em um regime de aumento.
- Número de Zeros (*NZ*): quantidade de zeros da curva de consumo que corresponde aos meses cujo consumo é inferior a 10% do valor médio da curva de consumo total.
- Número de Regimes na Faixa Média (*NRFM*): define-se uma faixa em torno do valor médio da curva de consumo que corresponde a 10% da diferença entre a amplitude máxima e a amplitude mínima. Quantificam-se os regimes que estão nessa faixa média.
- Número de Regimes Abaixo da Faixa Média (*NRAbFM*): quantidade de regimes que estão abaixo da faixa média.
- Número de Regimes Acima da Faixa Média (*NRAcFM*): quantidade de regimes acima da faixa média.

- ❖ Número de Regimes na Faixa do Regime Inicial (*NRFRI*): define-se uma faixa em torno do regime inicial denominada faixa inicial que corresponde a 10% da amplitude total da série de consumo.
- ❖ Número de Regimes Abaixo da Faixa do Regime Inicial (*NRAbFRI*): quantidade de regimes abaixo da faixa do regime inicial.
- ❖ Número de Regimes Acima da Faixa do Regime Inicial (*NRAcFRI*): quantidade de regimes acima da faixa do regime inicial.

5.3.4 Módulo de Agrupamento

Os clientes; cada um deles representados neste módulo pelos dezesseis atributos estatísticos extraídos a partir de seus respectivos históricos de consumos mensais, são agrupados em classes ou *clusters*.

A cada agrupamento, associa-se um índice denominado *GSA* (Grau de Suspeita do Agrupamento) que o identifica. O *GSA* representa o risco de os clientes pertencentes a um dado agrupamento de possuírem um histórico de consumo atípico ou suspeito. Esse índice varia entre [0,1]. O *GSA* é um dos índices utilizados para cálculo do *ISC* (Índice de Suspeita do Cliente) a ser abordado posteriormente na seção 5.3.8.

Conforme Silva, Spatti e Flauzino (2010), a identificação dos *clusters* é importante para o entendimento das relações entre os seus elementos constituintes. Para agrupar os clientes, utiliza-se a rede neural de Kohonen na forma de mapas auto-organizáveis bidimensionais ou SOM (ver seção 3.3).

De acordo com Haykin (1999), em problemas de classificação de padrões, o primeiro e mais importante passo é a seleção, isto é, a extração de características do espaço de entradas que, em geral, é executado de maneira não-supervisionada. O objetivo desse passo é selecionar um conjunto de padrões que contém o conteúdo de informação essencial dos dados de entrada os quais precisam ser categorizados. O mapa auto-organizável é adequado para a seleção de características a partir de dados de entrada gerados por um processo não-linear, por exemplo.

No entanto, antes de agrupar os clientes, é preciso treinar o SOM. No Quadro 6, estão as principais características e configurações utilizadas para treinamento e operação do SOM. Em conformidade com o exposto na seção 3.3, as entradas, isto é, o vetor contendo os dezesseis atributos é normalizado de forma unitária visando obter maior eficiência na fase de aprendizagem.

Quadro 6: Parâmetros do Mapa Auto-Organizável de Kohonen.

Treinamento	Aprendizado Competitivo
Tipo de treinamento	Não-Supervisionado
Arquitetura da rede	Reticulada
Topologia da rede	Grade Bidimensional 4 x 4, 5 x 5 e 6 x 6
Tipo de aprendizagem	*On-line*
Critério convergência	Variação da magnitude dos vetores de peso da rede
Função de ajuste dos vizinhos	Gaussiana
Topologia da vizinhança	Círculo de Raio Unitário, fixo e retangular
Taxa de treinamento	Decresce exponencialmente com aumento das épocas

Fonte: Informações da pesquisa do autor.

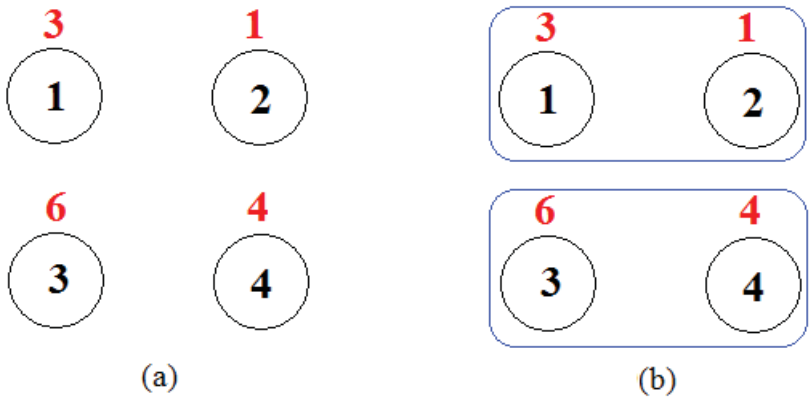
Além de agrupar os clientes em classes e de qualificar cada classe através do *GSA*; antes, este módulo encontra a quantidade de agrupamentos que é preciso para representar todos os clientes submetidos à análise. Isso porque, para o problema de perdas comerciais, não é possível saber *a priori* a quantidade ótima de classes que é preciso para representar os clientes em análise. Por exemplo, na Figura 12 (a) há um SOM elementar 2 x 2 constituído por quatro neurônios e cujas características são análogas ao SOM 4 x 4 da Figura 4 (seção 3.3). Os neurônios são representados pelos círculos. Neste caso, a capacidade máxima e mínima de agrupamentos é quatro e um, respectivamente. O número em vermelho acima dos círculos representa a quantidade de amostras associadas a cada neurônio após a fase de treinamento e de operação do SOM. Neste exemplo, existem quatorze amostras ou clientes em análise. A Figura 12 (b) contém o mapa de contexto que representa uma redução de agrupamentos de quatro para dois. Tal mapa é construído da seguinte forma: os neurônios vizinhos do neurônio um são os neurônios dois e três – $\Omega_1^{(1)} = \{2,3\}$. Entre os vizinhos do neurônio um aquele que mais se aproxima dele com respeito ao número de amostras é o neurônio dois. Neste instante, liga-se o neurônio um ao neurônio dois. Logo, há três agrupamentos: 1–2; 3 e 4. Esse processo é repetido para os neurônios restantes da grade e, ao final do processo, chega-se ao mapa de contexto esboçado na Figura 12 (b) com dois agrupamentos: 1–2 e 3–4.

Os quatorze clientes foram agrupados em duas classes. O próximo passo é qualificar cada um desses agrupamentos através do *GSA*. Antes, faz-se uma redução de variáveis de dezesseis para oito.

O módulo anterior – Módulo de Extração de Atributos Estatísticos – acabou por aumentar muito o conjunto de atributos da base de exemplos. A grande quantidade de atributos torna lento o processo de aprendizagem do Módulo de Classificação (seção 5.3.5).

Ademais, alguns atributos podiam ter baixo poder informativo, servindo apenas para reduzir o desempenho do classificador. A redução do número de variáveis é útil ao cálculo do *GSA* e à redução do número de entradas da rede PMC (Módulo de Classificação). Os atributos *NR*, *CV* e *NZ* são mantidos. A eles são adicionadas outras cinco novas variáveis criadas a partir dos treze atributos restantes conforme expressões em (5.3).

Figura 12: SOM elementar 2 x 2. (a) SOM após fase de treinamento e operação. (b) Formação do mapa de contexto.



Fonte: Informações da pesquisa do autor.

$$
\begin{aligned}
PQAR &= PQ - PA \\
NQAR &= NQ - NA \\
PTQA &= PTRQ - (PTRI + PTRA) \\
RFM &= NRAbFM - (NRFM + NRAcFM) \\
RFRI &= NRAbFRI - (NRFRI + NRAcFRI)
\end{aligned}
\tag{5.3}
$$

Sendo:

PQAR: Percentual de Queda e Aumento de Regime;

NQAR: Número de Queda e Aumento de Regime;

PTQA: Percentual de Tempo em regimes de Queda e Aumento;

RFM: Regimes na Faixa Média;

RFI: Regimes na Faixa do Regime Inicial.

Quanto mais positivas forem as cinco variáveis recém-criadas, maior é a tendência de redução do consumo ao longo do tempo. E quanto maiores forem as três variáveis restantes (*NR*, *CV* e *NZ*), mais atípico e anormal é o perfil de consumo do cliente.

O *GSA* é obtido da seguinte forma: calcula-se o valor máximo de cada uma das oito variáveis para cada agrupamento. Aquele que possuir os maiores valores máximos e em

maior quantidade, possuirá também o maior *GSA*. Logo, um agrupamento cujo *GSA* é alto, é constituído por clientes cujo perfil de consumo é atípico e/ou nos quais há uma tendência de redução do consumo de energia ao longo do tempo.

O SOM é utilizado também para criação de um conjunto de treinamento representativo para a rede neural PMC (Módulo de Classificação). Esse conjunto é composto por porcentagens iguais de clientes pertencentes a cada um dos agrupamentos criados neste módulo.

5.3.5 Módulo de Classificação

Este módulo utiliza a rede neural PMC para realizar classificação dos clientes em duas classes: clientes com perfil de consumo normal e clientes com perfil de consumo atípico ou suspeito. Um problema de classificação de padrões consiste em associar um padrão de entrada (amostra) a uma classe previamente definida. Este módulo recebe como entrada as oito variáveis normalizadas produzidas pelo Módulo de Agrupamento e possui saída binária.

A partir do histórico de inspeções obtêm-se o histórico de consumo mensal de clientes com perdas comerciais devido às fraudes ou erro na medição, por exemplo. A eles é adicionada uma porcentagem do histórico de consumo de clientes normais de cada um dos agrupamentos construídos no Módulo de Agrupamento. A esse conjunto de clientes denomina-se conjunto de treinamento e será utilizado para treinar a rede PMC. Dessa forma, a rede neural “aprende” quais são os perfis de consumo característico de clientes com algum dos diversos problemas que ocasionam perdas comerciais.

No Quadro 7, estão resumidas as principais características, topologias, técnicas de aprendizado empregadas na fase de treinamento e de operação da rede neural PMC.

A rede PMC possui apenas uma camada neural intermediária. Segundo Silva *et al*. (2010), tais redes são normalmente menos propensas a estacionar em mínimos locais, pois sua estrutura mais compacta reduz a complexidade geométrica da função que mapeia o erro quadrático médio. Além disso, um menor número de camadas intermediárias acelera a fase de treinamento. Logo, quanto menos camadas intermediárias, melhor em termos de simplicidade da arquitetura neural, de generalização e de tempo de processamento computacional. As redes PMC com apenas uma camada intermediária conseguem classificar padrões que estejam dispostos em regiões convexas e redes com duas camadas intermediárias são capazes de classificar padrões que estejam em quaisquer tipos de regiões geométricas, inclusive formatos não-convexos.

Similarmente ao *GSA*, a saída deste módulo também é utilizada para o cálculo do *ISC* que é uma das três entradas do Módulo Principal (seção 5.3.8).

Quadro 7: Parâmetros da rede neural PMC.

Treinamento	Regra Delta Generalizada com *Momentum*
Tipo de Treinamento	Supervisionado
Arquitetura da Rede	*FeedForward* ou Rede Alimentada Diretamente com Múltiplas camadas
Topologia	8 – 20 – 2
Tipo de Aprendizagem	*On-line* ou Sequencial
Função de Ativação	Sigmoidal (Logística e Tangente Hiperbólica)
Número de Camadas Neurais	3 camadas neurais
Critério de Convergência	Minimização do erro quadrático médio global
Amostras usadas no treinamento	70% das amostras
Seleção da Melhor Topologia	Validação Cruzada por amostragem aleatória
Seleção do número de saídas	Metodologia *one of c-classes*

Fonte: Informações da pesquisa do autor.

5.3.6 Módulo Especialista

Este módulo visa incorporar o conhecimento dos especialistas em perdas comerciais em um SIF (Sistema de Inferência *Fuzzy*). Utiliza três atributos estatísticos extraídos do histórico de consumo mensal pelo Módulo de Extração de Atributos Estatísticos. Esses atributos são selecionados para serem as entradas do SIF. A partir do valor numérico desses atributos é possível aos especialistas afirmar se determinado cliente possui um perfil de consumo anômalo ou não.

Este módulo consiste de um SIF do tipo Mamdani implementado no *Fuzzy Logic Toolbox 2*, um aplicativo que compõe o programa MATLAB® 7[6] e que é dedicado ao desenvolvimento de SIFs do tipo Mamdani e TSK (The MathWorks, 2007).

