Universidade Estadual de Campinas

Faculdade de Engenharia Elétrica e de Computação

Uso de Ferramentas de Aprendizado de Máquina para Prospecção de Perdas Comerciais em Distribuição de Energia Elétrica

Hamilton Melo Ferreira

Orientador: Prof. Dr. Fernando José Von Zuben

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Área de Concentração: Engenharia de Computação.

Campinas (SP) - Brasil

Janeiro de 2008

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

Ferreira, Hamilton Melo
F413u Uso de ferramentas de aprendizado de máquina para prospecção de perdas comerciais em distribuição de energia elétrica / Hamilton Melo Ferreira. --Campinas, SP: [s.n.], 2008.

Orientador: Fernando José Von Zuben.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Aprendizado do computador. 2. Sistemas de energia elétrica - Modelos matemáticos. 3. Sistema de suporte de decisão. 4. Mineração de dados (Computação). 5. Sistemas especialistas (Computação). I. Von Zuben, Fernando José. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Use of machine learning tools for prospecting commercial losses in electric energy distribution
Palavras-chave em Inglês: Machine learning, Electrical power systems - Mathematical models, Decision support system, Data mining (Computer), Expert systems (Computer)
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: João Luís Garcia Rosa, Christiano Lyra Filho, Romis Ribeiro de Faissol Attux
Data da defesa: 29/01/2008
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA – TESE DE MESTRADO

Candidato: Hamilton Melo Ferreira

Data da Defea: 29 de janeiro de 2008

Título da Tese: Uso de Ferramentas de Aprendizado de Máquina para Prospecção de Perdas Comerciais em Distribuição de Energia Elétrica

Prof. Dr. Fernando José Von Zuben (Matr. 263958): Fernando José Von Zuben

Prof. Dr. João Luís Garcia Rosa: João Luís G. Rosa

Prof. Dr. Christiano Lyra Filho:

Prof. Dr. Romis Ribeiro de Faissol Attux: Romis

Resumo

As concessionárias de energia elétrica deixam de faturar anualmente expressivos valores devido a perdas comerciais, as quais são originadas principalmente por fraudes cometidas por parte dos consumidores e por medidores defeituosos. A detecção automática dos pontos específicos onde ocorrem tais perdas é uma tarefa complexa, dada a grande quantidade de consumidores, a grande variedade de perfis de consumo de energia elétrica e o alto custo de cada inspeção. Este trabalho propõe o uso de técnicas de aprendizado de máquina para a incorporação de processamento inteligente na identificação das fontes de perdas comerciais, usando os dados reais fornecidos pela concessionária de energia elétrica AES Eletropaulo. Além da manipulação dos dados e análise de propostas alternativas presentes na literatura, quatro estratégias de classificação foram implementadas e comparadas, sendo que o algoritmo de indução C4.5 produziu os resultados mais consistentes em termos de especificidade e confiabilidade, tomadas como critérios de desempenho.

Abstract

The electric power concessionaires miss along the year significant amount of revenue due to commercial losses, which are mainly caused by frauds produced by consumers and defective sensors. The automatic detection of the specific sites where the losses are located is a complex task, given the high number of consumers, the great variety of electric power consumption profiles, and the high cost of each inspection. This work proposes the use of machine learning techniques capable of incorporating intelligent processing in the identification of the sources of commercial losses, using real data provided by the electric power concessionaire AES Eletropaulo. Besides data manipulation and analysis of alternative proposals presented in the literature, four classification strategies have been implemented and compared. The C4.5 algorithm has produced the most consistent results in terms of specificity and confiability, taken as performance criteria.

Dedicatória

Aos meus pais:
Maria Dirce e Wilmar (in memoriam),
ao meu irmão Alaor e
à minha namorada Marciane

Agradecimentos

Ao **meu pai Wilmar** (in memoriam), pela luz que abria novos horizontes nos momentos de indagações.

À **minha mãe Dirce**, pelo carinho, pelas orações e pela confiança e incentivo que sempre depositou em mim.

Ao **meu irmão Alaor**, braço amigo para qualquer necessidade.

À minha namorada **Marciane Milanski**, pelo seu carinho, compreensão, apoio, confiança, companherismo.

Ao meu amigo e orientador, **Prof. Dr. Fernando José Von Zuben**, pela incansável disposição em aceitar novos desafios, pela confiança depositada dentro e fora do LBiC e pelo apoio intelectual.

Aos **amigos do LBiC**, é um prazer estar ao lado de pessoas tão alegres, cheias de energia e sempre dispostas a aprender e trocar conhecimento.

Ao colega do LBiC **Wilfredo Jaime Puma Villanueva** por ter autorizado a utilização de seus dados nesse trabalho.

À **Inova** – Agência de Inovação da Unicamp e especialmente ao **Vernei Gialluca** pela confiança que deposita no LBiC.

A **AES Eletropaulo**,

por ter fornecido os dados para este trabalho,

pelo seu apoio financeiro,

pelo apoio técnico, principalmente **de Eduardo Bortotti Fagundes** e **Denílson Porto**.

Ao **Denílson Porto** que confiou em nós e não mediu esforço para a realização deste trabalho.

Ao **CNPq** e à **ANEEL** pelo incentivo à pesquisa brasileira e pelo apoio financeiro durante a realização deste trabalho.

Aos professores da Unicamp, especialmente aos da FEEC, sempre dispostos a compartilhar seus conhecimentos e suas experiências.

À Unicamp, à FEEC, juntamente com todos os seus funcionários, pela estrutura e apoio oferecidos aos seus alunos.

Índice

CAPÍTULO 1

Aprendizado de Máquina em Perdas Comerciais

Este capítulo de apresentação do trabalho realizado contém uma breve introdução à área de aprendizado de máquina e à aplicação de suas técnicas a um dos principais problemas enfrentados pelas concessionárias de energia elétrica, as perdas comerciais. No final do capítulo, é descrita a estrutura da dissertação.

1.1. Introdução

É indiscutível que a quantidade de dados armazenados cresce de maneira vertiginosa e, como cada vez mais surgem novos dispositivos de armazenamento e de comunicação de dados, com maior capacidade tanto de armazenamento como de processamento e com preços inferiores, não há limite à vista para o fim dessa expansão.

Além disso, existe também uma consciência de que informação vale "ouro". No entanto, nem sempre a existência de dados significa informação, pois é necessário extrair "algo de útil" desses dados.

Diante disso, por exemplo, os departamentos de marketing das empresas de todos os segmentos analisam as informações procurando descobrir onde se encontram os seus clientes e o que eles desejam, as instituições financeiras desejam saber qual o risco em conceder crédito para um determinado cliente e os cientistas analisam as expressões gênicas para descobrir o fundamento genético de certas enfermidades. Enfim, existem inúmeras aplicações que

mineram os dados na busca de uma informação útil (que seria equivalente a uma pepita de ouro).

Entretanto, a análise de grandes bases de dados para buscar a informação desejada é uma tarefa árdua, mesmo com a ajuda dos computadores, pois envolve quantidades elevadas de itens e operações a serem aplicadas sobre esses itens de dados, além da dificuldade adicional causada pelo fato dos dados serem, muitas vezes, repetidos, incompletos e/ou inconsistentes.

Diante deste cenário, estabeleceu-se a área da computação chamada de **aprendizado de máquina**, cuja principal motivação, embora não seja a única, está na possibilidade de desenvolvimento de técnicas computacionais capazes de viabilizar a extração automática de conhecimento a partir de bases de dados. Nos algoritmos de aprendizado de máquina, as relações entre os atributos dos dados não estão pré-definidas. O objetivo é justamente "aprender" essas relações a partir do próprio conjunto de dados de entrada, sem que haja intervenção direta de um ser humano.

Uma das áreas que necessita de algoritmos de aprendizado de máquina para melhorar a sua eficiência é a de **Perdas Comerciais das Concessionárias de Distribuição de Energia Elétrica**, como é explicado na subseção 1.3.

As perdas comerciais saem do âmbito do "fio" (da rede elétrica) e fixam-se principalmente no âmbito dos consumidores, dos medidores, do cadastro e do faturamento. Como exemplo, pode-se citar o **popular "gato na rede elétrica"** (Figura 1.1), onde os clientes da distribuidora de energia elétrica fazem ligações elétricas diretamente na rede de distribuição, sem passar pelo "relógio".

Figura 1.1: Ligações clandestinas na rede elétrica.

O objetivo específico deste trabalho é auxiliar o Departamento de Perdas Comerciais da AES Eletropaulo a detectar os locais onde as perdas comerciais ocorrem. Para isso, propõe-se o desenvolvimento de ferramentas de aprendizado de máquina, as quais trabalhariam com os dados armazenados na base de dados da AES Eletropaulo, porém não se limitando a esta base de dados.

Nas próximas subseções, são discutidos em mais detalhes conceitos relacionados a aprendizado de máquina e também a perdas comerciais em distribuição de energia elétrica.

1.2. Aprendizado de Máquina

Um algoritmo consiste em uma seqüência de passos bem definida e determinística. Todo programa de computador implementa algum tipo de algoritmo, como os programas que extraem de um banco de dados as informações desejadas em forma de relatório. Neste cenário, o programador conhece a priori a estrutura dos dados e o procedimento de acesso a cada item da base de dados.

No entanto, existem cenários em que não se sabe antecipadamente que tipo de manipulação deve ser realizado junto à base de dados, caracterizando assim problemas mal definidos.

As técnicas de aprendizado de máquina trabalham com problemas mal definidos, em que não é possível analisar todas as alternativas possíveis, dada a grande quantidade de possibilidades de encaminhamento até a solução. Assim, um algoritmo baseado em aprendizado de máquina extrai informações da base de dados sem que uma pessoa tenha decidido previamente quais os dados que devem ser acessados.

A diferença entre os dois tipos de algoritmo é que, no primeiro caso, o programador sabe qual é a saída desejada e quais são as operações necessárias para se atingir o objetivo, enquanto que no algoritmo de aprendizado de máquina geralmente não se sabe antecipadamente qual a saída e nem a seqüência de passos que deverá ser executada para resolver o problema.

Para explicar essa diferença será usado, como exemplo, o problema da concessão de crédito bancário. O objetivo deste problema é estimar qual é o risco de uma determinada pessoa se tornar inadimplente se lhe for concedido um empréstimo bancário. Para isso, espera-se determinar um conjunto de regras a partir da base de dados existente, contendo dados de muitos clientes, tanto de inadimplentes como daqueles que nunca atrasaram o pagamento de seus financiamentos.

Assim, o objetivo de um algoritmo de aprendizado de máquina é descobrir, por exemplo, as regras que determinam o risco do cliente ser ou não inadimplente. Uma vez determinadas quais são essas regras, é possível implementá-las facilmente em um programa de computador.