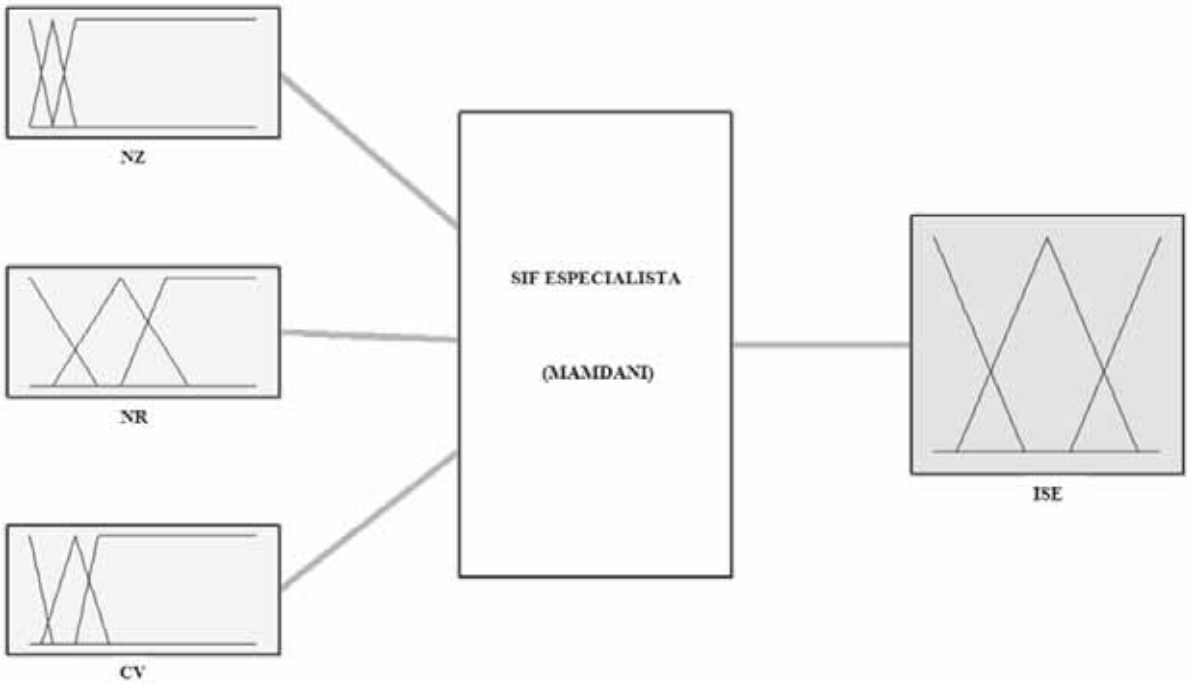
O modelo Mamdani foi selecionado em vez do modelo TSK (Takagi-Sugeno-Kang), porque este último é baseado em dados, paramétrico. A partir do conjunto de entradas e saída desejadas, ajustam-se os coeficientes do polinômio de saída. A presente metodologia já utiliza técnicas baseadas em dados que são as redes neurais PMC e SOM. Além disso, o modelo TSK é utilizado preferencialmente em problemas mais previsíveis e com menor grau de liberdade como em problemas envolvendo sistemas de controle, por exemplo. Em contra-

[6] MATLAB® 7 é uma marca registrada da The MathWorks, Inc.

partida, o sistema Mamdani é baseado unicamente na experiência e conhecimento do especialista e, por isso, é mais intuitivo e possui maior grau de liberdade.

Na Figura 13, exibe-se uma visão geral do SIF especialista desenvolvido. As entradas deste módulo são: o Número de Zeros (*NZ*), o Número de Regimes (*NR*) e o Coeficiente de Variação (*CV*). Tais entradas foram escolhidas, porque são mais intuitivas, mais fáceis de serem entendidas pelo especialista e algumas já são usadas na prática nas concessionárias, como o *NZ*, por exemplo. Além disso, outras variáveis de entrada podem ser facilmente adicionadas. Este módulo possui uma variável de saída, o *ISE* (Índice de Suspeita do Especialista). Logo, a partir de regras simples envolvendo essas três variáveis de entrada e a variável de saída, espera-se identificar históricos de consumo atípicos possivelmente resultantes de clientes com perdas comerciais.

Figura 13: Visão geral do SIF especialista.



Fonte: Informações da pesquisa do autor.

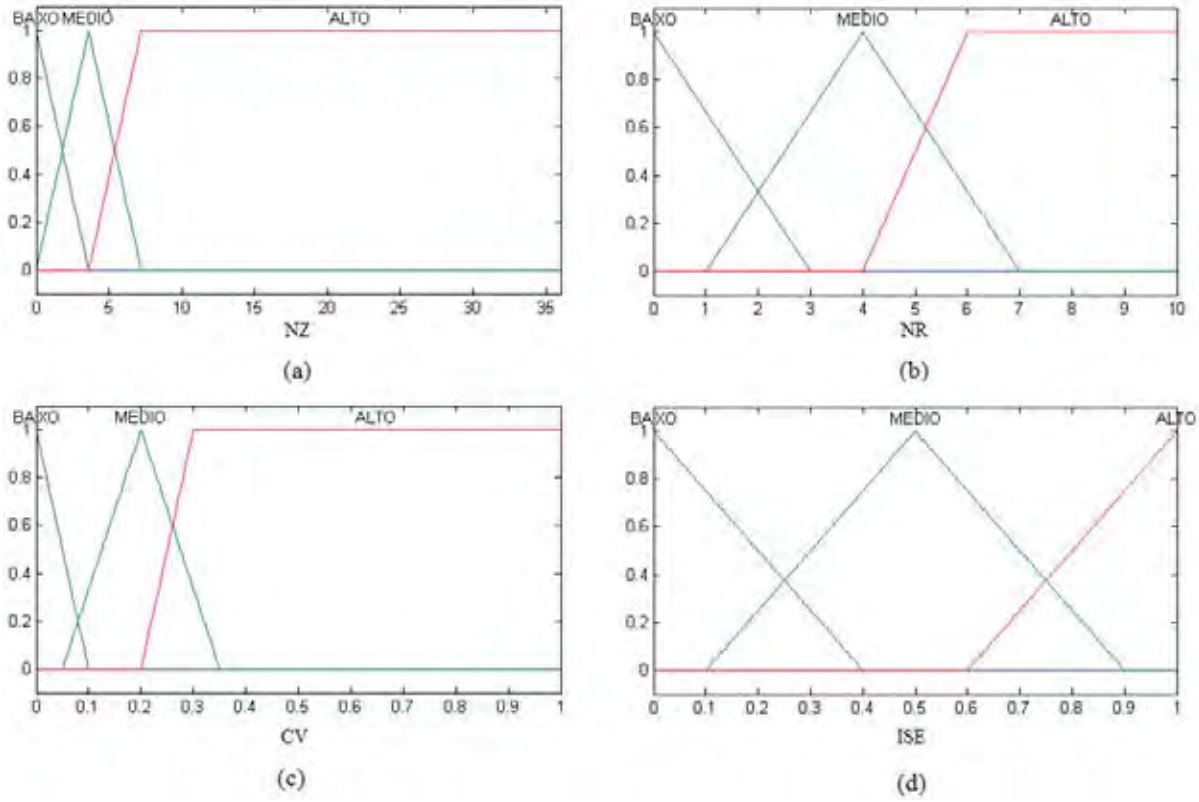
As Figuras 14 (a), (b), (c) e (d) contêm as partições *fuzzy* para as três variáveis de entrada e para a variável de saída, respectivamente. Tais partições mostram as funções de pertinência e as variáveis linguísticas pertencentes a cada variável *fuzzy* do SIF desenvolvido. Elas foram criadas a partir do conhecimento empírico do especialista em inspeções.

No Quadro 8, constam as principais características do SIF especialista desenvolvido tais como: tipo de inferência, funções de pertinência utilizadas e método de defuzificação.

Sabe-se que as regras que compõem a base de conhecimento do SIF especialista tornam-se ineficazes, pois o perfil de clientes com perdas comerciais, principalmente daqueles que praticam ilícitos sofre adaptações com o passar do tempo. Logo, as regras devem ser periodicamente atualizadas pelos especialistas.

A saída deste módulo será utilizada para o cálculo do *ISC* (seção 5.3.8).

Figura 14: Partições *fuzzy* do SIF especialista. (a), (b), (c) Variáveis de entrada. (d) Variável de Saída.



Fonte: Informações da pesquisa do autor.

Quadro 8: Parâmetros do SIF Especialista.

SIF	Mamdani
Funções de Pertinência	Triangulares e Trapezoidais
Inferência	*Max-Min*
Variáveis de Entrada	Três (*NZ*, *CV*, *NR*)
Variáveis de Saída	Uma (*ISE*)
Método de Defuzificação	Centróide

Fonte: Informações da pesquisa do autor.

5.3.7 Módulo de busca

Os módulos anteriores como o Módulo de Extração de Atributos Estatísticos, Módulo de Agrupamento, o Módulo de Classificação e o Módulo Especialista analisaram exclusivamente desvios e anomalias no histórico de consumo mensal de cada cliente e no histórico das inspeções realizadas.

Este módulo visa ampliar a análise através da pesquisa no banco de dados por atributos adicionais tais como o *IPCA* e o *ISA* que correspondem, respectivamente, às perdas comerciais percentuais do alimentador em relação às perdas globais e à vulnerabilidade da área que poderia favorecer o surgimento de fraudes, por exemplo. Esses dois atributos não são pontuais, são por região, por área na qual se encontram os clientes em análise. Eles compõem as entradas do Módulo Principal. Além do *IPCA* e do *ISA*, este módulo busca também características pontuais de cada cliente tais como o Consumo Irregular (*CI*) e o Religamento (*R*).

Outra busca interessante implementada neste módulo e não encontrada na literatura consultada refere-se à Lista de Nomes Suspeitos (*LNS*) e a Lista de Atividades Suspeitas (*LAS*). Informações como o nome completo e a atividade dos clientes flagrados em ilícitos durante inspeções em campo são armazenadas nas *LNS* e *LAS*. Durante a análise de cada cliente, é feita uma busca para averiguar se o sobrenome e a atividade de cada cliente constam nessas listas suspeitas. A presença do cliente na *LNS* e/ou na *LAS* não garante que ele pratique ilícito, mas é um indício a mais. Por exemplo: um cliente hipotético chamado *Jota Eme Beltrano* está sendo analisado neste instante. O sobrenome *Beltrano* consta na *LNS*, isto é, alguém portador desse sobrenome e, provavelmente, parente do senhor *Jota Eme*, foi surpreendido em uma das inspeções em campo por prática de ilícito. É plenamente plausível inferir que o suposto parente do senhor *Jota Eme* tenha comunicado a ele que existe uma forma bastante viável de “economizar energia”. Isso não garante que o senhor *Jota Eme* esteja praticando ilícito como seu parente, mas é um indício a mais que o torna mais suspeito que os demais clientes. São excluídos da *LNS*, os sobrenomes mais comuns tais como Silva, Oliveira, etc.

O raciocínio para a *LAS* é análogo ao da *LNS*. Supõem-se, por exemplo, que em um bairro foi flagrado um cliente comercial, um açougue com uma fraude embutida, por exemplo. Nesse instante a atividade “açougue” é inserida na *LAS* e os demais clientes comerciais dessa área cuja atividade é “açougue” possuem maior indício de igualmente estarem praticando ilícito. É plausível inferir isso, pois comerciantes do mesmo ramo, em geral, mantêm relações entre si e a prática de fraude por um deles impele os demais a praticarem também por

uma questão de sobrevivência no mercado, já que a prática de fraude implica em uma grande economia financeira.

Os parâmetros *CI*, *R*, *LNS* e *LAS* obtidos neste módulo são binários e entram no cálculo do *ISC* a ser abordado no Módulo Principal.

Para implementação deste módulo, foram utilizados *drivers* do tipo ODBC para conectar o programa desenvolvido no MATLAB® 7 ao banco de dados desenvolvido no programa Microsoft Office Access® 2010 no qual estão armazenados todos os dados de entrada detalhados na seção 5.3.1.

5.3.8 Módulo Principal

Este módulo final pretende reunir todos os parâmetros obtidos nos módulos anteriores que são pontuais, isto é, dizem respeito somente ao cliente em análise. A união desses parâmetros produz o *ISC* (Índice de Suspeita do Cliente). Logo, o valor do *ISC* depende unicamente das características particulares de cada cliente. O *ISC* é obtido por meio da expressão (5.4). P_1 pondera o parâmetro obtido a partir do conhecimento especialista acerca do histórico de consumo mensal; P_2 pondera os parâmetros intrínsecos e não óbvios obtidos através das redes neurais SOM e PMC a partir do histórico de consumo mensal e de inspeções e, em último, P_3 pondera os parâmetros obtidos a partir das buscas nos dados cadastrais e no histórico de inspeções. Esses pesos são arbitrados empiricamente pelo especialista conforme o seu julgamento acerca da confiabilidade dos dados de entrada. Por exemplo, se o histórico de inspeções, não estiver confiável, estiver desatualizado ou com poucos registros, o peso P_2 deve ser menor que os demais, pois a integridade do histórico de inspeções é crucial para o bom desempenho dos módulos de agrupamento e de classificação.

$$ISC = \frac{P_1(ISE) + P_2\left(\frac{GSA + PMC}{2}\right) + P_3\left(\frac{LNS + LAS + CI + R}{4}\right)}{\sum_{i=1}^{3} P_i} \tag{5.4}$$

A seguir, a descrição de cada uma das variáveis da expressão (5.4):

ISC: Índice de Suspeita do cliente;

P_1: peso referente aos parâmetros obtidos a partir da saída do Módulo Especialista;

P_2: peso referente aos parâmetros obtidos nos Módulos de Agrupamento e de Classificação;

P_3: peso referente às saídas do Módulo de Busca;

GSA: Grau de Suspeita do Agrupamento (Módulo de Agrupamento);

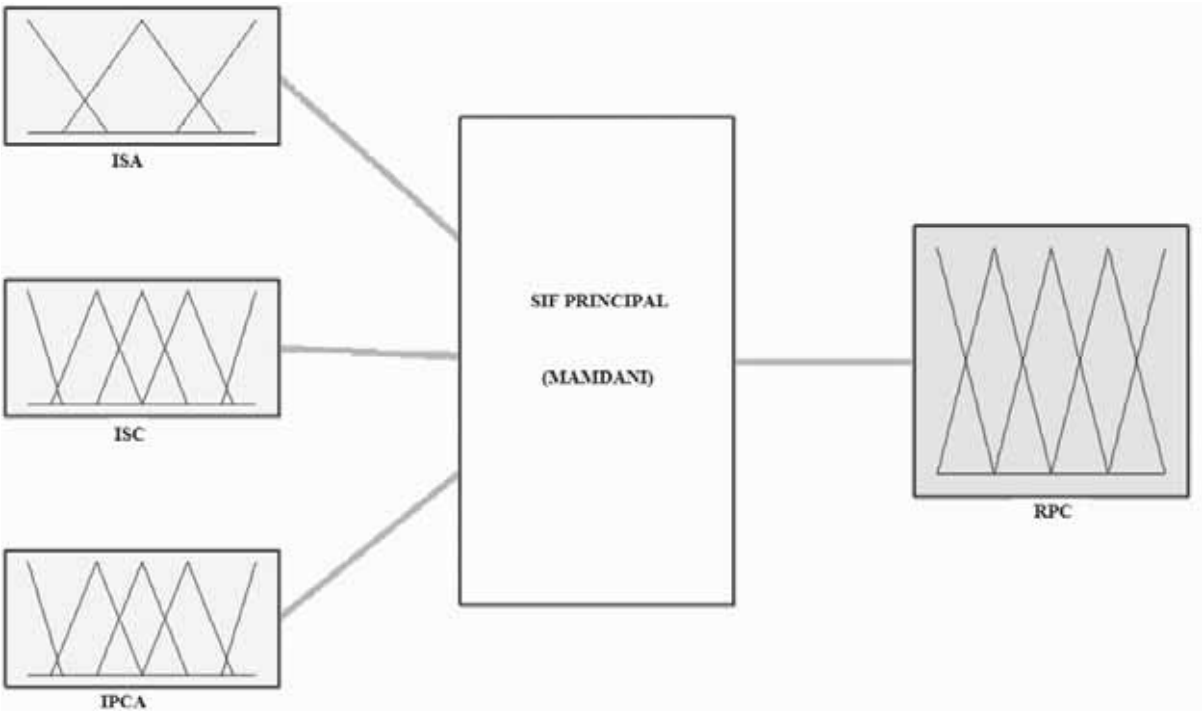
PMC: saída da rede neural PMC (Módulo de classificação);
ISE: saída do SIF Especialista (Módulo Especialista);
LNS: Lista de Nomes Suspeitos (Módulo de Busca);
LAS: Lista de Atividades Suspeitas (Módulo de Busca);
CI: Consumo Irregular (Módulo de Busca);
R: Religamento (Módulo de Busca).

Ao *ISC* são adicionados outros dois parâmetros: o *ISA* e o *IPCA*. Esses três atributos constituem as entradas do SIF Principal esboçado na Figura 15. A saída desse módulo e da metodologia é o *RPC* (Risco de Perda Comercial) que é um número no intervalo [0,1] que indica a possibilidade de um dado cliente de possuir perda comercial.

Ao contrário do *ISC*, o *ISA* e o *IPCA* são parâmetros não pontuais, por área. O *IPCA* é um parâmetro determinístico e depende unicamente da eficiência do método utilizado para cálculo das PTs (Perdas Técnicas) e o *ISA* é um parâmetro puramente empírico e depende da experiência e conhecimento do especialista a respeito das características de consumo dos clientes em cada área ou região.

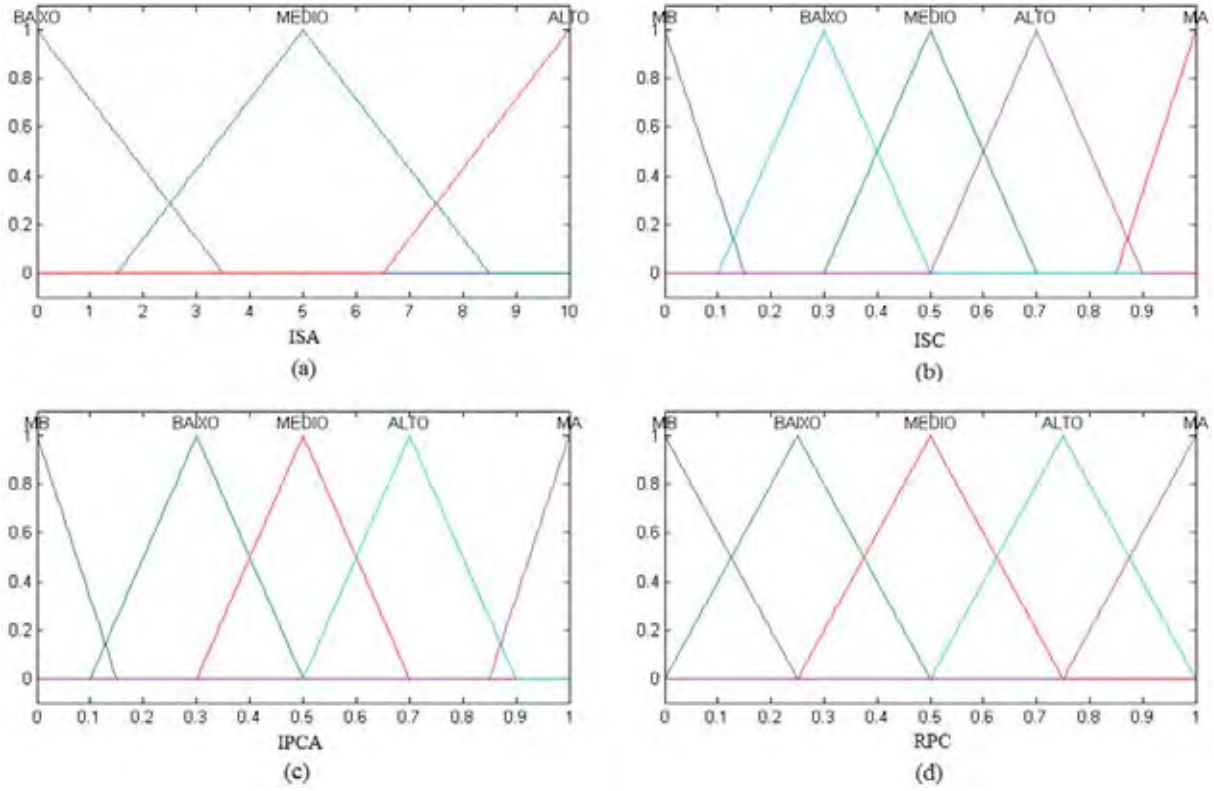
Nas Figuras 16 (a), (b), (c) e (d) exibem-se as partições *fuzzy* das três variáveis de entrada e da variável de saída, respectivamente. Elas foram criadas a partir do conhecimento prático do especialista em inspeções.

Figura 15: Visão geral do SIF principal.



Fonte: Informações da pesquisa do autor.

Figura 16: Partições *fuzzy* do SIF principal. (a), (b), (c) Variáveis de entrada. (d) Variável de Saída.



Fonte: Informações da pesquisa do autor.

5.4 CONSIDERAÇÕES FINAIS

A construção desta metodologia para detecção perdas comerciais em cada cliente do SDEE visou utilizar todas as entradas que se tem disponível bem como unir o conhecimento do especialista com informações intrínsecas obtidas via técnicas oriundas da área de SI (Sistemas Inteligentes). Adicionalmente, foram realizadas buscas na base de dados a respeito das perdas comerciais do alimentador e da vulnerabilidade da região em que se encontram os clientes. Nesse ponto, algumas considerações são pertinentes:

- Não se sabe com precisão quais são todas as regras que determinam um perfil suspeito. Nesse contexto, surge a necessidade de utilizar metodologias oriundas da área de SI a fim de descobrir relações ocultas entre os dados de entrada que possivelmente identifiquem um cliente com perdas comerciais.
- A base de regras é dinâmica e deve ser atualizada constantemente através, por exemplo, do conhecimento advindo das inspeções realizadas em campo.
- A metodologia apresentada possui um caráter empírico muito forte. Portanto, a habilidade e experiência do especialista para elaboração das funções de pertinên-

cia, da base de regras e para calibração dos parâmetros é crucial para o funcionamento adequado da metodologia proposta.

No capítulo subsequente, a metodologia exposta neste capítulo será implementada e testada. Os dados de entrada são reais e provêm de um alimentador da rede de distribuição de energia elétrica.

6 TESTES E RESULTADOS

Neste capítulo, as simulações foram executadas por classes de consumo e por alimentador. Cada um dos módulos que compõem a metodologia exposta no capítulo 5 foi programado utilizando o MATLAB® 7. O banco de dados foi desenvolvido no sistema de gerenciamento de banco de dados Microsoft Office Access® 2010. Foram utilizados *drivers* ODBC para conectar o programa desenvolvido no MATLAB® 7 ao banco de dados. Os testes foram realizados em um processador AMD Atholon™ II Dual-Core; 2,1 GHz e 3 GB de memória RAM.

Os dados completos do sistema de distribuição são: dados cadastrais, o histórico de consumo mensal e alguns dados técnicos adicionais como o alimentador[7] da rede de distribuição a que pertence cada cliente. Esses dados são reais e pertencem a um município interior do Estado de São Paulo. Trata-se de uma cidade de porte médio com aproximadamente 200.000 habitantes e de 560 km^2 de área. Possui 81.000 UCs (Unidades Consumidoras) distribuídas em sete classes de consumo: residencial (88,03%), comercial (9,46%), industrial (1,65%), rural (0,34%), iluminação pública (0,42%), serviço público (0,07%) e consumo próprio (0,02%).

Os dados cadastrais e o histórico de consumo mensal são reais. Esse último compreende 36 meses e inicia-se em novembro de 2003 até outubro de 2006. O histórico de inspeções, o *IPCA* (Índice de Perdas Comerciais do Alimentador) e o *ISA* (Índice de Suspeita da Área) foram arbitrados coerentemente com o intuito de testar a metodologia proposta.

As simulações e testes apresentados foram realizados em clientes das classes residencial, comercial e industrial que possuem a maior parte das perdas comerciais e nas quais a detecção e localização precisa das mesmas são mais difíceis (DANTAS, 2006).

6.1 CONSTRUÇÃO DOS HISTÓRICOS DE INSPEÇÕES

Entre os dados reais obtidos não há histórico de inspeções. Dessa forma, a fim de testar a metodologia proposta, construíram-se históricos de inspeções. São compostos por cinco grupos distintos de clientes comumente encontrados em históricos de inspeções reais. Eles estão apresentados no Quadro 9 no qual estão descritos os estados das variáveis de entrada para cada grupo. *"B"* significa baixo índice de suspeita, isto é, uma entrada que correspon-

[7] Nos dados reais, é fornecida a subestação a qual cada cliente está vinculado. Neste capítulo, utiliza-se a terminologia alimentador em lugar de subestação para testar a metodologia, porque ela foi criada para avaliar clientes por alimentadores da rede de distribuição de energia elétrica.

de a um cliente normal e "*A*" significa alto índice de suspeita, isto é, uma entrada característica de um cliente com algum problema que resulta em acréscimo das perdas comerciais. Se um cliente possuir "*B*" no campo *Dados Cadastrais*, por exemplo, significa que as variáveis de entrada referentes aos *Dados Cadastrais* (seção 5.3.1) desse cliente estão normais e correspondem a um cliente comum, normal, aparentemente sem qualquer problema que implique em elevação das perdas comerciais. Os quatro primeiros grupos correspondem a clientes com alguma anormalidade. O último grupo é composto por clientes normais. Eles são descritos a seguir:

- Grupo 1: Grupo de clientes suspeitos com extrema dificuldade para serem detectados. Todos os indicadores pontuais (os quais tem maior poder preditivo) são do tipo *"B"*. Mesmo esse grupo tendo todos os indicadores por área (os quais possuem menor poder preditivo) do tipo *"A"* não reduz a grande dificuldade de detectar os clientes vinculados a ele. Este grupo simula UCs com fraudes muito bem disfarçadas, como exemplo, novos loteamentos nos quais existem fraudes embutidas já previamente instaladas. Os clientes desse grupo são denominados cliente irregular com fraude. Para a maior parte das referências consultadas, os clientes desse grupo certamente seriam considerados totalmente normais.
- Grupo 2: Grupo de clientes suspeitos cuja detecção é muito difícil. Os clientes desse grupo possuem *Dados Cadastrais "B"* e *Histórico de Consumo Mensal "A"*. *Dados Cadastrais "A"* indica que o cliente possui um histórico de irregularidades e sabe-se que, em geral, os fraudadores são reincidentes e sempre buscam formas mais elaboradas de furtar energia a fim de não serem descobertos. No entanto, esse não é o caso dos clientes desse grupo os quais possuem *Dados Cadastrais "B"*. Portanto, este grupo simula clientes com medidor avariado, por exemplo, que não têm intenção de lesar a distribuidora de energia. Os clientes desse grupo são denominados cliente irregular com defeito.
- Grupo 3: Grupo de clientes suspeitos cuja detecção é difícil. É composto por clientes propensos a realizar fraudes (*Dados Cadastrais* "*A*"). No entanto, tais fraudes são muito bem camufladas, pois a fraude não causa grandes alterações no histórico de consumo mensal (*Histórico de Consumo Mensal* "*B*"). Devido a isso, os módulos que implementam redes neurais não conseguem prever anormalidades para UCs desse grupo. Os clientes desse grupo são denominados cliente irregular com fraude.