É claro que algumas regras simples e que fazem parte do senso comum, como do tipo "se desempregado então o risco de ser inadimplente é alto", já foram implementadas em algoritmos tradicionais e validadas ou descartadas na prática, já que o senso comum nem sempre retrata as tendências e a realidade de cenários em que muitos fatores estão envolvidos nos processos de tomada de decisão.

No entanto, como o mundo está cada vez mais competitivo e os processos e sistemas cada vez mais intricados e integrados, as empresas procuram melhorar a maneira como executam suas operações sobre processos e sistemas. Isto faz com que elas busquem regras mais complexas como "se a pessoa é solteira, reside na Asa Sul de Brasília, exerce uma determinada profissão e o nível da atividade industrial está em alta, então o risco é X", pois os bancos (no caso do exemplo) estão atrás da fatia dos clientes com maior interesse em consumo e que não serão inadimplentes. A crise financeira nos Estados Unidos da América, iniciada em 2008, começou justamente porque os bancos emprestaram elevadas quantias (através do crédito imobiliário) para pessoas que muito provavelmente não irão pagar as suas dívidas.

Além disso, no caso de fraudes, seja na área elétrica, seja em cartões de crédito, os fraudadores também estão se tornando mais sofisticados, ou seja, os algoritmos de detecção de fraudes mais simples deixaram ou vão deixar de ser competitivos.

Um outro exemplo na área de saúde: já se sabe que fumar aumenta o risco de ter câncer. Porém, agora resta descobrir outros fatores, além de fatores genéticos, que aumentam a

tendência de pessoas que nunca fumaram também desenvolverem câncer e algumas que sempre fumaram não desenvolverem a doença.

Observe que, tanto no exemplo de determinação do risco de crédito como na descoberta de outros fatores cancerígenos, não se sabe qual a saída, sendo este o objetivo dos algoritmos de aprendizado de máquina.

Existem vários algoritmos e meta-heurísticas, como Redes Neurais Artificiais (RNAs) e Algoritmos Genéticos, o indutor de árvore de decisão C4.5 e SVM (do inglês "Support Vector Machine"), que são classificados como técnicas de aprendizado de máquina.

Este leque de opções se sustenta pelo fato de que não é possível saber qual vai ser o desempenho relativo de cada ferramenta de aprendizado de máquina, pois há dependência de peculiaridades de cada aplicação. Logo, o que se faz é executá-las, checar os seus resultados e compará-los. É justamente esta a proposta deste trabalho.

A seguir, é descrito qual foi o problema que motivou esse trabalho e é exposta a razão da escolha de algoritmos de aprendizado de máquina.

1.3. Perdas comerciais em distribuição de energia elétrica

Perdas comerciais são as partes do faturamento das concessionárias de energia elétrica perdidas devido às fraudes no sistema elétrico (o popular "gato na rede elétrica"), aos problemas com os medidores e aos erros em procedimentos internos.

O controle das perdas das distribuidoras de energia elétrica, tanto no âmbito técnico quanto no comercial, é hoje, sem dúvida, um dos principais elementos responsáveis pela eficiência da corporação. No lado das perdas técnicas, muito já se fez em pesquisa e desenvolvimento de técnicas e algoritmos para sua mensuração e minimização, sendo que nos últimos anos foram desenvolvidos reconfiguradores de redes, sistemas de proteção e alocação de capacitores, que incorporam definições e funcionalidades das áreas de planejamento e operação dos sistemas de distribuição. Porém, muito pouco se fez no âmbito das perdas comerciais, razão pela qual esta foi uma das áreas escolhidas como prioritárias no ano de 2005

pela ANEEL, que estimou a perda comercial do setor brasileiro de distribuição de energia elétrica em R$ 5,1 bilhões em 2005 (Francisco, 2006).

Este não é o primeiro trabalho na área de perdas comerciais em distribuição de energia, existindo tanto contribuições de pesquisadores brasileiros (Cometti & Varejão, 2005) como de outros países (Jiang et al., 2002).

Basicamente as perdas comerciais são agrupadas nos seguintes segmentos, em relação a sua causa:

a) Causas internas às distribuidoras, como erros de leitura, de faturamento, de cadastramento de consumidores e de equipamentos obsoletos ou com fadigas.

b) Causas externas às distribuidoras e relacionadas aos hábitos de classes de consumidores, como fraudes em equipamentos e ligações clandestinas feitas por clientes, além de falta de informações necessárias no processo de ligação, quando não realizado pela empresa.

c) Causas externas às distribuidoras e relacionadas com problemas sociais e governamentais, como invasões e crescimento de áreas de baixa renda, como favelas. A diferença desta causa em relação ao item anterior é que, neste caso, o fraudador não é cliente da empresa, pois não tem nem cadastro nem medidor instalado, e também não tem condições financeiras para pagar a conta.

Não está no escopo deste trabalho detectar as áreas em que o fraudador não é cliente da empresa, como ocorre com as favelas e invasões. Nessas áreas, a fonte de perdas é facilmente detectável, mas a regularização já envolve questões sociais e legais. Para atuar junto a esse tipo de problema, a AES Eletropaulo faz um trabalho conjunto com as prefeituras para a regularização de ligações clandestinas.

Em relação às fraudes cometidas pelos clientes, elas não se limitam às ligações diretas na rede elétrica, pois os clientes estão utilizando artifícios mais difíceis de serem detectados, como alterações internas nos medidores de energia, para que registrem valores inferiores ao

consumido. Como já foi mencionado, na área de fraude existe uma certa competição entre empresas e fraudadores, cada lado buscando aprimorar mais e mais seus métodos.

Um outro caso que se insere na área de perdas comerciais é a utilização de medidores com defeitos ou com fadiga. Neste caso, o cliente não está fraudando. Porém, a empresa continua tendo perda de receita pelo uso de tais medidores. O medidor de energia elétrica não pertence ao cliente, e sim à empresa de distribuição de energia elétrica. Assim, a empresa tem o cadastro de cada medidor em cada cliente. Mesmo assim, é difícil detectar os medidores com problemas, pois eles sofrem deterioração em função das condições dos locais onde estão instalados, como: entrada de água no medidor, ação de insetos, sobre-tensão na rede. Além disso, cada medidor apresenta mais ou menos defeitos em função também do fabricante, modelo, série e tempo de uso.

As ligações clandestinas, juntamente com as fraudes, causam outras perdas às distribuidoras, além das relacionadas ao faturamento. Como, na maioria das vezes, essas ligações são feitas precariamente, elas podem provocar curtos-circuitos, fazendo com que toda uma região fique sem energia elétrica, além de aumentarem o risco de incêndio e morte de pessoas. Outras vezes, um cliente (tipicamente industrial ou comercial) que necessitaria de uma tensão maior frauda a companhia usando uma tensão não adequada ao seu consumo, aumentando, assim, a intensidade da corrente elétrica na rede e, conseqüentemente, as perdas técnicas. Ou seja, neste caso, a perda comercial provoca um aumento na perda técnica.

As perdas comerciais são caracterizadas como pontuais e dispersas, representando, no entanto, a parcela junto a qual as empresas têm obtido o maior grau de recuperação em relação às perdas totais ocorridas na empresa. Diferentemente da área de perdas técnicas, onde é possível saber com boa precisão onde ocorre a perda de energia e seu valor, em perda comercial não é possível precisar qual consumidor está fraudando a distribuidora, qual o impacto da fraude ou qual medidor está com problema. Sabe-se que existe perda comercial pela diferença entre a energia distribuída e a faturada, descontando-se as perdas técnicas. Porém, é necessária uma inspeção minuciosa, talvez envolvendo todos os clientes, para checar se algum cliente está fraudando ou se o medidor está com defeito, o que é infactível devido à grande quantidade de clientes. Além do mais, no caso de fraude, um cliente pode remover a

ligação clandestina ao perceber a proximidade de um fiscal. E nada impede que um cliente volte a fraudar a empresa logo após ele ter sido descoberto e ter a sua situação regularizada.

Para se conseguir a diminuição das perdas comerciais, é necessário lidar com o conjunto de fatores descritos abaixo de forma integrada, o que caracteriza um problema de alta complexidade.

Fatores a serem considerados na prospecção das perdas comerciais:

(i) Os dados estão espalhados por diferentes processos e departamentos. Por exemplo, tem um departamento que é responsável por entregar a energia ao cliente, enquanto outro tem como função cobrar a energia que foi consumida, existindo também os departamentos de manutenção (um para os equipamentos e outro para a rede elétrica), além do departamento que estuda melhorias na rede. As bases de dados desses departamentos geralmente não estão unificadas, fazendo com que haja diferentes sistemas de armazenamento e modelos relacionais para atender às necessidades específicas de cada departamento. Com isso, não é considerada a utilização dos seus dados por outros departamentos, o que acarreta, por exemplo, em discrepâncias na definição dos processos de coleta, pré-processamento e armazenagem dos dados, dificultando, assim, a integração das bases de dados.

(ii) O perfil dos fraudadores varia com o tempo, e a simples inspeção (ou a falta dela) pode alterar o perfil do consumidor. Se um cliente é inspecionado (mesmo sem ser detectada nenhuma irregularidade), ele vai ponderar melhor quais são os riscos de fraudar a companhia e, muitas vezes, isto vai afetar também as pessoas próximas, como vizinhos, parentes e indústrias do mesmo ramo. O raciocínio é similar no caso de falta de inspeções, mas com efeito inverso. Assim, o ideal é inspecionar todos os clientes, embora o custo seja proibitivo.

(iii) O custo de serviços de prospecção em campo é elevado, devido ao tempo empregado e aos recursos alocados na preparação e transporte dos técnicos.

(iv) A diversidade de consumidores é elevada.

(v) O adensamento de carga é elevado.

Como conseqüência, há um alto custo de prospecção e um alto valor não faturado devido às perdas comerciais e há uma grande quantidade de dados distribuídos e variantes no tempo a serem manipulados e analisados. Assim sendo, por mais experiência que tenham os gestores de perdas da empresa, são necessárias ferramentas computacionais inteligentes para lidar com esse tipo de problema.

A proposta deste trabalho é desenvolver ferramentas inteligentes com capacidade de "aprender" as regras baseando-se nos históricos e nas necessidades da empresa. O algoritmo aprenderia, por exemplo, uma regra do tipo "medidor do fabricante *X*, fabricado dentro do período *P*, instalado com transformadores de corrente do tipo *T*, apresentam defeitos nas regiões *R* com probabilidade *Pr*", a qual seria utilizada para detectar os medidores defeituosos ou com fadiga.