- Grupo 4: Grupo de clientes suspeitos e mais fáceis de serem detectados. Possui clientes propensos a realizar fraudes e, provavelmente, já o fizeram anteriormente. Todos os indicadores pontuais são *"A"*. Os clientes desse grupo são denominados cliente irregular com fraude.
- Grupo 5: Grupo composto por clientes normais e cujos indicadores são todos *"B"*, exceto o *IPCA* que é *"A"*.

Quadro 9: Composição dos dados de entrada para cada um dos cinco grupos que compõem o histórico de inspeções.

Grupos	Indicadores Pontuais		Indicadores por Área	
	Dados Cadastrais	Histórico de Consumo Mensal	Índice de Perdas Comerciais do Alimentador (IPCA)	Índice de Suspeita da Área (ISA)
1	B	B	A	A
2	B	A	A	B
3	A	B	A	B
4	A	A	A	B
5	B	B	A	B

Fonte: Informações da pesquisa do autor.

Os históricos de inspeções arbitrados neste capítulo possuem a mesma proporção comumente encontrada em históricos de inspeções reais, isto é, para cada dez clientes, nove são normais (pertencem ao grupo 5) e um cliente possui alguma anormalidade (pertence a um dos grupos restantes do Quadro 9). Os históricos de inspeções construídos são utilizados em todas as simulações realizadas neste capítulo em clientes das classes residencial, comercial e industrial.

Observa-se que o *IPCA* é igual a *"A"* em todos os grupos, porque, intuitivamente, os alimentadores com as maiores perdas globais são os primeiros a serem investigados.

Todas as simulações deste capítulo foram realizadas em UCs pertencentes ao alimentador *A5* cujo *IPCA* é 0,8. Tal alimentador possui 12.872 UCs conectadas a ele o que corresponde a 15,77% do total de UCs que há no município em análise.

Em resumo, os cinco grupos arbitrados possuem diferentes níveis de dificuldade de detecção. Pretende-se, portanto, avaliar o desempenho da metodologia proposta frente a cada um dos grupos criados.

6.2 CLIENTES RESIDENCIAIS

A quantidade de clientes em cada um dos grupos que compõem o histórico de inspeção utilizado na simulação para clientes residenciais consta no Quadro 10. No total, são 11.758 UCs residenciais pertencentes ao alimentador *A5*.

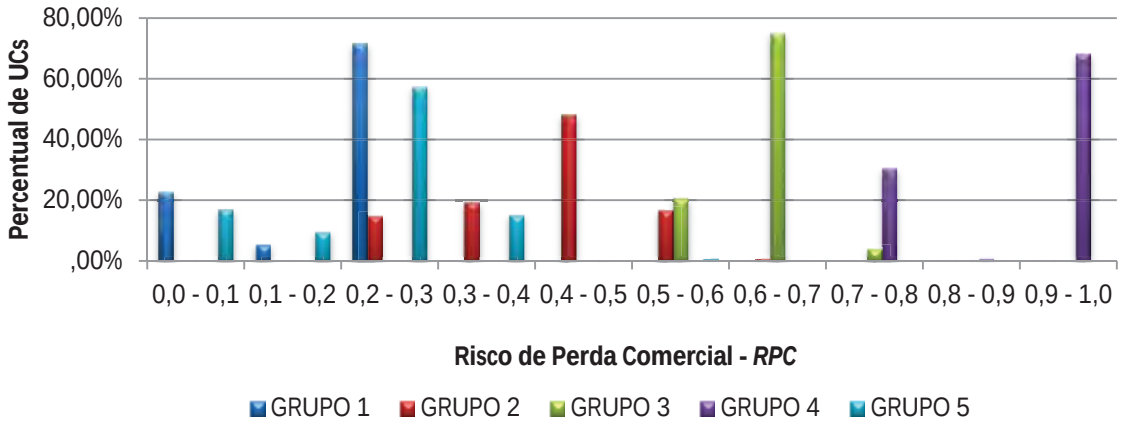
Quadro 10: Parâmetros dos grupos que compõem o histórico de inspeções para clientes residenciais.

Grupos	Nº UCs	Indicadores Pontuais		Indicadores por Área	
		Dados Cadastrais	Histórico de Consumo	Índice de Perdas Comerciais do Alimentador (IPCA)	Índice de Suspeita da Área (ISA)
1	294 (≈ 2,5%)	B	B	A	A
2	294 (≈ 2,5%)	B	A	A	B
3	294 (≈ 2,5%)	A	B	A	B
4	294 (≈ 2,5%)	A	A	A	B
5	10.582 (≈ 90%)	B	B	A	B

Fonte: Informações da pesquisa do autor.

A Figura 17 contém um histograma com a distribuição do *RPC* (Risco de Perda Comercial) para os grupos do Quadro 10. A maioria dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *RPC* nos seguintes intervalos: [0,2; 0,3], [0,4; 0,5], [0,6; 0,7], [0,9; 1,0] e [0,2; 0,3], respectivamente. Os grupos 1, 3, 4 e 5 possuem *RPC* satisfatório, ou seja, elevado para os três primeiros grupos (que são grupos suspeitos) e baixo para o último (grupo de clientes normais). Apenas o grupo 2 obteve um *RPC* mediano que é considerado baixo para um grupo composto por clientes suspeitos.

Figura 17: Distribuição do *RPC* em intervalos iguais para cada um dos grupos que compõem o histórico de inspeção dos clientes residenciais.



Fonte: Informações da pesquisa do autor.

Observa-se que o grupo 1, apesar de ser composto por clientes suspeitos e de ter obtido *a priori* um baixo *RPC*, diante da extrema dificuldade para detecção de suspeitos desse grupo, considera-se que o resultado foi igualmente satisfatório. Observa-se também que o grupo 4 exibe o maior *RPC*, porque tal grupo possui todos indicadores pontuais *"A"* conforme exposto no Quadro 9 (Seção 6.1).

Em ultimo, observa-se que a resposta da simulação computacional apresentada na Figura 17 está em concordância com a descrição dos cinco grupos (Seção 6.1) que compõem o histórico de inspeção dos clientes residenciais.

6.2.1 Análise dos resultados da simulação para clientes residenciais

Aproximadamente 91,35% das UCs (≈11.758) do alimentador *A5* são da classe residencial e representam 16,36% das UCs do município em análise o qual possui 71.870 clientes da classe residencial.

Todas as 11.758 UCs da classe residencial do alimentador *A5* consomem em média 21,495 GWh/ano. As 1.176 UCs pertencentes aos quatro grupos suspeitos consomem em média 1,825 GWh/ano. Assume-se que a perda comercial do alimentador seja igual ao consumo das UCs dos grupos suspeitos (o qual supostamente não foi faturado) e considera-se o *IPCA = 0,8*; a partir da expressão (5.1), tem-se 2,281 GWh/ano para as perdas globais sendo que 0,456 GWh/ano são perdas técnicas. Logo, anualmente, 10,47% da energia transportada pelo alimentador *A5* para clientes da classe residencial é perdida devido às perdas técnicas (1,98%) e às perdas comerciais (8,49%).

No Quadro 11 tem-se a indicação por grupo da quantidade de clientes residenciais com maior e menor *RPC*. Ao todo, há 703 clientes suspeitos com alto *RPC* (maior ou igual a 0,5), sendo 639 deles pertencentes a um dos grupos suspeitos e 64 deles são clientes normais. Portanto, 703 clientes residenciais serão inspecionados, sendo que 639 possuem perdas comerciais, isto é, 90,9% das UCs a serem visitadas possuem algum problema que eleve as perdas comerciais o que é uma alta taxa de sucesso (ver Quadro 3, Seção 2.3). As 639 UCs com perdas comerciais têm consumo médio supostamente não faturado de 0,9597 GWh/ano. Portanto, a detecção desses clientes implica em redução das perdas globais que passa de 10,47% para 6,15% em relação à energia total transportada pelo alimentador *A5*.

A tarifa para consumidores residenciais com consumo entre 101 e 220 kWh é de 0,248 R$/kWh. Logo, em termos financeiros, a recuperação de 0,9597 GWh/ano representa um acréscimo na receita da concessionária de, em média, R$ 238 milhões/ano. Desse valor,

R$ 28,56 milhões/ano será recolhido na forma de imposto – ICMS – para uma alíquota de 12%.

Quadro 11: Eficiência da metodologia em relação aos clientes residenciais analisados.

Grupos	*RPC* < *0,5*		*RPC* ≥ *0,5* (Inspecionar)	
	Nº de UCs	UCs %	Nº de UCs	UCs %
1	294	100,0%	0	0,0%
2	243	82,7%	51	17,3%
3	0	0,0%	294	100,0%
4	0	0,0%	294	100,0%
5	10.518	99,4%	64	0,6%

Fonte: Informações da pesquisa do autor.

Observa-se que todos os clientes suspeitos do Grupo 1 do Quadro 11 possuem baixo *RPC*. Isso é porque os clientes desse grupo têm indicadores pontuais típicos de clientes normais, por isso, sua detecção é muito difícil.

Destaca-se a alta confiabilidade da metodologia, pois ela atribuiu baixo *RPC* a 99,4% dos clientes pertencentes ao grupo 5 (clientes normais); logo, se o programa atribuir alto *RPC* a um cliente, há uma grande chance de que o mesmo possua algum problema que eleve as perdas comerciais na rede de distribuição.

6.2.2 Análise dos módulos da metodologia para clientes residenciais

O valor do *RPC*, que representa a resposta da simulação, conforme Figura 11 (seção 5.3.1), depende de três variáveis: o *ISA*, o *IPCA* e o *ISC*. O *ISA* é uma variável muito ampla e seu valor depende do conhecimento empírico do especialista a respeito dos clientes da região em análise. O *IPCA* é um índice determinístico e seu valor depende da metodologia adotada para cálculo das perdas técnicas[8]. Em contrapartida, o desempenho das redes neurais PMC e SOM e do SIF especialista além dos dados cadastrais dos clientes refletem diretamente no valor do *ISC* (expressão 5.4, seção 5.3.8). Admite-se que os dados cadastrais dos clientes estão íntegros e atualizados. Nesse contexto, o valor satisfatório do *RPC* depende dos índices: *PMC* (saída da rede neural PMC – Módulo de Classificação), *GSA* (Grau de Suspeita do Agrupamento – Módulo de Agrupamento) e do *ISE* (Índice de Suspeita Especialista – Módulo Especialista) os quais serão avaliados nesta seção.

[8] Na literatura consultada, a maioria dos trabalhos obtêm as perdas comerciais a partir das perdas técnicas obtidas *a priori*. No entanto, Dantas (2006), por exemplo, estima as perdas comercias sem cálculo prévio das perdas técnicas.

O desempenho do Módulo de Classificação pode ser avaliado a partir de métricas extraídas da matriz de confusão (seção 3.4): a especificidade e a confiabilidade negativa. A especificidade proporciona uma noção da cobertura do sistema classificador, isto é, o percentual do conjunto de fraudadores que o sistema consegue identificar. A confiabilidade negativa proporciona uma noção de precisão das inspeções indicadas, isto é, o percentual de sucessos na identificação de reais fraudadores em relação ao total de inspeções recomendadas. Um classificador com alta especificidade e baixa confiabilidade possui boa cobertura, mas baixa taxa de sucesso nas inspeções. Em contrapartida, um classificador com alta confiabilidade e baixa especificidade possui boa taxa de sucesso, mas não identifica uma parcela significativa dos fraudadores. É preferível, portanto, ter uma distribuição equilibrada de desempenho entre essas duas métricas.

No Quadro 12, tem-se a matriz de confusão (seção 3.4) referente ao Módulo de Classificação para os clientes residenciais. As expressões (6.1), (6.2) e (6.3) representam, respectivamente, a confiabilidade negativa, a especificidade e a taxa de acerto.

O Módulo de Classificação obteve valores altos para a confiabilidade negativa e para a taxa de acerto e um valor razoável para a especificidade. Esse valor inferior para a especificidade foi devido ao histórico de consumo mensal dos clientes residenciais normais e anômalos serem muito próximos o que confere maior dificuldade ao Módulo de Classificação. Além disso, o sistema foi projetado de tal forma que possua a maior confiabilidade possível com o objetivo de minimizar inspeções em UCs normais. Isso implica em uma pequena queda da especificidade, porque tais métricas são inversamente proporcionais, ou seja, uma elevação da confiabilidade negativa ocasiona uma queda na especificidade. Em último, em geral, as redes neurais possuem maior confiabilidade negativa do que especificidade (COMETTI; VAREJÃO, 2005; FERREIRA, 2008).

Quadro 12: Matriz de confusão para o Módulo de Classificação referente à simulação dos clientes residenciais.

Matriz de Confusão		Classe Predita	
		N	F
Classe Real	N	$Q_{NN} = 11.099$	$Q_{NF} = 71$
	F	$Q_{FN} = 191$	$Q_{FF} = 397$

Fonte: Informações da pesquisa do autor.

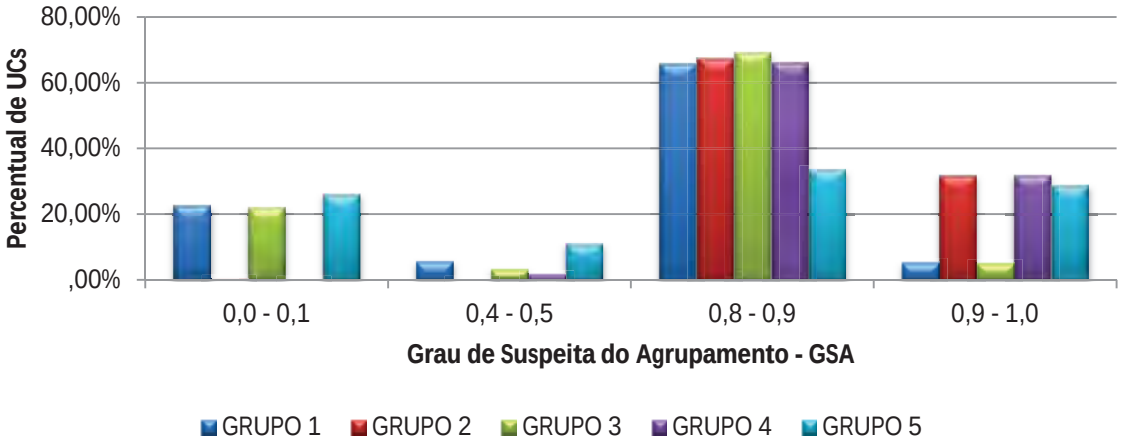
$$c = \frac{Q_{FF}}{Q_{FF} + Q_{NF}} = \frac{397}{397 + 71} = 0{,}848 \quad (6.1)$$

$$e = \frac{Q_{FF}}{Q_{FF} + Q_{FN}} = \frac{397}{397 + 191} = 0{,}675 \quad (6.2)$$

$$Taxa\_Acerto = \frac{Q_{NN} + Q_{FF}}{Q_{NN} + Q_{FF} + Q_{NF} + Q_{FN}} = \frac{11.099 + 397}{11.099 + \ 397 + 71 + 191} = 0{,}978 \quad (6.3)$$

O desempenho do Módulo de Agrupamento pode ser avaliado a partir da Figura 18 que representa o *GSA* de todos os grupos de clientes residenciais simulados. A maior parte dos clientes pertencentes aos grupos 1, 2, 3, 4 e 5 possuem *GSA* no intervalo [0,8; 0,9]. Destaca-se a homogeneidade do histórico de consumo dos clientes residenciais simulados, porque a maioria das UCs de cada um dos grupos tem o *GSA* encerrado em um único intervalo. Tal comportamento é típico de regiões em que há baixas perdas comerciais.

Figura 18: Distribuição do *GSA* para cada grupo que compõe o histórico de inspeção dos clientes residenciais.



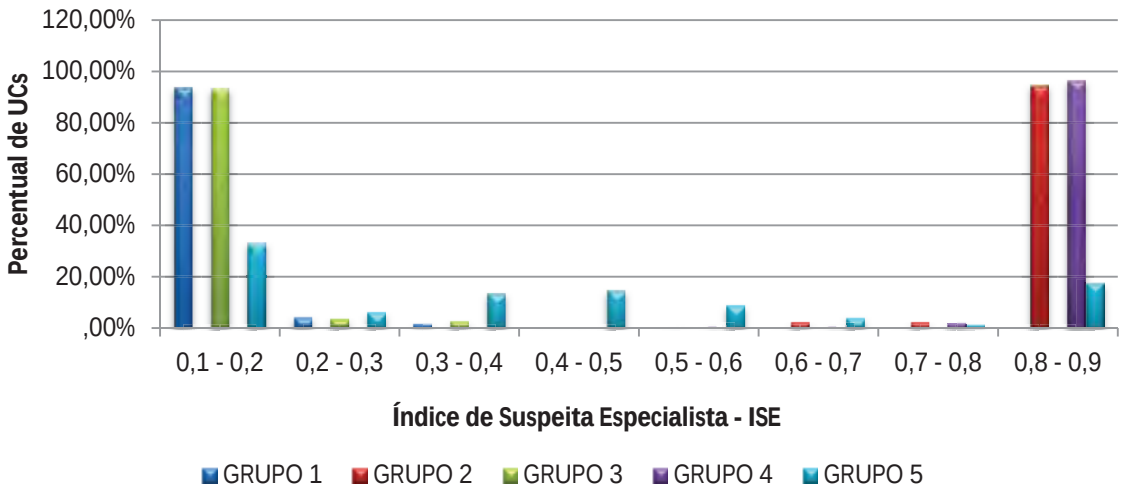
Fonte: Informações da pesquisa do autor.

O desempenho do Módulo de Especialista é avaliado a partir da Figura 19 que contém os *ISE* de todos os grupos simulados. A maior parte dos clientes pertencentes aos grupos 1, 2, 3, 4 e 5 possui o *ISE* nos intervalos [0,1; 0,2], [0,8; 0,9], [0,1; 0,2], [0,8; 0,9] e $[0{,}3;\ 0{,}5] \cup [0{,}8;\ 0{,}9]$, respectivamente.

O *ISE* avalia o grau de anomalia de históricos de consumo mensais. Os grupos portadores de históricos de consumo anômalos (grupos 2 e 4 – ver Quadro 9, Seção 6.1) têm alto *ISE*; os demais, possuem baixo *ISE*. Apenas uma parte das UCs do grupo 5 possuem alto

ISE devido à grande homogeneidade dos históricos de consumo. Portanto, a saída do Módulo Especialista é coerente com as características de cada um dos grupos criados.

Figura 19: Distribuição do *ISE* para os grupos de clientes residenciais.



Fonte: Informações da pesquisa do autor.

6.3 CLIENTES COMERCIAIS

As simulações com clientes da classe comercial foram realizadas em 752 UCs do alimentador *A5*. O Quadro 13 contém o número de clientes em cada um dos grupos que compõem o histórico de inspeção utilizado na simulação para clientes comerciais.

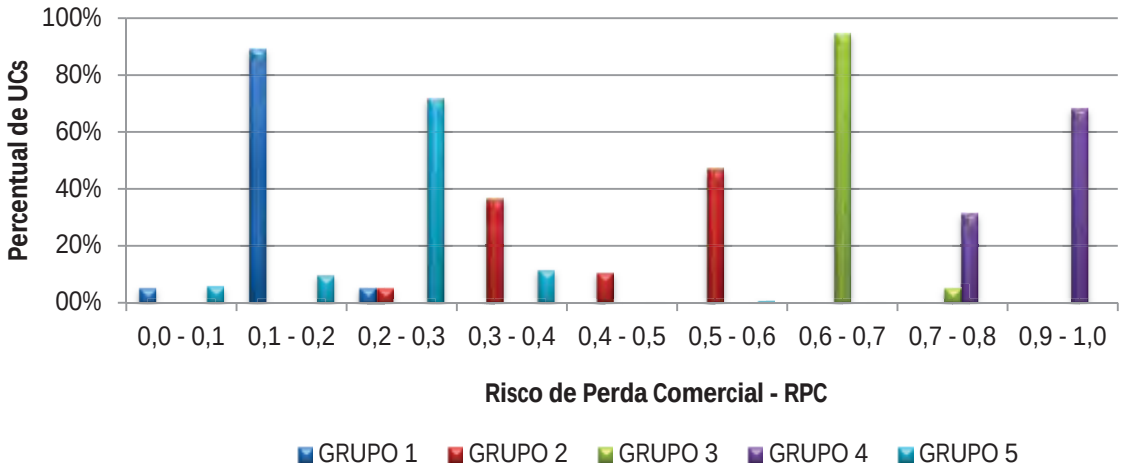
Na figura 20, há um histograma com a distribuição do *RPC* para cada um dos grupos do Quadro 13. A maior parte dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *RPC* nos intervalos [0,1; 0,2], [0,5; 0,6], [0,6; 0,7], [0,9; 1,0] e [0,2; 0,3], respectivamente. Todos os grupos possuem *RPC* em conformidade as características de cada grupo exibidas no Quadro 13. A análise desses resultados é análoga àquela realizada na Seção 6.2 para os clientes da classe residencial.

Quadro 13: Parâmetros dos grupos que compõem o histórico de inspeções para clientes comerciais.

Grupos	Nº de UCs	Indicadores Pontuais		Indicadores por Área	
		Dados Cadastrais	Histórico de Consumo	Índice de Perdas Comerciais do Alimentador (IPCA)	Índice de Suspeita da Área (ISA)
1	19 (≈ 2,5%)	B	B	A	A
2	19 (≈ 2,5%)	B	A	A	B
3	19 (≈ 2,5%)	A	B	A	B
4	19 (≈ 2,5%)	A	A	A	B
5	676 (≈ 90,0%)	B	B	A	B

Fonte: Informações da pesquisa do autor

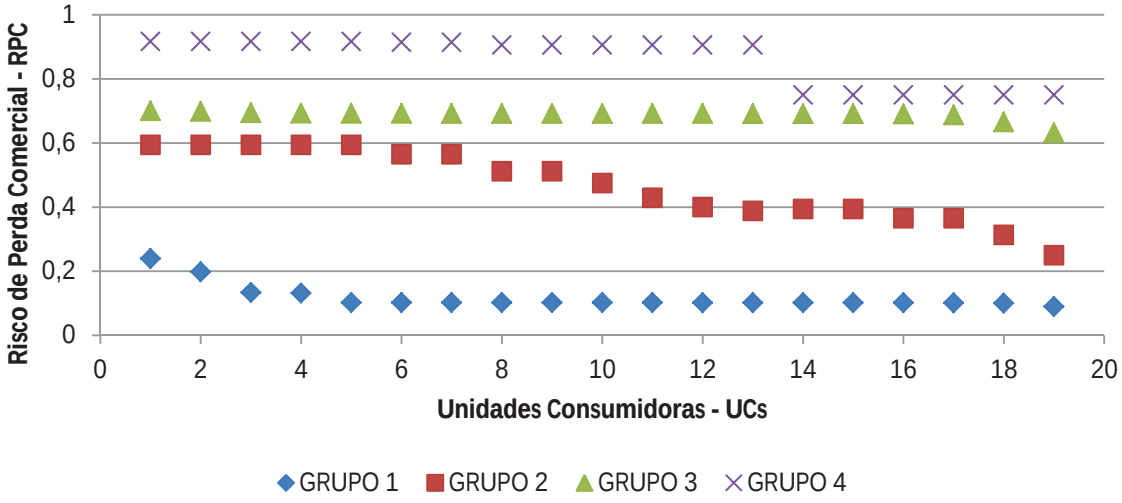
Figura 20: Risco de Perda Comercial por intervalos de todos os grupos de UCs comerciais.



Fonte: Informações da pesquisa do autor.

Na Figura 21 tem-se o *RPC* apenas para os clientes dos grupos suspeitos. Observa-se visualmente que os grupos que possuem maior número de indicadores pontuais anômalos exibem maior *RPC* e vice-versa.

Figura 21: Risco de Perda Comercial para as UCs dos grupos suspeitos da classe comercial.



Fonte: Informações da pesquisa do autor.

6.3.1 Análise dos resultados da simulação para clientes comerciais

Aproximadamente 5,84% (752) das UCs do alimentador *A5* são da classe comercial e representam 9,74% das UCs comerciais do município em análise. As 752 UCs da classe comercial conectadas ao alimentador *A5* consomem em média 8,602 GWh/ano. As 76 UCs suspeitas, isto é, que pertencem aos Grupos 1, 2, 3 e 4 consomem em média 0,762 GWh/ano. A partir da expressão (5.1), considerando que toda a energia consumida pelos clientes suspei-

tos não foi faturada e assumindo *IPCA = 0,8*; tem-se um valor de 0,952 GWh/ano para as perdas globais no alimentador sendo 0,19 GWh/ano referente às perdas técnicas. Portanto, a cada ano, em média, 11,07% da energia transportada pelo alimentador *A5* para os clientes da classe comercial é perdida na forma de perdas técnicas (2,21%) e de perdas comercias (8,86%).

No Quadro 14 tem-se a indicação dos clientes com maior *RPC* para cada um dos cinco grupos de clientes comerciais simulados. Ao todo há 52 clientes suspeitos com alto *RPC* (maior ou igual a 0,5), sendo 47 deles pertencentes a um dos grupos suspeitos do Quadro 13 e 05 deles são clientes normais. Logo, de todos os 52 clientes com alto *RPC* os quais serão inspecionados (6,9% dos clientes comerciais do alimentador *A5*), 47 possuem perdas comerciais, isto é, 90,4% dos clientes a serem visitados possuem perdas comerciais o que representa uma alta taxa de sucesso. Esses 47 clientes possuem consumo médio de 0,297 GWh/ano. Supõe-se que toda a energia consumida por tais clientes suspeitos não é faturada. Dessa forma, a descoberta desses clientes implica em uma redução das perdas globais do alimentador *A5* de 11,07% para 6,99% em relação ao consumo dos clientes da classe comercial.