1.4. Estrutura do Trabalho

Este trabalho é composto por cinco capítulos. O segundo capítulo aborda o processo de recuperação das perdas comerciais da AES Eletropaulo e a melhoria proposta por este trabalho. O terceiro capítulo descreve os dados utilizados atualmente na área de perdas comerciais e as métricas de desempenho. O quarto capítulo descreve e analisa os testes realizados com quatro ferramentas computacionais (C4.5, RNA, SVM e Naive Bayes) e o último capítulo é dedicado às considerações finais e às perspectivas futuras da pesquisa.

CAPÍTULO 2

Recuperação de Perdas Comerciais

Neste capítulo, são descritos o processo de recuperação de perdas comerciais empregado pela AES Eletropaulo e o projeto de redução de perdas em que este trabalho está inserido.

2.1. O Caso AES Eletropaulo

A AES Eletropaulo é uma concessionária que distribui energia elétrica para 24 municípios da região metropolitana de São Paulo (incluindo a Capital), cuja área abrange 4.526 km² e concentra a região socioeconômica mais importante do país, totalizando uma população de 16,5 milhões de habitantes e fazendo com que a AES Eletropaulo seja, em faturamento, a maior distribuidora de energia elétrica da América Latina.

Em 2006, os clientes cativos da AES Eletropaulo consumiram 31,65 mil GWh, o que correspondeu a 35% do consumo estadual, a 9,4% do nacional e a uma receita líquida de R$ 8,3 bilhões. A distribuidora possui 15 mil clientes corporativos, os quais respondem por 40% do faturamento.

Em setembro de 2005, a AES Eletropaulo registrou 13,01% de perdas (técnicas e comerciais), divididas em 5,60% de perdas técnicas e 7,41% de perdas comerciais. Anualmente, a AES Eletropaulo deixa de faturar R$ 500 milhões (sem a incidência de impostos) por causa das ligações clandestinas de energia, estimadas em 477 mil na área de concessão da AES Eletropaulo. As ligações informais consomem 1.780 GWh ano. Essa quantidade de energia é equivalente ao consumo residencial anual dos municípios de São

Bernardo, São Caetano, Santo André e Diadema, que juntos têm 2 milhões de habitantes (FRANCISCO, 2006).

A AES Eletropaulo tem um departamento dedicado exclusivamente a diminuir as perdas comerciais, possuindo uma equipe experiente que procura constantemente contratar empresas e grupos de pesquisa externos à empresa para auxiliá-los. No entanto, como se trata de um problema complexo, acredita-se que haja espaço para novas contribuições nesta área, permitindo assim a redução das perdas comerciais.

O objetivo estratégico do departamento de Perdas Comerciais é reduzir a ocorrência de perdas comerciais, e não apenas detectá-las depois que elas ocorreram. Para isto, idealizam e implementam políticas e ações específicas para diminuir a ocorrências de perdas comerciais, como:

- Trabalho conjunto com o poder público para evitar roubo de energia em favelas e áreas invadidas.
- Ações de conscientização da população sobre os riscos de ligações clandestinas.
- Instalação de medidores nos postes, e não nas casas dos consumidores.

Todas essas ações têm um custo, de modo que elas devem ser priorizadas segundo a relação custo-benefício.

Já o objetivo tático é detectar os pontos de perdas comerciais, tomar ações corretivas e, no caso de fraudes, cobrar retroativamente a energia consumida e não paga. É nesta frente de atuação que esse trabalho se insere.

O processo de recuperação de perdas comerciais, conforme descrito na Figura 2.1, consiste em analisar dados de diferentes departamentos e selecionar locais para serem inspecionados com o objetivo de detectar fraudes por parte dos clientes ou anomalias nos medidores.

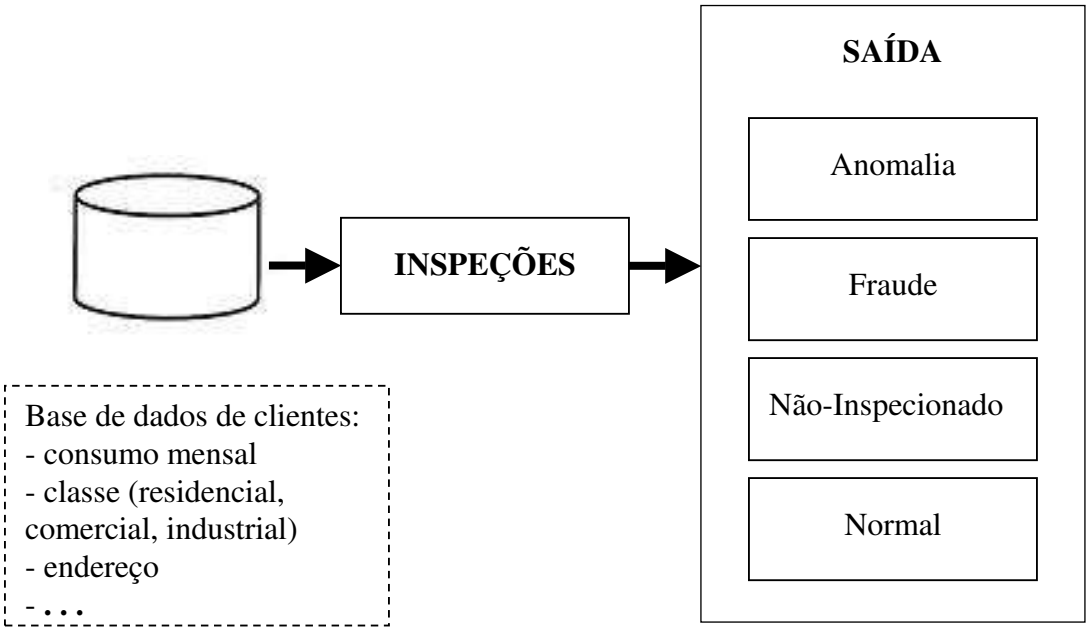

Figura 2.1: Fluxo de dados para as inspeções.

Atualmente, os dados estão espalhados em diversos sistemas. No entanto, a AES Eletropaulo tem investido em uma nova base de dados e, em breve, a empresa espera conseguir ter um acesso mais eficaz e integrado aos dados.

As inspeções são rotuladas pelo seu código de retorno:

- Anomalia: medidor com defeito.
- Fraude: o cliente está fraudando a empresa.
- Não-Inspecionado: não foi possível realizar a inspeção por diversos motivos, como endereço não encontrado ou local fechado. Dependendo do laudo (como exemplo, local fechado), o inspetor volta numa data futura para tentar inspecionar o local.
- Normal: não existe problema.

Nos último 3 anos, foram realizadas mais de 1 milhão e 200 mil inspeções. A Figura 2.2 mostra como elas foram rotuladas segundo o código de retorno.

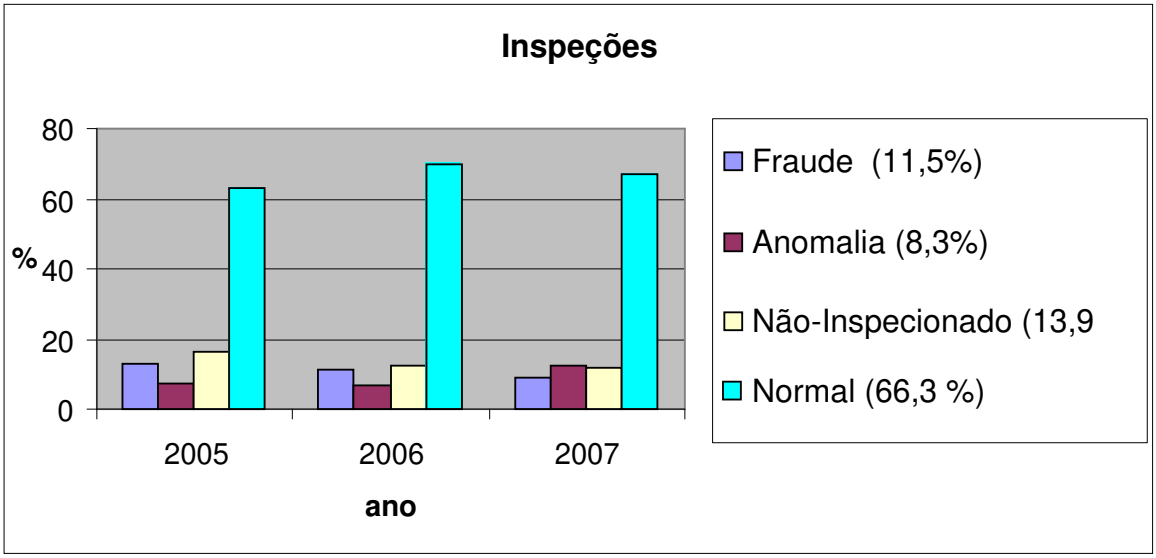

Figura 2.2: Histograma dos códigos de retorno das inspeções realizadas entre 2005 e 2007. Ao lado dos rótulos, consta a porcentagem média entre 2005 e 2007 para cada classe.

Os dados apresentados na Figura 2.2 mostram, portanto, que na média 19,8% das inspeções (fraude mais anomalia) obtiveram sucesso. Assim, a empresa está investindo sem retorno em aproximadamente 80% das inspeções, o que representa um alto custo.

2.2. O Processo de Inspeção

Os locais a serem inspecionados são escolhidos através de várias análises estatísticas e baseadas em conhecimento de especialistas expresso na forma de regras, sendo que a informação disponível é o conteúdo da base de dados da empresa. Além disso, os programadores das inspeções monitoram o código de retorno das inspeções. Assim, se as primeiras inspeções dentre um conjunto de inspeções programadas não fornecerem resultados satisfatórios, as inspeções ainda pendentes são canceladas e um novo lote de inspeções é programado. Ou seja, o departamento está muito atento ao desempenho das inspeções.

Mesmo com todo esse esforço para que as inspeções sejam eficazes, o índice de acerto médio nos últimos três anos é de 19,8%. Visando aumentar o índice de acerto, busca-se um conjunto mais eficaz de regras, ou até outras estratégias, para selecionar os clientes a serem inspecionados.

Trabalhos anteriores (FRANCISCO & FAGUNDES, 2007), feitos com dados da própria AES Eletropaulo, indicam que a fraude está diretamente relacionada a renda, aspectos culturais e condições sócio-econômicas da população, ou seja, variáveis mutuamente relacionadas e associadas à localização geográfica.

A experiência dos programadores de inspeção e o bom-senso também sugerem que a tendência em fraudar está muito relacionada com a vizinhança, pois um vizinho pode induzir o outro a cometer a fraude. No entanto, mesmo considerando que essas premissas são verdadeiras, ainda assim é uma tarefa árdua descobrir esses locais, como mostra a Figura 2.3.