Quadro 14: Eficiência da metodologia em relação aos clientes comerciais analisados.

Grupos	***RPC* < *0,5***		***RPC* ≥ *0,5* (Inspecionados)**	
	Nº de UCs	**UCs %**	**Nº de UCs**	**UCs %**
1	19	100,0%	0	0,0%
2	10	52,6%	09	47,4%
3	0	0,0%	19	100,0%
4	0	0,0%	19	100,0%
5	671	99,3%	05	0,7 %

Fonte: Informações da pesquisa do autor.

Observa-se que todos os clientes do Grupo 1 do Quadro 14 têm baixo *RPC*. Isso é porque as UCs desse grupo têm indicadores pontuais típicos de clientes normais, por isso, são muito dificilmente detectadas.

Destaca-se também a alta confiabilidade da metodologia, pois ela atribuiu baixo *RPC* a 99,3% dos clientes pertencentes ao grupo 5 (clientes normais); logo, se o programa atribuir alto *RPC* a um cliente, há grande chance de que o mesmo seja um cliente anômalo.

Em último, observa-se que o faturamento de grandes clientes, isto é, clientes pertencentes ao grupo A é mais complexo e depende das especificidades contratuais de cada cliente, do horário de consumo, da demanda contratada, etc. Em geral, inúmeras UCs da classe

comercial pertencem ao grupo A. Por isso, não será feita uma estimativa de economia financeira ocasionada pelo combate às perdas comerciais como foi feito para os clientes da classe residencial cuja política tarifária é mais simplificada.

6.3.2 Análise dos módulos da metodologia para clientes comerciais

Esta seção visa avaliar os parâmetros *PMC*, *GSA* e *ISE* que influenciam diretamente no valor do *ISC* e, consequentemente, na eficácia das simulações para os clientes comerciais. No Quadro 15, apresenta-se a matriz de confusão referente ao Módulo de Classificação para os clientes comerciais.

Quadro 15: Matriz de confusão para o Módulo de Classificação referente à simulação dos clientes comerciais.

Matriz de Confusão		Classe Predita	
		N	F
Classe Real	N	$Q_{NN} = 708$	$Q_{NF} = 06$
	F	$Q_{FN} = 12$	$Q_{FF} = 26$

Fonte: Informações da pesquisa do autor.

As expressões (6.4), (6.5) e (6.6) representam, respectivamente, a confiabilidade negativa, a especificidade e a taxa de acerto para os dados do Quadro 15. Estes valores indicam que o Módulo de Classificação obteve desempenho satisfatório principalmente em relação à confiabilidade negativa e a taxa de acerto. Similarmente à análise anterior para os clientes residenciais observa-se novamente que a confiabilidade negativa é superior à especificidade o que é uma característica das redes neurais e da metodologia proposta.

$$c = \frac{Q_{FF}}{Q_{FF} + Q_{NF}} = \frac{26}{26 + 06} = 0{,}813 \tag{6.4}$$

$$e = \frac{Q_{FF}}{Q_{FF} + Q_{FN}} = \frac{26}{26 + 12} = 0{,}684 \tag{6.5}$$

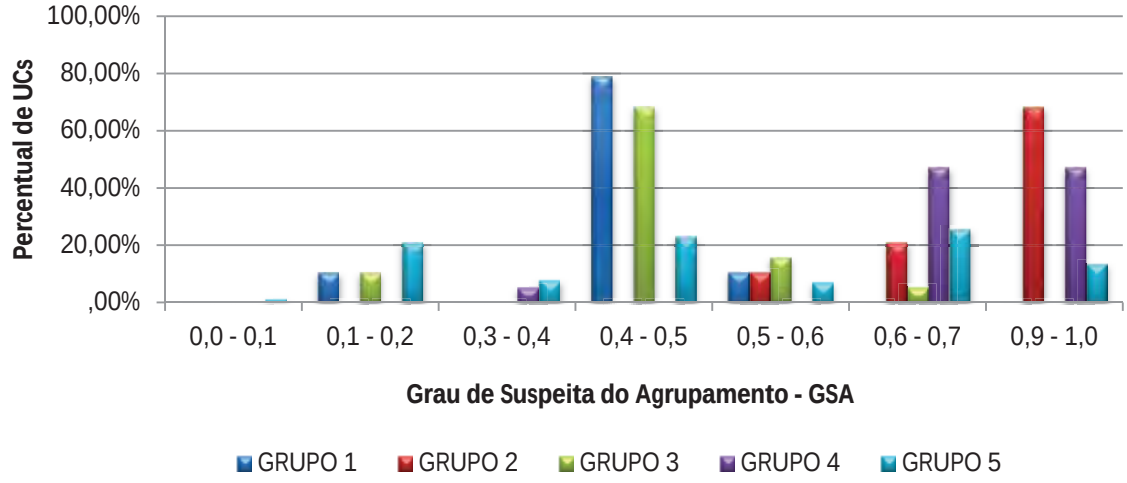
$$Taxa\_Acerto = \frac{Q_{NN} + Q_{FF}}{Q_{NN} + Q_{FF} + Q_{NF} + Q_{FN}} = \frac{708 + 26}{708 + \ 26 + 06 + 12} = 0{,}976 \tag{6.6}$$

O desempenho do Módulo de Agrupamento pode ser avaliado através da Figura 22 que representa os *GSA* de todos os grupos de clientes comerciais. A maior parte dos clien-

tes dos grupos 1, 2, 3, 4 e 5 possuem *GSA* nos intervalos [0,4; 0,5], [0,9; 1,0], [0,4; 0,5], [0,6; 0,7] ∪ [0,9; 1,0] e [0,1; 0,2] ∪ [0,4; 0,5] ∪ [0,6; 0,7], respectivamente.

Destaca-se o bom desempenho do Módulo de Agrupamento em relação às características dos grupos 1, 2, 3 e 4. Observa-se, no entanto, que parte do grupo 5 obteve alto *GSA* para um grupo composto por clientes normais.

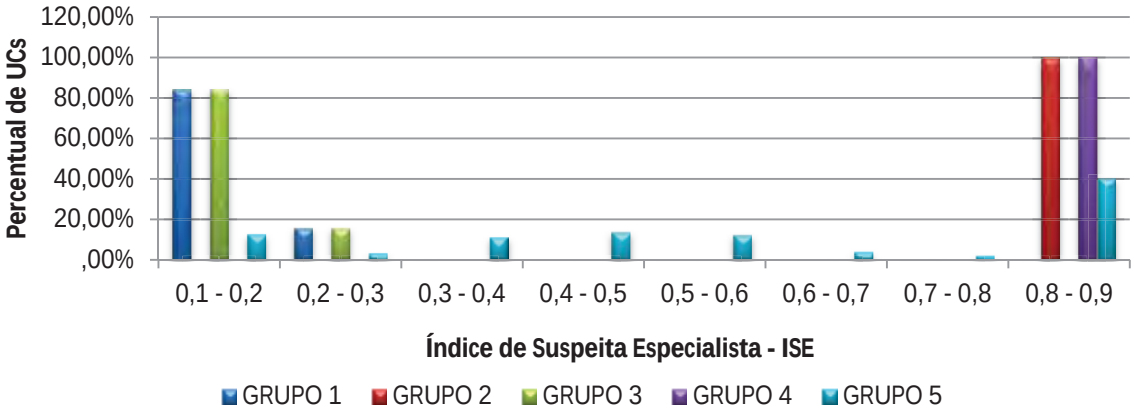
Figura 22: Distribuição do *GSA* para cada grupo que compõe o histórico de inspeção das UCs comerciais.



Fonte: Informações da pesquisa do autor.

O desempenho do Módulo de Especialista é avaliado a partir da Figura 23 que contém os *ISE* de todos os grupos de clientes comerciais. A parte majoritária dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *ISE* nos intervalos [0,1; 0,2], [0,8; 0,9], [0,1; 0,2], [0,8; 0,9] e [0,3; 0,6] ∪ [0,8; 0,9], respectivamente. Destaca-se o bom desempenho do Módulo Especialista dada a coerência do *ISE* obtido frente as características de cada grupo. No entanto, parte do grupo 5 (clientes normais) obteve alto *ISE* devido à grande homogeneidade do histórico de consumo mensal.

Figura 23: Distribuição do *ISE* para os grupos que compõem o histórico de inspeção dos clientes comerciais.



Fonte: Informações da pesquisa do autor.

6.4 CLIENTES INDUSTRIAIS

As simulações com clientes industriais foram realizadas com 294 UCs da classe industrial pertencentes ao alimentador *A5*. No Quadro 16, tem-se o número de clientes em cada um dos grupos que compõem o histórico de inspeção utilizado na simulação para os clientes industriais.

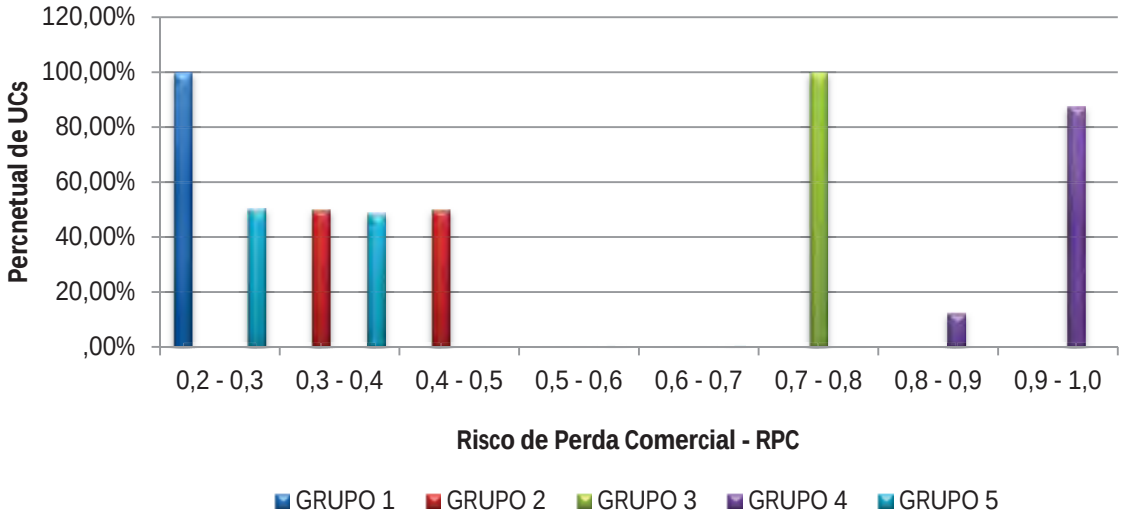
Quadro 16: Parâmetros dos grupos que compõem o histórico de inspeções para clientes industriais.

Grupos	Nº UCs	Indicadores Pontuais		Indicadores por Área	
		Dados Cadastrais	Histórico de Consumo	Índice de Perdas Comerciais do Alimentador (IPCA)	Índice de Suspeita da Área (ISA)
1	08 (≈ 2,7%)	B	B	A	A
2	08 (≈ 2,7%)	B	A	A	B
3	08 (≈ 2,7%)	A	B	A	B
4	08 (≈ 2,7%)	A	A	A	B
5	262 (≈ 89,2%)	B	B	A	B

Fonte: Informações da pesquisa do autor.

A Figura 24 contém o histograma com a distribuição do *RPC* para os grupos do Quadro 16. A maior parte dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *RPC* nos intervalos [0,2; 0,3], [0,3; 0,5], [0,7; 0,8], [0,9; 1,0] e [0,2; 0,4], respectivamente. O *RPC* obtido pelos grupos estão em conformidade com os parâmetros de cada grupo. A exceção é o grupo 2 cujo *RPC* deveria ser superior, porque tal grupo é composto por clientes anômalos.

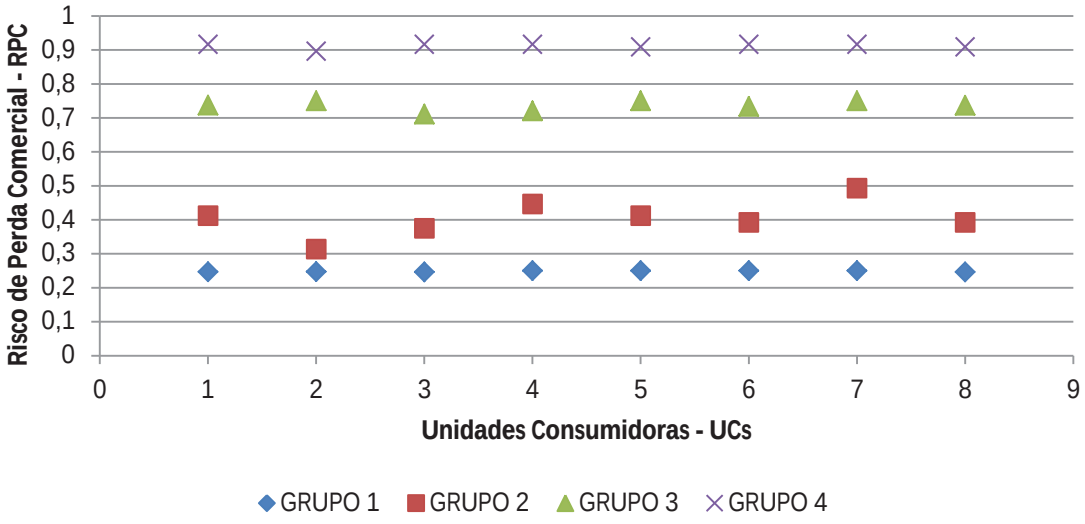
Figura 24: Risco de Perda Comercial por intervalos de todos os grupos de UCs industriais.



Fonte: Informações da pesquisa do autor.

Na Figura 25 tem-se o *RPC* apenas para os clientes industriais dos grupos suspeitos. Observa-se que os grupos que têm maior número de indicadores pontuais anômalos têm maior *RPC* e vice-versa.

Figura 25: Risco de Perda Comercial para as UCs dos grupos suspeitos da classe industrial.



Fonte: Informações da pesquisa do autor.

6.4.1 Análise dos resultados da simulação para clientes industriais

Aproximadamente 2,28% das UCs (294) do alimentador *A5* pertencem à classe industrial e representam 21,8% das UCs industriais do município em análise.

As 294 UCs da classe industrial pertencentes ao alimentador *A5* consomem em média 4,284 GWh/ano. Já as 32 UCs pertencentes aos grupos anômalos consomem em média 0,184 GWh/ano. A partir da expressão (5.1) obtém-se 0,23 GWh/ano para as perdas globais sendo 0,046 GWh/ano relativo às perdas técnicas. Para tal considera-se que toda a energia consumida pelos grupos anômalos não foi faturada (são perdas comerciais) e que o *IPCA = 0,8*.

Portanto, a cada ano, em média, 5,37% da energia transportada pelo alimentador *A5* para os clientes da classe industrial é perdida na forma de perdas técnicas (1,07%) e de perdas comercias (4,30%).

No Quadro 17 tem-se a indicação dos clientes com maior e com menor *RPC* para cada grupo de cliente industrial simulado. Ao todo existem 18 clientes industriais suspeitos com alto *RPC* (maior ou igual a 0,5), sendo 16 deles pertencentes a um dos grupos suspeitos do Quadro 16 e 02 deles são clientes normais. Logo, de todos os 18 clientes com alto *RPC* e que por isso serão inspecionados, 16 possuem perdas comerciais, ou seja, 88,8% dos clientes

a serem inspecionados possuem perdas comerciais o que representa uma alta taxa de sucesso.

Esses 16 clientes industriais que possuem perdas comerciais têm um consumo médio de 0,103 GWh/ano. Supõe-se que toda a energia consumida por tais clientes suspeitos não é faturada. Dessa forma, a descoberta desses clientes implica em uma redução das perdas globais do alimentador *A5* de 5,37% para 2,96% em relação ao consumo total dos clientes da classe industrial.

Quadro 17: Eficiência da metodologia em relação aos clientes industriais analisados.

Grupos	*RPC* < *0,5*		*RPC* ≥ *0,5* (Inspecionados)	
	Nº de UCs	UCs %	Nº de UCs	UCs %
1	08	100,0%	0	0,0%
2	08	100%	0	0,0%
3	0	0,0%	08	100,0%
4	0	0,0%	08	100,0%
5	260	99,2%	02	0,8 %

Fonte: Informações da pesquisa do autor.

A partir do Quadro 17 observa-se novamente a grande confiabilidade da metodologia, porque 99,2% das UCs do grupo 5 têm *RPC* baixo (inferior a 0,5). No entanto, todos as UCs do grupo 2 têm erroneamente baixo *RPC* para um grupo composto por clientes com perdas comerciais. Isso é devido ao histórico de consumo mensal desses clientes terem grande homogeneidade. Isso é um indício de que há poucas perdas comerciais nessa classe de clientes. Tal conclusão é pertinente visto que a concessionária possui mais informaões a respeito dos clientes da classe industrial e toma medidas preventivas a fim de minimizar ao máximo as perdas comercias nesta classe composta por grandes consumidores.

De maneira análoga aos clientes comerciais, não será realizada uma avaliação da economia financeira advinda do combate às perdas comerciais em virtude das especificidades contratuais de cada cliente da classe industrial.

6.4.2 Análise dos módulos da metodologia para clientes industriais

Esta seção visa avaliar os parâmetros *PMC*, *GSA* e *ISE* que influem no valor do *ISC* e no resultado final das simulações para os clientes industriais.

No Quadro 18, tem-se a matriz de confusão referente ao Módulo de Classificação para os clientes industriais. As expressões (6.7), (6.8) e (6.9) representam, respectivamente, a confiabilidade negativa e a especificidade e a taxa de acerto para os dados do Quadro 16. Os

valores para esses três parâmetros indicam que o Módulo de Classificação obteve um desempenho satisfatório, ou seja, uma alta taxa de acerto das inspeções indicadas e um boa cobertura em relação à descoberta de todos clientes com perdas comerciais em uma região específica.

Quadro 18: Matriz de confusão para o Módulo de Classificação referente à simulação dos clientes industriais.

Matriz de Confusão		Classe Predita	
		N	F
Classe Real	N	$Q_{NN} = 276$	$Q_{NF} = 02$
	F	$Q_{FN} = 03$	$Q_{FF} = 13$

Fonte: Informações da pesquisa do autor.

$$c = \frac{Q_{FF}}{Q_{FF} + Q_{NF}} = \frac{13}{13 + 02} = 0{,}867 \tag{6.7}$$

$$e = \frac{Q_{FF}}{Q_{FF} + Q_{FN}} = \frac{13}{13 + 03} = 0{,}813 \tag{6.8}$$

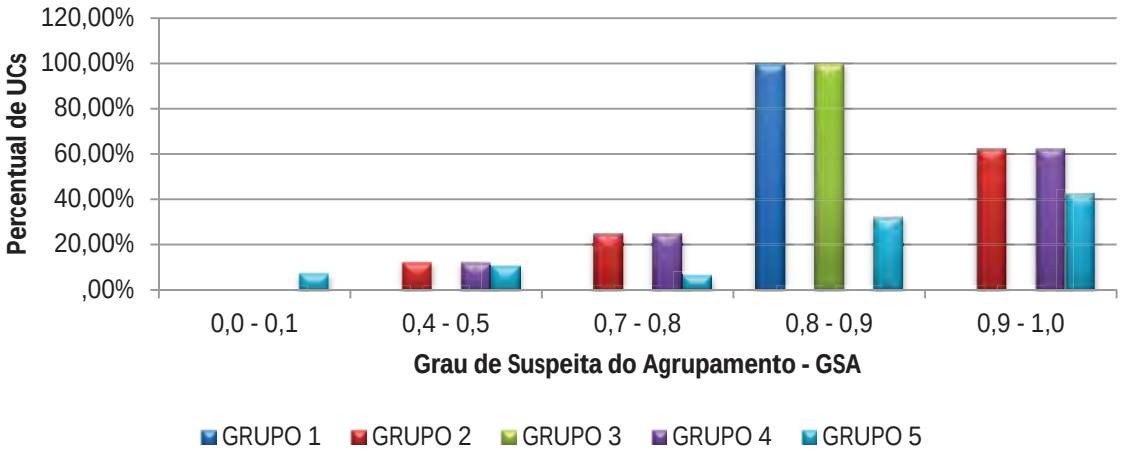
$$Taxa\_Acerto = \frac{Q_{NN} + Q_{FF}}{Q_{NN} + Q_{FF} + Q_{NF} + Q_{FN}} = \frac{276 + 13}{276 + \ 13 + 02 + 03} = 0{,}983 \tag{6.9}$$

O desempenho do Módulo de Agrupamento pode ser avaliado através da Figura 26 que representa os *GSA* de todos os grupos de clientes industriais. A maior parte dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *GSA* nos intervalos [0,8; 0,9], [0,9; 1,0], [0,8; 0,9], [0,9; 0,1] e [0,8; 1,0], respectivamente. Destaca-se o bom desempenho do Módulo de Agrupamento em relação a todos os grupos, exceto ao grupo 5 que obteve alto *GSA* para um grupo composto por clientes normais. Isso é devido a grande homogeneidade dos históricos de consumo mensais dos clientes industriais.

O desempenho do Módulo Especialista é avaliado a partir da Figura 27 que contém os *ISE* de todos os cinco grupos de clientes industriais. A parte majoritária dos clientes dos grupos 1, 2, 3, 4 e 5 possuem *ISE* nos intervalos [0,1; 0,2], [0,8; 0,9], [0,3; 0,5], [0,8; 0,9] e [0,8; 0,9], respectivamente.

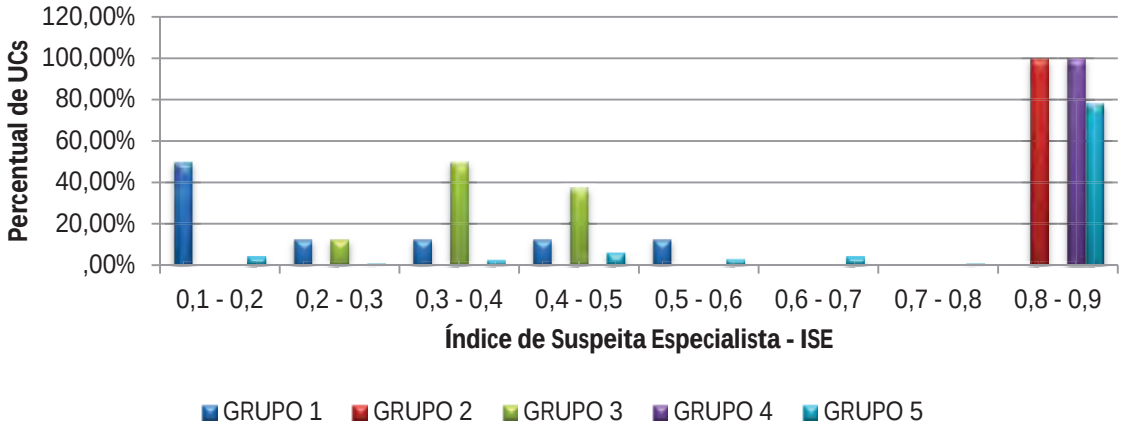
O valor do *ISE* fornecido pelo Módulo Especialista é coerente com os parâmetros de cada grupo. A exceção é o grupo 5 cujo *ISE* é alto para um grupo de clientes normais. Isso também é devido a grande regularidade do histórico de consumo mensal dos clientes industriais simulados.

Figura 26: Distribuição do *GSA* para cada grupo que compõe o histórico de inspeção dos clientes industriais.



Fonte: Informações da pesquisa do autor.

Figura 27: Distribuição do *ISE* para os grupos que compõem o histórico de inspeção dos clientes industriais.



Fonte: Informações da pesquisa do autor.

O Quadro 19 exibe um resumo das principais características e resultados das simulações como o número de UCs por classe de consumo, a quantidade de UCs suspeitas que existem nos quatro grupos anômalos arbitrados e a quantidade de UCs a serem inspecionadas que são aquelas cujo *RPC* é superior a 0,5. Têm-se também as perdas globais, as perdas comerciais e as perdas técnicas em cada classe. As perdas comerciais correspondem ao consumo (que não é faturado) de todas as UCs suspeitas. A energia recuperada refere-se à redução das perdas comerciais ocasionada pela inspeção em todas as UCs cujo *RPC* é superior a 0,5 e que pertencem a um dos grupos suspeitos.

A quantidade de clientes das três classes simuladas representam 99,5% do total de clientes do alimentador *A5*. A partir da análise do Quadro 19 observa-se que todas as UCs

dessas classes consomem em média 34,381 GWh/ano. Isso corresponde a 77,4% da energia transportada pelo alimentador *A5*. As perdas globais do alimentador *A5* considerando apenas as classes simuladas é de 3,464 GWh/ano que equivale a 7,8% da energia transportada por ele. A energia total recuperada a partir da redução das perdas comerciais através das inspeções aos clientes com alto *RPC* é de 1,36 GWh/ano que corresponde a uma redução considerável de 39,3% das perdas globais considerando apenas as três classes simuladas.

Quadro 19: Resumo dos parâmetros e resultados das simulações do alimentador *A5*.

Parâmetros e Resultados	**Unidades Consumidoras**		
	Residencial	**Comercial**	**Industrial**
Nº de UCs do Alimentador *A5*	11.758	752	294
Nº de UCs Suspeitas	1.176	76	32
Nº de UCs Inspecionadas ($RPC \geq 0{,}5$)	703	52	18
Consumo Total [GWh/ano]	21,495	8,602	4,284
Perdas Globais [GWh/ano]	2,282	0,952	0,230
Perdas Comerciais [GWh/ano]	1,825	0,762	0,184
Perdas Técnicas [GWh/ano]	0,456	0,190	0,046
Energia Recuperada [GWh/ano]	0,960	0,297	0,103

Fonte: Informações da pesquisa do autor.

No Quadro 20 estão resumidas as características gerais das simulações realizadas neste capítulo e alguns parâmetros que podem ser calibrados. Tem-se o tempo das simulações, a quantidade de épocas de treinamento, o número de agrupamentos e os pesos P_1 que pondera o *ISE*, P_2 que pondera o *GSA* e o *PMC* e, finalmente, P_3 que pondera *CI*, *R*, *LNS* e *LAS* (seção 5.38).