O histograma apresentado na Figura 2.3, cuja taxa de acerto é calculada pela Equação 2.1, mostra a variação na taxa de acerto em função da região e do período. Observa-se, por exemplo, que a região do ABC, que apresentou alto índice de acerto em 2005, teve um índice baixo em 2007.

$$taxa\_de\_acerto = \frac{\sum fraudes + \sum anomalias}{\sum inspeções} \tag{2.1}$$

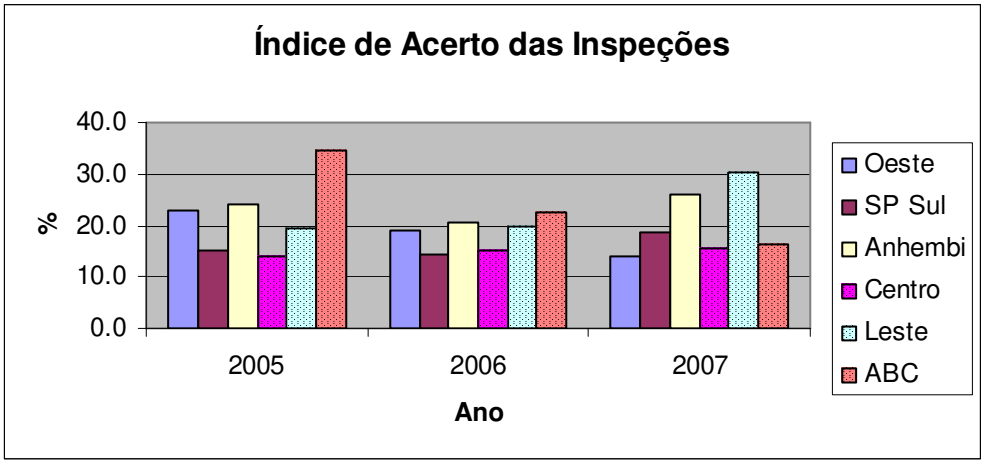

Figura 2.3: Índices anuais de acerto das inspeções por região.

É possível dividir as 6 grandes regiões mostradas no gráfico acima em sub-regiões e fazer um gráfico semelhante ao da Figura 2.3. Porém, ainda assim não foi possível determinar uma correlação entre as regiões e as fraudes mais as anomalias.

Enfim, pode-se levantar um alto número de análises estatísticas tanto em relação às regiões geográficas quanto às classes do cliente (residencial, industrial, comercial, público) ou ao tipo de atividade. No entanto, mesmo assim não se conseguiu descobrir um padrão.

Um dos fatores que contribui para isto é que o próprio trabalho de detecção de fraudes provoca a sua diminuição. Assim como uma pessoa fraudadora pode induzir o seu vizinho a fazer o mesmo, quando alguma fraude é descoberta, os vizinhos também ficam sabendo e, muitas vezes, desistem de cometê-la.

2.3. Descrição do Projeto

2.3.1. Objetivo

Esta dissertação está vinculada a um projeto de pesquisa e desenvolvimento com a AES Eletropaulo, apoiado pela ANEEL. O objetivo inicial proposto era aumentar a taxa de acertos nas inspeções. No entanto, percebeu-se que a AES Eletropaulo deseja ter maior retorno financeiro em suas inspeções. Isto significa que é preferível ter um retorno financeiro alto, mesmo que a uma taxa de acerto menor. Esta é a razão pela qual a política da AES Eletropaulo é inspecionar todos os clientes industriais anualmente, mesmo com um baixo índice de sucesso (2%).

Assim, conforme é mostrado na Figura 2.4, o problema passa a ser modelado como uma tarefa de otimização, cujo objetivo é aumentar a receita, levando em conta também o custo de cada inspeção. Esta constatação é de extrema relevância, pois insere um viés na forma com que o produto final decorrente da execução do projeto será avaliado. A classificação ainda continua necessária, pois ela fornece o risco de fraude. No entanto, será necessário um módulo de otimização que considere o risco de fraude e a taxa de retorno financeiro ao escolher os locais a serem inspecionados. Não faz parte do escopo deste trabalho o desenvolvimento do módulo de otimização.

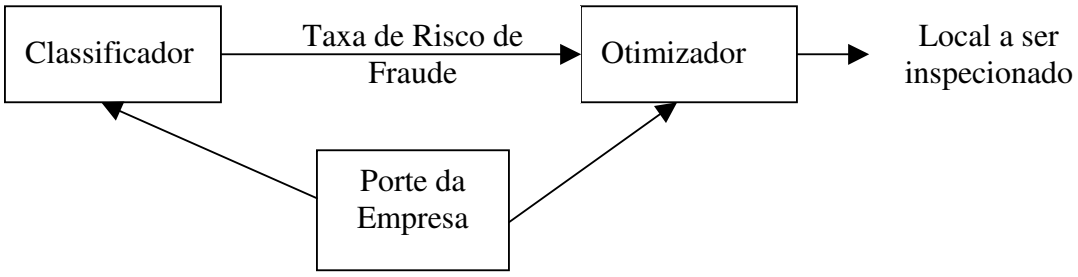

Figura 2.4: Novo paradigma para abordar o problema de perdas comerciais

É conveniente salientar que a necessidade de se buscar o maior retorno financeiro ao selecionar os locais a serem inspecionados não é só de interesse da AES Eletropaulo (uma empresa privada), mas também da ANEEL e de toda a sociedade, porque a conta de consumo de energia elétrica contém uma parcela referente às perdas comerciais que é medida em valores monetários, e não em número de clientes fraudulentos.

2.3.2.Metodologia

A duração prevista para este projeto é de 18 meses, e no momento o projeto se encontra no seu 5º mês de execução. A conclusão desta dissertação marca o encerramento do primeiro ciclo. A metodologia de desenvolvimento de software adotada é o paradigma evolutivo (CARVALHO & CHIOSSI, 2001), com as atividades de desenvolvimento e validação sendo executadas em paralelo e com liberações de versões intermediárias dos softwares para validação em campo, conforme descrito na Figura 2.5. De fato, como não existe uma metodologia fechada para resolver o problema descrito, é necessário trabalhar em conjunto com a AES Eletropaulo para se detalhar melhor os objetivos e entender todas as restrições do problema.

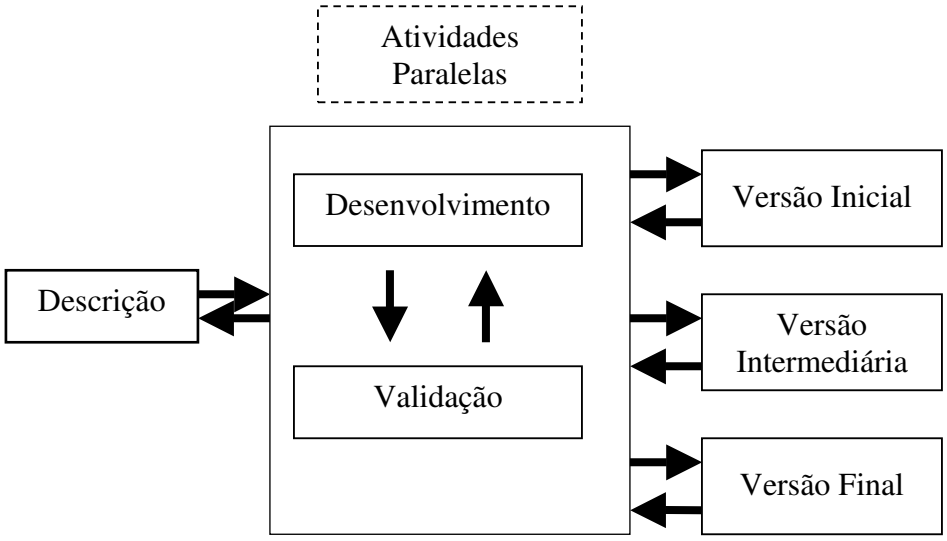

Figura 2.5: Desenvolvimento Evolutivo Exploratório

2.3.3.Descrição

Inicialmente foi enviada para a Unicamp uma base de dados real, porém reduzida, para permitir uma familiarização com os dados. Posteriormente, foi enviada toda a base de dados utilizada pelo departamento de Perdas Comerciais da AES Eletropaulo para o treinamento das ferramentas de classificação.

Em relação aos classificadores, uma das principais dificuldades encontradas na sua síntese é evitar que ocorra sobre-ajuste durante o treinamento, isto é, os classificadores ficam excessivamente moldados em função das exceções e ruídos contidos nos dados de treinamento. Obtém-se, assim, um excelente desempenho quando os classificadores são medidos em função dos próprios dados de treinamento. Porém, o desempenho cai bastante quando eles são testados com outros dados, ou seja, o classificador perde a capacidade de generalização, tornando-se específico para os dados de treinamento.

Um classificador com pouca capacidade de generalização não tem muita utilidade. O objetivo é conseguir uma ferramenta que responda corretamente para novos dados e não para dados antigos (utilizados no treinamento) cujas respostas já são conhecidas.

Em relação à qualidade da solução, ela depende tanto da qualidade da ferramenta de classificação como da qualidade dos dados de entrada. Além disso, uma ferramenta de classificação pode ter desempenho melhor para um determinado tipo de dados e pior para outro tipo. Existem também diferenças de desempenho entre as ferramentas de classificação quando se considera uma maior ou menor quantidade de dados.

Toda essa preocupação em entender as particularidades das empresas, em reproduzir o ambiente de produção da empresa, em estudar o seu processo corrente e em liberar versões intermediárias do produto vinculado ao projeto, é porque é necessário se aproximar ao máximo do ambiente real para que o projeto traga retorno, tanto para a AES Eletropaulo e ANEEL como para a sociedade em geral.

Além disso, esse projeto depende muito da participação dos técnicos da AES Eletropaulo, pois são eles que conhecem o problema a fundo, tendo uma vasta experiência acumulada, e também eles serão os futuros usuários da metodologia ou software a ser implementado.

2.4. Objetivo desta dissertação no contexto do projeto

O objetivo deste trabalho é avaliar quais ferramentas de classificação e bases de dados são mais promissores para melhorar o índice de acerto das inspeções da AES Eletropaulo.

Para tanto, buscou-se entender as particularidades da empresa e de seu ambiente de produção e estudar o seu processo corrente de combate às perdas comerciais. Deve-se salientar que o projeto em que se insere esta dissertação contempla o fato de que os produtos de software devem ser liberados em versões intermediárias, visando maximização de desempenho.

Conforme indicado na Figura 2.6, este trabalho propõe especificamente sintetizar e realizar testes com quatro diferentes ferramentas de classificação: C4.5, Redes Neurais Artificiais, Comparadores Naive Bayes e SVM. São considerados diferentes tipos de dados de entrada para descobrir qual (quais) ferramenta(s) combinada(s) com qual (quais) tipos de dados serão mais promissores para as outras fases do projeto.

Uma apresentação resumida de cada ferramenta de classificação será fornecida no Capítulo 4 desta dissertação.

A Figura 2.6 mostra que parte da base de dados da AES Eletropaulo foi trazida para a Unicamp (Base de dados filtrada). Essa parte da base de dados refere-se aos dados utilizados pelas ferramentas correntes do departamento de Perdas Comerciais. Para manipular esses dados, foi desenvolvida uma ferramenta para gerar os arquivos de entrada no formato de cada classificador a partir da base de dados. No final do processo, a saída de cada classificador foi analisada.

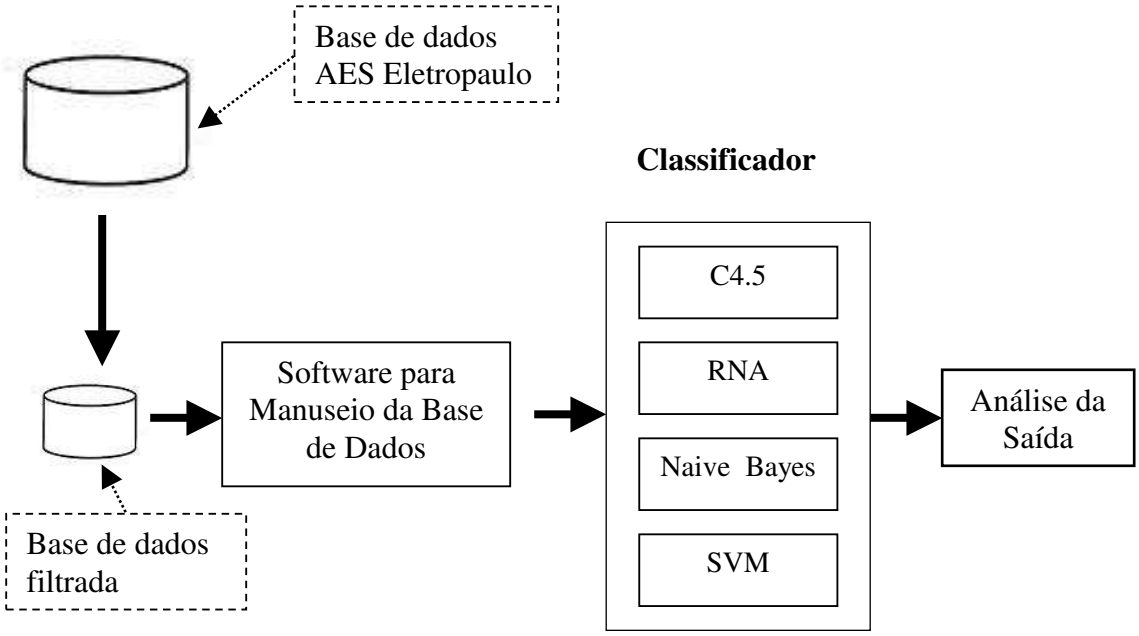

Figura 2.6: Acesso aos dados e ferramentas de classificação utilizadas neste trabalho

CAPÍTULO 3

Base de Dados e Metodologia

Este capítulo descreve os conjuntos de dados disponíveis e que servirão de entrada para as ferramentas de classificação, os tipos de atributos existentes nesses conjuntos de dados, as classes de saída (comumente utilizadas) para as ferramentas de classificação e a metodologia que foi utilizada para a medição dos resultados e comparação das ferramentas de classificação.

3.1. Base de Dados

São duas as fontes de dados para os algoritmos de aprendizado de máquina utilizados:

- base de dados com atributos dos clientes;

- resultado das inspeções.

3.2. Conjunto de Dados de Entrada

Este trabalho utiliza três conjuntos de dados de entrada:

a) Série

É uma série temporal composta pela quantidade de energia elétrica em kWh consumida pelos clientes nos últimos 59 meses. Esta é a quantidade de meses com que o departamento de perdas comerciais da AES Eletropaulo trabalha no dia-a-dia e, se fosse necessário, seria possível ter um histórico mais longo. Porém, isso exigiria um esforço maior dos técnicos de informática da empresa.

b) Características

Este conjunto é composto por características extraídas das séries de consumo. As características são novos atributos extraídos a partir dos dados crus, os quais correspondem às quantidades de energia mensal consumida. O objetivo é propor atributos mais informativos do que os originais. Por exemplo, ao tratar os consumos mensais de maneira independente perde-se a relação de ordem entre eles, isto é, as ferramentas de classificação propostas não são capazes de perceber que o consumo do mês de agosto, por exemplo, é anterior ao mês de setembro e posterior ao mês de julho, considerando-se o mesmo ano. Para as ferramentas, tanto faz comparar, por exemplo, os consumos de julho/2007 com junho/2006 ou dezembro/2005. Porém, a ordem pode ser incorporada direta ou indiretamente com a geração de novos atributos.

c) Genérico

Este é um conjunto de atributos que normalmente já são utilizados pelos programadores de inspeções da AES Eletropaulo para auxiliá-los a determinar os locais a serem inspecionados. Estes atributos podem ser classificados como dinâmicos e estáticos, conforme eles são alterados ou não, respectivamente, no decorrer do tempo.

Esses conjuntos de dados são discutidos nas subseções a seguir.

3.2.1. Conjunto Série

Há diversos fatores que influenciam no consumo de energia elétrica, como as condições meteorológicas (se está frio, é gasta mais energia com aquecedor e chuveiro, enquanto que em dias quentes o ar-condicionado é ligado) e a compra de um novo eletrodoméstico, que pode acarretar tanto aumento (compra de um novo aparelho) como diminuição do consumo (no caso de substituição por um aparelho mais econômico). Além disso, há o período de férias, em que ocorre queda de consumo porque a família viajou ou há aumento de consumo porque recebeu visitas.

Assim, pode-se observar que existem inúmeros fatores que influenciam nas séries de consumo e esses fatores são difíceis de serem monitorados. Mas, mesmo assim, este projeto partiu da idéia de analisar mudanças de perfis nas séries de consumo com o objetivo de detectar fraudes e anomalias, pois trabalhos anteriores como JIANG et al. (2002) e COMETTI & VAREJÃO (2005) já aplicaram séries de consumo em perdas comerciais. Este último trabalho foi realizado com dados da Espírito Santo Centrais Elétricas (ESCELSA) e os resultados dos testes aplicados são fornecidos no Capítulo 4.

3.2.2. Conjunto Características

A palavra "extração" no contexto de nosso problema consiste em considerar outros valores ou medidas estatísticas a partir dos valores originais que representam uma instância do problema (cliente ou usuário). Assim, por exemplo, a partir dos valores do consumo elétrico de um determinado cliente, é possível extrair características ou atributos que o representem. A quantidade ou dimensão de atributos extraídos pode ser menor, igual ou maior do que a quantidade ou dimensão dos valores originais. Sendo que o número de atributos é independente da quantidade de valores de consumo do cliente. Por exemplo, poderíamos ter 59 ou mais observações de consumo de um cliente e apenas 17 atributos associados. Em aprendizado de máquina, busca-se trabalhar com o menor número de atributos, o que implica que cada atributo deve agregar um alto grau de significado para a aplicação pretendida. Em problemas de classificação, por exemplo, ser um atributo de alto significado implica ter alta capacidade discriminante.

Neste trabalho, foram extraídas apenas 17 características das séries de consumo, principalmente através do procedimento de detecção de regimes. Outros trabalhos (COMETTI & VAREJÃO, 2005; JIANG et al., 2002) também fizeram extração de características de séries de consumo elétrico, porém as características foram baseadas em outras informações, como coeficientes de amplitude e fase da série de Fourier, coeficientes de wavelets e coeficientes de polinômios aproximados pelo método de quadrados mínimos. Essas outras características citadas neste parágrafo não foram utilizadas neste trabalho. No entanto, poderiam ter sido utilizadas para comparar os desempenhos.

A seguir, será explicado cada um dos 17 atributos extraídos.

Atributo 1º: Número de regimes

Este atributo conta o número de regimes detectados na série de consumo do cliente. Na Figura 3.1, mostra-se um exemplo em que foram detectados 5 regimes, que é a quantidade de patamares da curva de regimes.

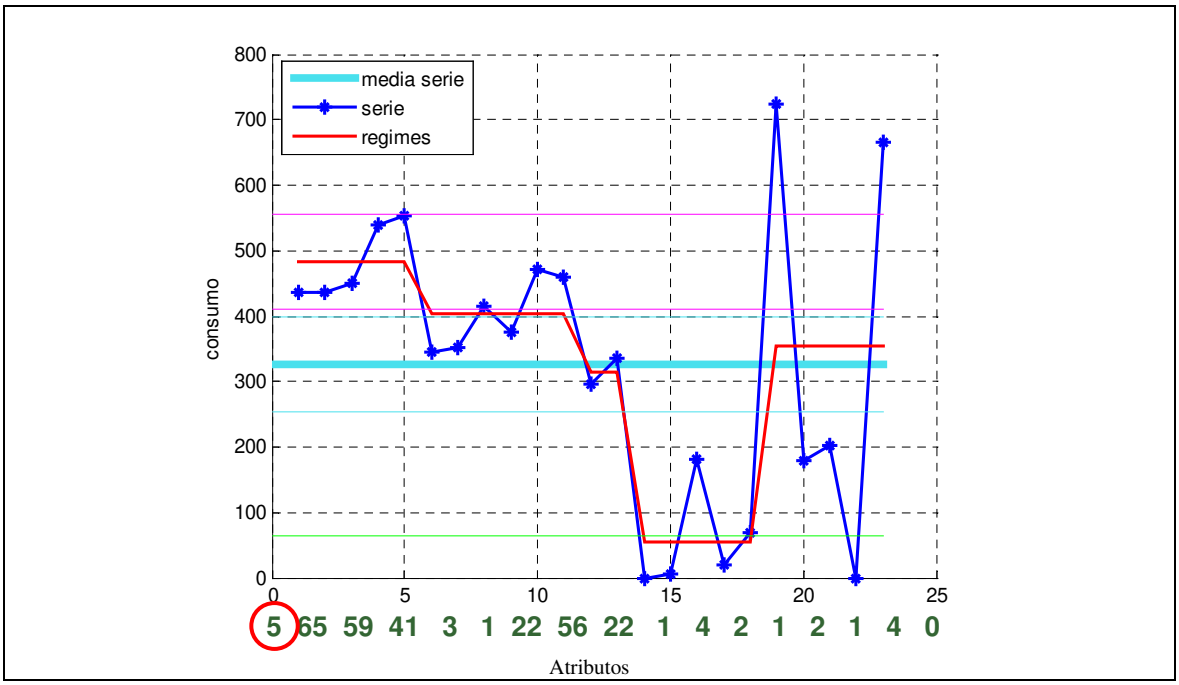

Figura 3.1: Número de regimes.