Do Quadro 20, depreende-se que o tempo de simulação é quase todo empregado no treinamento das redes neurais PMC e SOM. No entanto, após o treinamento das redes neurais, o tempo computacional das simulações reduz sensivelmente.

Observa-se também que o peso P_3 é superior aos demais, porque, nessas simulações, considera-se que os dados cadastrais são totalmente íntegros e mais confiáveis do que aqueles advindos do Módulo Especialista, do Módulo de Classificação e do Módulo de Agrupamento os quais têm pesos P_1 e P_2 idênticos em todos os casos simulados.

Quadro 20: Resumo dos principais parâmetros das simulações executadas no alimentador *A5*.

Parâmetros	Unidades Consumidoras		
	Residencial	**Comercial**	**Industrial**
Tempo para treinamento das redes neurais	1h 16min 8s	15min 44s	2min 25s
Tempo para simulação	58,3s	43,11s	6,6s
Nº de épocas para treinamento da rede neural PMC	111	468	2.063
Nº de épocas para treinamento da rede neural SOM	2.454	6.968	3.669
Nº de Agrupamentos	06	08	07
$[P_1, P_2, P_3]$	[1, 1, 2]	[1, 1, 2]	[1, 1, 2]

Fonte: Informações da pesquisa do autor.

6.5 CONSIDERAÇÕES FINAIS

Neste capítulo, obteve-se o *RPC* para o histórico de inspeções arbitrado o qual é constituído por cinco grupos característicos encontrados em históricos de inspeções reais. Os resultados estão expostos nas Figuras 17, 20 e 24 para os clientes residenciais, comerciais e industriais, respectivamente. Eles mostraram-se satisfatórios frente às dificuldades tais como:

- O histórico de consumo mensal dos clientes é muito homogêneo (típico de regiões com baixa incidência de perdas comerciais). A consequencia é a redução da eficiência dos Módulos de Classificação, do Módulo de Agrupamento e do Módulo Especialista.
- A base de conhecimento do Módulo Especialista e do Módulo Principal são gerais e foram elaboradas sem a presença do especialista que conhece a fundo o município de onde vieram os dados utilizados nas simulações.

Apesar de o resultado geral ter sido considerado satisfatório, algumas melhorias podem ser implementadas tais como: busca de melhor sintonização dos pesos $[P_1, P_2, P_3]$, aperfeiçoamento da metodologia para obtenção do *GSA* o qual caracteriza cada um dos agrupamentos e, em último, composição de uma base de regras para os Módulos Especialista e Módulo Principal mais específica da região em que foram colhidos os dados de entrada para as simulações.

7 CONCLUSÕES E SUGESTÕES PARA FUTUROS TRABALHOS

Neste trabalho, foi idealizada e programada computacionalmente uma metodologia para detecção das perdas comerciais em clientes ativos das empresas distribuidoras de energia elétrica. Atenção maior foi dedicada aos clientes das classes residencial, comercial e industrial os quais são responsáveis pela fatia majoritária das perdas comerciais no sistema de distribuição de energia elétrica. Ademais, dentre todas as classes de clientes, a detecção das perdas comerciais nessas classes é mais dificultosa e; por isso, é necessário o emprego de ferramentas sofisticadas a fim de aumentar as chances de localização de clientes fraudadores e, consequentemente, proporcionar a redução ao máximo das perdas comerciais.

O intuito da metodologia desenvolvida é utilizar todos os indicadores que se tem disponível (indicadores pontuais e indicadores por área) como dados de entrada para as técnicas de sistemas inteligentes. A saída fornecida é o *RPC* (Risco de Perda Comercial). Esse índice é associado a cada cliente analisado e representa a possibilidade do mesmo possuir algum problema que implique em acréscimo das perdas comerciais. Quanto maior a precisão desse indicativo, maior será a taxa de sucesso das inspeções em campo aos clientes suspeitos e menores serão os gastos da concessionária com as mesmas.

A maioria dos trabalhos da literatura avaliada utilizam somente os dados cadastrais e o histórico de consumo mensal. Ademais, eles utilizam ora técnicas orientadas à extração automática de informações ora técnicas orientadas ao conhecimento do especialista. Neste trabalho, no entanto, combinaram-se, em uma mesma metodologia, técnicas de extração automática de informações com técnicas que dependem do conhecimento especialista.

A obtenção do *RPC* deu-se a partir de outros indicadores obtidos da seguinte forma:

- Extração automática de conhecimento a partir do histórico de consumo mensal (indicadores *PMC* e *GSA*);
- Avaliação automática do histórico de consumo mensal via aquisição de conhecimento especialista (indicador *ISE*);
- Conhecimento especialista acerca da região em análise (indicador *ISA*);
- Busca por nomes e atividades suspeitas (indicadores *LNS* e *LAS*);
- Utilização de outros dados de entrada como *IPCA* (Índice de Perda Comercial do Alimentador), *CI* (Consumo Irregular) e *R* (Religamento).

As simulações com dados reais apresentaram resultados satisfatórios em virtude da redução das perdas comerciais. No entanto, tais resultados poderiam melhorar sensivelmente através do conhecimento do especialista específico da região de origem dos dados para calibração mais refinada dos pesos para cálculo do *ISC* – expressão (5.4). A metodologia apresentada possui forte caráter empírico. Portanto, a habilidade e a experiência do especialista para elaboração das funções de pertinência, da base de regras e para calibração dos parâmetros é crucial para o funcionamento adequado da metodologia proposta.

A seguir, algumas sugestões para trabalhos futuros:

- Não considerar os consumos mensais de maneira independente, isto é, relacionar o consumo em kWh ao mês em que ele ocorre. Se um dado consumidor tem o costume de viajar durante as férias escolares (meses de dezembro, janeiro e julho) então é normal que ocorra uma queda significava no consumo mensal de energia nesse período. O contrário também é verdade.
- Considerar as relações temporais existentes entre os atributos. Por exemplo, levar em conta o fato de janeiro ser posterior a dezembro e anterior a fevereiro.
- Elaborar Módulo de Explicação que justifique por que motivo determinado cliente possui alto *RPC*.
- Além de localizar onde há maior chance de existir as perdas comerciais, traçar uma rota que minimize os custos e que maximize o retorno das inpeções, isto é, $\max \frac{\text{Retorno Financeiro (inspeções)}}{\text{Custo das Inspeções}}$.
- Elaborar uma técnica a fim de evitar falsos positivos, isto é, existem consumidores cuja curva de consumo mensal apresenta comportamento suspeito, no entanto, em alguns casos, tal comportamento é consequência, por exemplo, de sua atividade econômica.
- Estimar a carga de cada consumidor e a partir daí estimar o seu consumo mensal.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL. **Metodologia de tratamento regulatório para perdas não-técnicas de energia elétrica**. Brasília, DF: SER/ANEEL, 2008. (Nota Técnica, n. 342/2008).

BASTOS, P. R. F. M. **Diagnóstico de perdas comerciais de energia elétrica na distribuição usando rede Bayesiana.** 2011. 171 f. Tese (Doutorado em Engenharia Elétrica) – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2011.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDEMIR, T. B. **Redes neurais artificiais**: **t**eoria e aplicações. Rio de Janeiro: LTC, 2007. 226 p.

CABRAL, J. E. et al. Fraud detection in electrical energy consumers using rough sets. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 2004, The Hague. **Proceedings…** Piscataway: IEEE, 2004. p. 3625-3629.

COMETTI, E. S.; VAREJÃO, F. M. **Melhoramentos da identificação de perdas comerciais através da análise computacional inteligente do perfil de consumo e dos dados cadastrais dos consumidores**. Vitória: UFES/NINFA, 2005. 11 p. Disponível em: <http://ninfa.inf.ufes.br/public_files/mip/ArtigoRelatorioFinal.pdf>. Acesso em: 15 fev. 2011.

DANTAS, P. R. P**. Avaliação de perdas de energia elétrica não-técnicas:** metodologia aplicada no município de Salvador/BA. 2006. 98 f. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia e Arquitetura, Universidade Salvador, Salvador, 2006.

DOMINGOS, J. L. **Uma contribuição à modelagem nebulosa.** 1998. 109 f. Dissertação (Mestrado em Engenharia Elétrica) – Centro de Ciências Exatas e Tecnologia, Departamento de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, 1998.

FERREIRA, H. M. **Uso de ferramentas de aprendizado de máquina para prospecção de perdas comerciais em distribuição de energia elétrica.** 2008. 89 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, São Paulo, 2008.

GILAT, A. **MATLAB**: com aplicações em engenharia. Porto Alegre: Bookman, 2006. 360 p.

GUERREIRO, J. I.; LEÓN, C.; BISCARRI, F. Increasing the efficiency in non-technical losses detection in utility companies. In: IEEE MEDITERRANEAN ELECTROTECHNICAL CONFERENCE – MELECON, 15., 2010, Valletta. **Proceedings…** Piscataway: IEEE, 2010. p. 136-141.

HAYKIN, S. **Redes neurais:** princípios e aplicações. Porto Alegre: Bookman, 2001. 893 p.

JIANG, R. J. Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In: IEEE/PES TRANSMISSION AND DISTRIBUTION CONFERENCE AND EXHIBITION: Asia Pacific, 2002, Dallas. **Proceedings…** Piscataway: IEEE, 2002. p. 06-10.

McCULLLOCH, W. S; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, New York, n. 9, p. 127-147, 1943.

MATSUMOTO, E. Y. **MATLAB® 7**: fundamentos. São Paulo: Érica, 2004. 376 p.

NIZAR, A. H.; DONG, Z. Y.; ZHANG, P. Detection rules for non-technical losses analysis in power utilities. IEEE POWER AND ENERGY SOCIETY GENERAL MEETING, 2008, Pittsburgh. **Proceedings...** Piscataway: IEEE, 2008, 8 p.

OLIVEIRA, M. E. **Avaliação de metodologias de cálculo de perdas técnicas em sistemas de distribuição de energia elétrica.** 2009. 137 f. Tese (Doutorado em Engenharia Elétrica) - Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2009.

PENIN, C. A. S. **Combate, prevenção e otimização das perdas comerciais de energia elétrica.** 2008. 227 f. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.

PERIM, G. T. et al. Uma abordagem baseada em conhecimento para identificação de perdas elétricas. In: CONGRESSO BRASILEIRO DE EFICIÊNCIA ENERGÉTICA, 2., Vitória. **Anais…** Vitória: Associação Brasileira de Eficiência Energética, 2007. 6 p.

RAMOS, C. C. O. et al. New insights on nontechnical losses characterization through evolutionary-based feature selection. **IEEE Transactions on Power Delivery**, Piscataway, v. 27, n. 1, p. 140-146, Jan. 2012.

RAUBER, T. W. et. al. Extração e seleção de características na identificação de perdas comerciais na distribuição de energia elétrica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., São Leopoldo. **Anais...** São Leopoldo: Sociedade Brasileira de Computação, 2005. 4 p.

REZENDE, S. O. et al. **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Manole, 2005. 525 p.

SANDRI, S.; CORREA, C. Lógica nebulosa. In: ESCOLA DE REDES NEURAIS, 5., 1999, São José dos Campos. **Anais...** São José dos Campos: Conselho Nacional de Redes Neurais, 1999. p. c073-c090. Disponível em: <http://www.ele.ita.br/cnrn/minicursos-5ern/logneb.pdf>. Acesso em: 15 fev. 2011.

SHAW, I. S.; SIMÕES, M. G. **Controle e modelagem Fuzzy**. São Paulo: Edgard Blücher, 1999. 165 p.

SILVA, I. N. Avanços em tecnologias de sistemas inteligentes e aplicações. In: FELTRIN, A. P. et al. (Eds.). **Tutoriais do XVIII Congresso Brasileiro de Automática, Bonito, Mato Grosso do Sul**. São Paulo: Sociedade Brasileira de Automática, 2010. v. 1, p. 65-96.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais**: para engenharias e ciências aplicadas. São Paulo: Artliber, 2010. 398 p.

SMITH, T. B. Electricity theft: a comparative analysis. **Energy Policy,** Oxford, v. 32, p. 2067-2076, 2004.

SIMPSON, P. K. **artificial neural systems**: foudations, paradigms, applications and implementations. New York: Pergamon Press, 1989.

SPERANDIO, M.; COELHO, J.; QUEIROZ, H. L. Identificação de agrupamentos de consumidores de energia elétrica através de Mapas Auto-Organizáveis. In: SEMINÁRIO BRASILEIRO SOBRE QUALIDADE DA ENERGIA ELÉTRICA, 5., 2003, Aracaju. **Anais...** Aracaju: Núcleo de Estudos e Pesquisas do Nordeste, 2003.

THE MATHWORKS INC. **Fuzzy Logic Toolbox 2**: user's guide. Natick, 2007. Disponível em: <http://www.mathworks.com/help/pdf_doc/fuzzy/fuzzy.pdf>. Acesso em: 15 fev. 2011.

WIDROW, B.; LEHR, M. 30 years and adaptive neural networks: perceptron, madaline and backpropagation. **Proceedings of the IEEE**, Piscataway, v. 78, n. 9, p. 1415-1442, Sep. 1990.

ZADEH, L. Fuzzy sets. **Information and Control**, Maryland Heights, v. 8, p. 338-353, 1965.