Atributo 2º: Coeficiente de Variação

O segundo atributo é o coeficiente de variação (CV), que é uma estatística adimensional (independente da escala da série). Seu cálculo é dado pela seguinte equação:

$$CV = \frac{100 \times DesvioPadrão}{média}$$

Tanto o desvio padrão como a média de consumo empregados para se obter CV foram calculados sobre um período de 24 meses. O desvio padrão mede a variabilidade da série e ele

é dividido pela média para se conseguir um valor relativo que permite comparar esse atributo entre séries diferentes.

Na Figura 3.2, é apresentado um exemplo em que as observações estão mais afastadas do valor médio da série e, na Figura 3.3, é apresentado outro exemplo em que as observações não se afastam muito da média, refletindo num valor bem menor para o atributo.

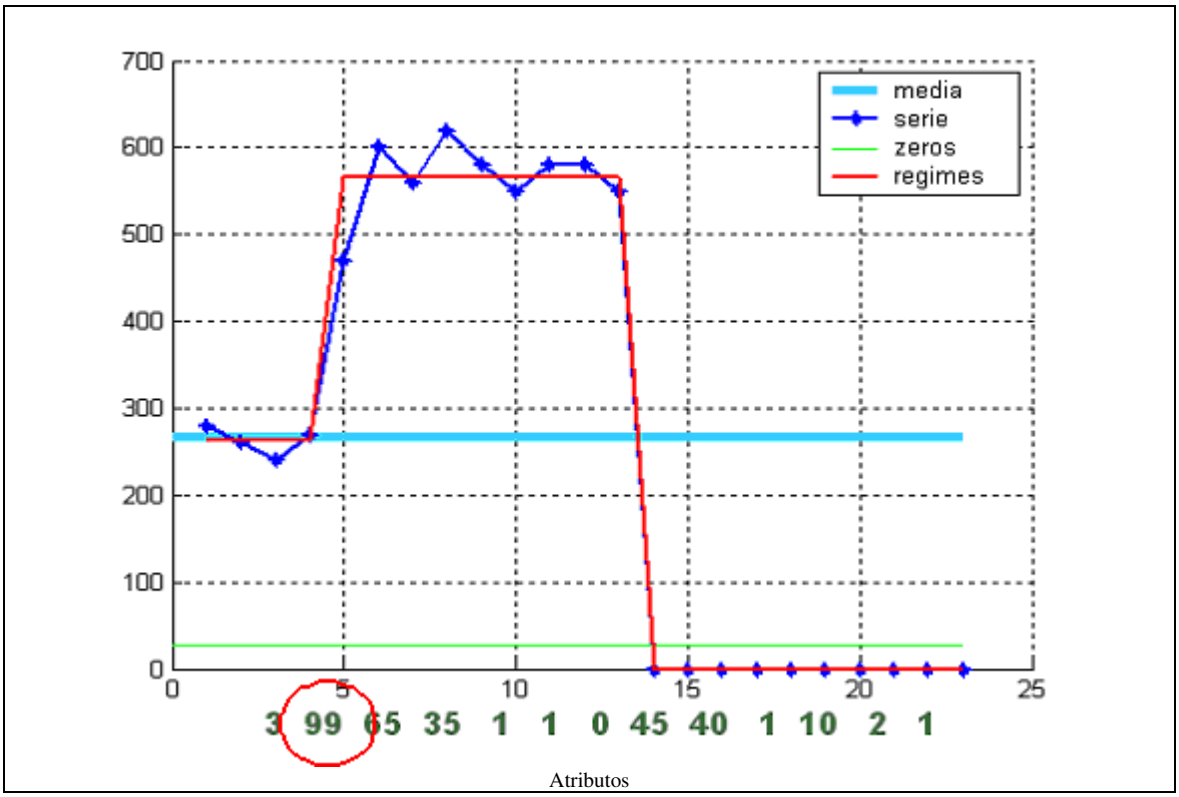

Figura 3.2: Coeficiente de variação com as observações mais afastadas do valor médio da série.

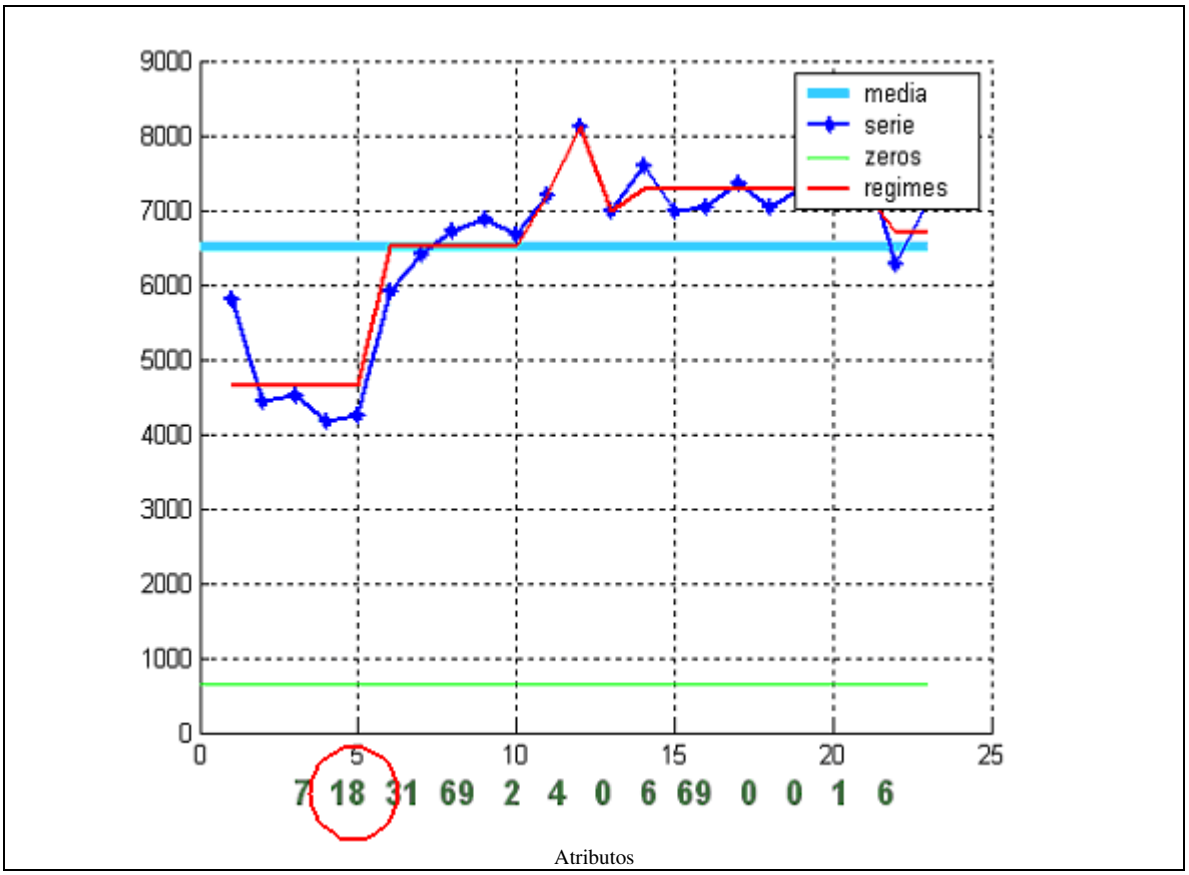

Figura 3.3: Coeficiente de variação com as observações menos afastadas do valor médio da série.

Atributos 3º e 4º: Porcentagem das Quedas e Aumentos em Relação ao Regime

Os terceiro e quarto atributos medem, respectivamente, as porcentagens de quedas e aumentos relativos, ou seja, cada vez que se experimenta uma quebra de regime verifica-se se esta quebra produziu queda ou aumento. Se as porcentagens de queda e de aumento forem iguais, então o regime final coincide com o regime inicial. O exemplo da Figura 3.4 mostra que a porcentagem de quedas acabou sendo maior do que a porcentagem de aumentos. Isto implica que, ao final, o cliente experimentou uma tendência de queda no consumo global.

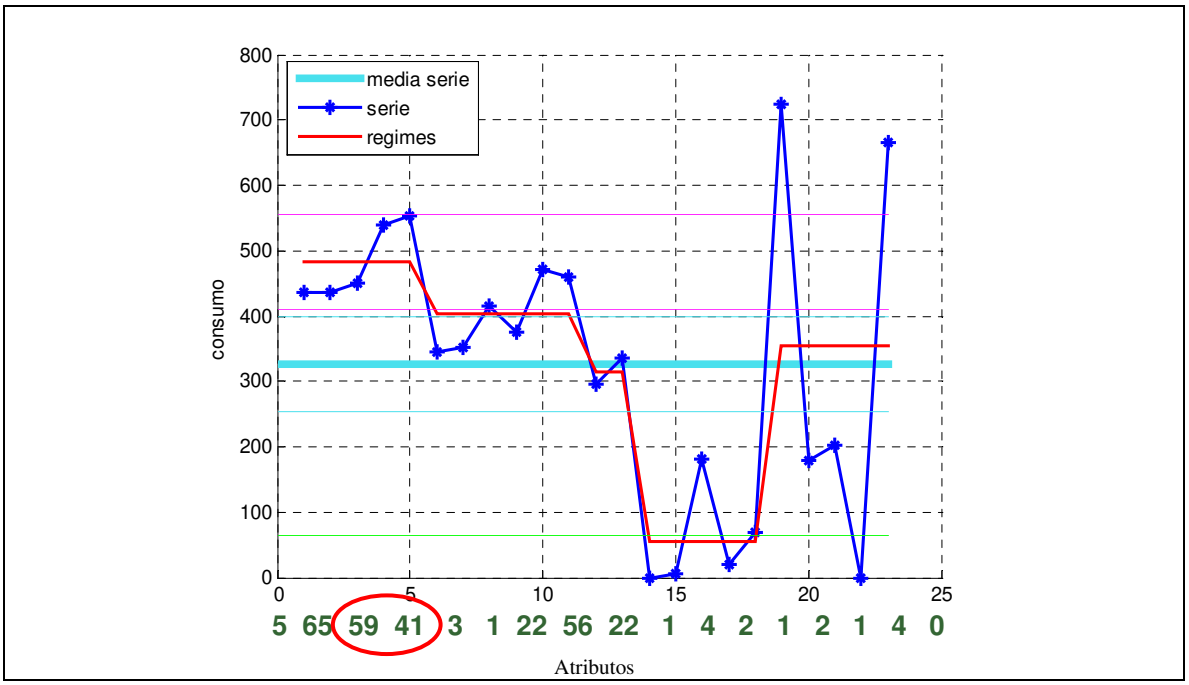

Figura 3.4: Porcentagens de quedas e aumentos de consumo em relação ao regime.

Atributos 5º e 6º: Número de Quedas e Aumentos

Estes atributos guardam certa semelhança com os dois anteriores, com a diferença de que os anteriores medem a porcentagem e estes dois apenas contam o número de quedas e de aumentos. Pode haver casos em que ocorram mais quedas do que aumentos e, mesmo assim, haja tendência de alta ao final. Na Figura 3.5, são ilustrados esses dois atributos.

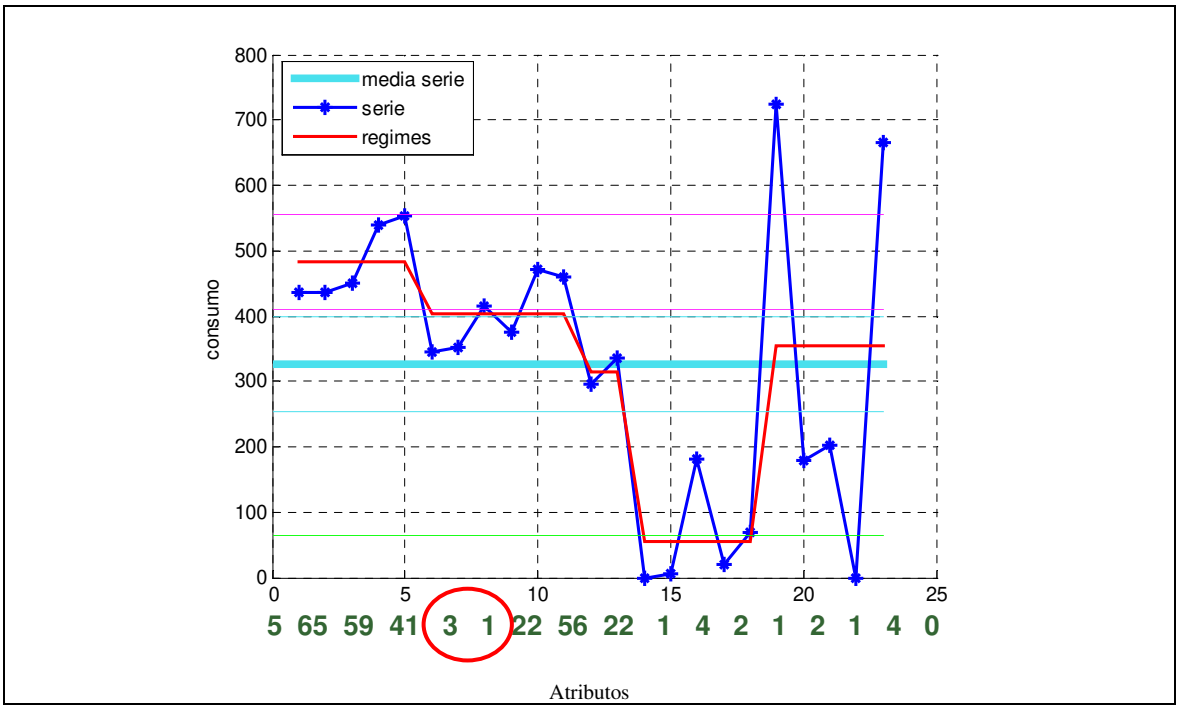

Figura 3.5: Número de quedas e aumentos.

Atributos 7º, 8º e 9º: Porcentagem do Tempo no Regime Inicial, nas Quedas e nos Aumentos.

Estes três atributos medem as porcentagens em relação ao tempo total (no caso 24 observações de consumo) em que o consumo esteve no regime inicial, em regimes de queda relativa e em regimes de aumento relativo. O exemplo da Figura 3.6 mostra estes atributos. Nele, vemos que o único regime de aumento relativo foi o regime final, o qual representou 22% e que coincidiu com a porcentagem do regime inicial.

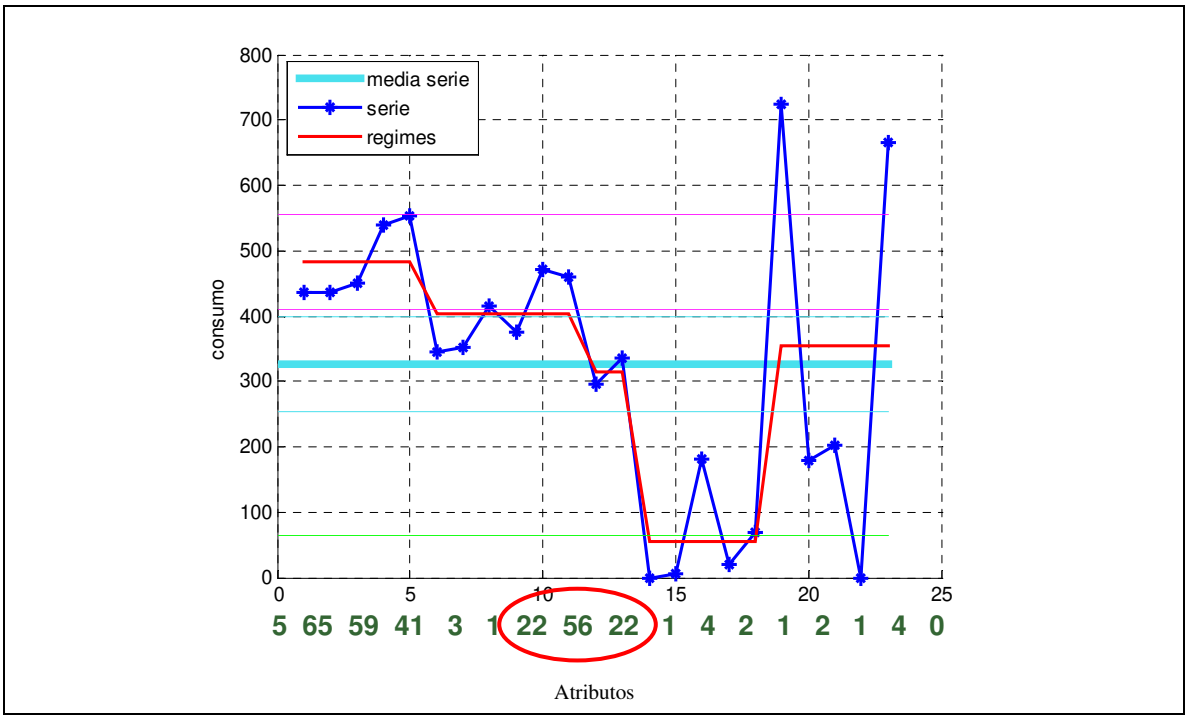

Figura 3.6: Porcentagem do tempo no regime inicial, nas quedas e nos aumentos.

Atributos 10º e 11º: Presença de Zeros e Número de Zeros

O décimo atributo assume valor binário 0 ou 1, indicando a ausência ou presença de valores considerados como zero nas observações dos consumos de um cliente. O décimo primeiro atributo informa o número de zeros considerados. Visando inserir certa folga na decisão de considerar uma observação como zero, considera-se uma faixa no eixo dos valores do consumo. No caso, considerou-se 10% da faixa que compreende desde o valor zero até a média da série. Assim, observações que estiverem nessa faixa serão consideradas como zero. A Figura 3.7 ilustra um exemplo em que a faixa zero está abaixo da linha verde e foram consideradas 4 observações como zero. Se o atributo 11º é zero, significa que não há presença de zeros.

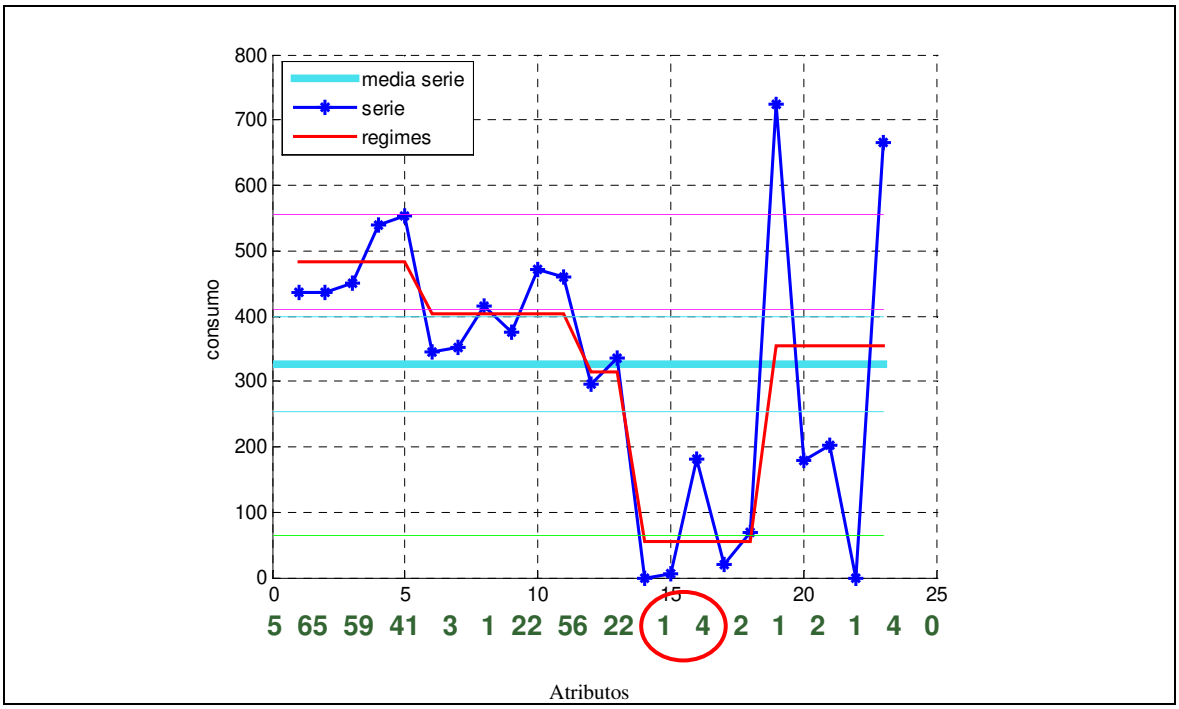

Figura 3.7: Presença de zeros e número de zeros.

Atributos 12º, 13º e 14º: Número de Regimes na Faixa da Média, Abaixo da Faixa da Média e Acima da Faixa da Média.

É definida uma faixa em torno do valor médio da série de consumo, valendo 10% da diferença entre a amplitude máxima e a mínima. Estes três atributos contabilizam o número de regimes que estão na faixa, abaixo da faixa e acima da faixa da média, conforme o exemplo da Figura 3.8.

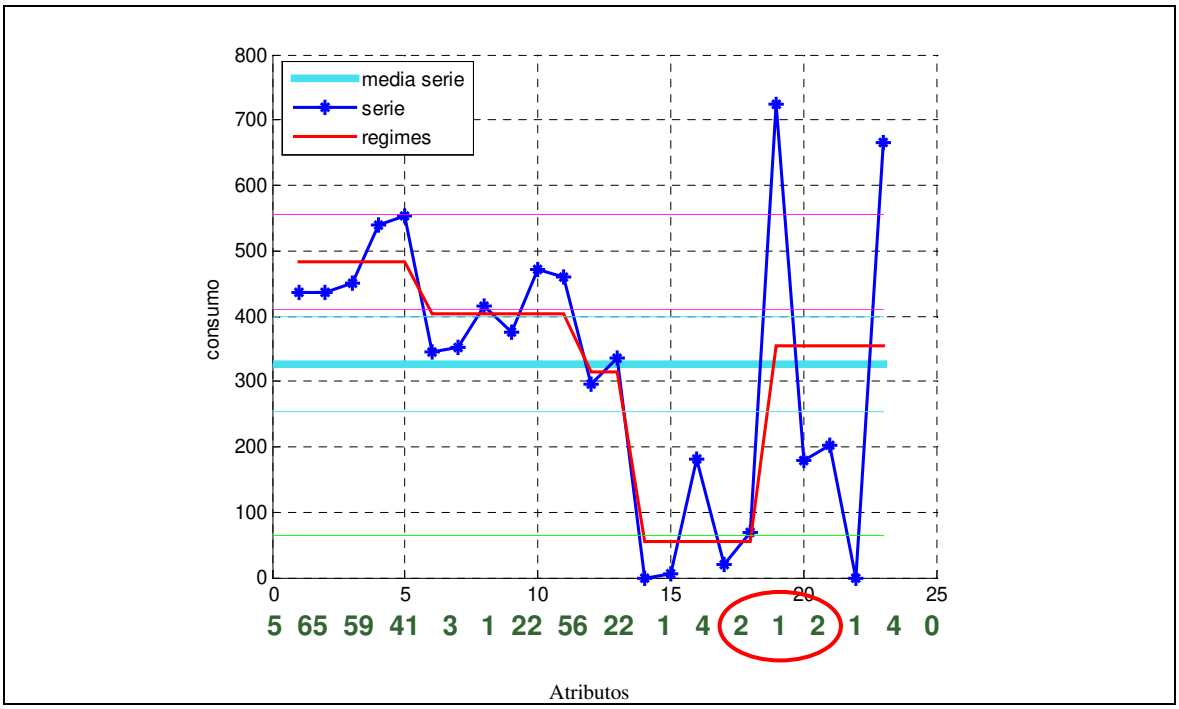

Figura 3.8: Número de regimes na faixa, abaixo da faixa e acima da faixa da média.

Atributos 15º, 16º e 17º: Número de Regimes na Faixa do Regime Inicial, Abaixo da Faixa do Regime Inicial e Acima da Faixa do Regime Inicial.

Análogo aos três atributos anteriores, dessa vez é definida uma faixa em torno do regime inicial detectado, com o valor de 10% da amplitude total. Estes três atributos contabilizam o número de regimes que estão na faixa, abaixo da faixa e acima da faixa do regime inicial, respectivamente, como exemplificado na Figura 3.9. No caso, não existem regimes acima da faixa do regime inicial.

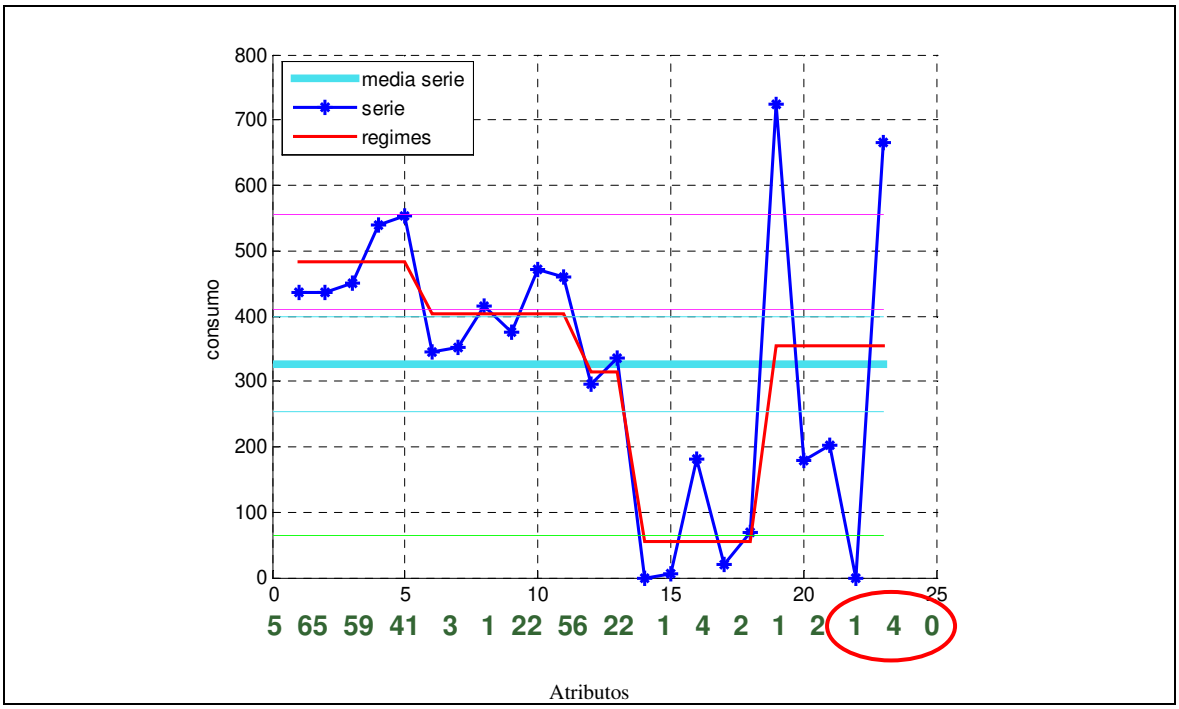

Figura 3.9: Número de regimes na faixa, abaixo da faixa e acima da faixa do regime inicial.

3.2.3. Conjunto Genérico

Além das séries de consumo e de suas características, resolveu-se também trabalhar com outros atributos, os quais já são utilizados pela AES Eletropaulo. Abaixo, a lista desses atributos:

- Unidade: os clientes foram agrupados em 6 grandes regiões geográficas (Oeste, Sul, Anhembi, Centro, Leste, ABC).
- Local: os clientes foram agrupados em 85 regiões geográficas.
- Classe: residencial, comercial, industrial, rural, poder público, iluminação pública, serviço público e consumo próprio.
- Atividade: 1174 atributos distintos. Exemplo: extração de sal-marinho e sal-gema, produção de ferro-gusa e ferro-esponja, fabricante de máquinas para indústria de marmoraria e artefatos de cimento e olarias cerâmicas.

- Situação: consumidor ligado com medidor, ligado sem medidor, desligado com medidor, desligado sem medidor e excluído do cadastro.
- Tipo de Ligação: definitiva com medição, definitiva sem medição, provisória com medição e provisória sem medição.
- Irregularidade de leitura: possui 99 códigos. Exemplos: mudança do nome ou número do logradouro, insetos dentro ou envolvendo o medidor.
- Consumo Irregular: atributo binário que indica se o consumidor tem ou não histórico de fraude.
- Religamentos: indica se o consumidor já se auto-religou após ter sua ligação cortada pela companhia. Valores: 0 – nenhuma auto-religação; 1 – uma auto-religação; 2 – mais do que uma auto-religação.
- Faturamento: monofásico, bifásico ou trifásico.
- Quantidade de carga declarada: carga que o cliente indica no momento da ligação, que está relacionada com a quantidade de aparelhos ou máquinas que ele possui.
- Média diária de consumo em kWh.
- Valor da última leitura em kWh.

Observe que, nesse conjunto, o atributo "valor da última leitura" foi extraído diretamente da série de consumo e "média diária de consumo" é uma característica também extraída da série de consumo.

3.3. Classes de saída

As inspeções podem ser classificadas em quatro classes, as quais já foram descritas no Capítulo 2: Normal, Fraude, Anomalia e Não-Inspecionado.

É objetivo do projeto também diminuir o número de Não-Inspecionados. No entanto, este trabalho não aborda este tópico, pois é necessário fazer um levantamento junto à Eletropaulo sobre as razões do código de retorno "Não-Inspecionado", como casa fechada ou

casa demolida. Neste trabalho, o código de retorno "Não-inspecionado" simplesmente significa que não se sabe qual a real situação do consumidor.

É necessário entender primeiro as razões do "Não-inspecionado" para depois analisar se existe a necessidade de fazer uma busca inteligente na base de dados. Por exemplo, o "Não-inspecionado" poderia ser um problema de procedimento do tipo "cadastro do cliente não-atualizado".

Também as inspeções rotuladas como Anomalia não foram tratadas por esse trabalho, pois é necessária uma investigação prévia a respeito dos atributos que influem nesta classe. A classe Anomalia só foi utilizada nos estudos relatados na seção do Capítulo 4 referente ao algoritmo C4.5, que mostram justamente a importância de se distinguir bem a classe Anomalia.

3.4. Tipos de atributo

De acordo com KLÖSGEN & ZYTKOW (2002), os atributos podem ser distinguidos por seus tipos: numérico e categórico, sendo que os atributos categóricos ainda são subdivididos em: ordinal, binário e nominal.

Os atributos numéricos contêm valores pertencentes ao conjunto dos números reais, por exemplo, o consumo mensal de energia elétrica. Os atributos nominais têm um domínio de valores de cardinalidade finito, onde não existe uma implicação de ordem natural entre os valores desse domínio. Por exemplo, região, raça e estado civil. Os valores do atributo região, Norte, Sul, Leste e Oeste, poderiam ser codificados respectivamente em 1, 2, 3 e 4. No entanto, esses números são meramente identificadores e não têm nenhum significado ordinal, pois não se pode dizer que o Oeste é maior do que o Leste, ou que o Norte está mais próximo do Sul do que do Oeste.

Diferentemente dos atributos numéricos, nos atributos categóricos não é possível aplicar as operações matemáticas tradicionais (como soma e multiplicação). Assim, a diferenciação dos atributos é importante para que o algoritmo saiba como armazenar e tratar os dados.

O atributo binário é um atributo nominal cujo domínio contém apenas dois valores. A sua diferenciação é importante pois, além do indicativo do seu tamanho, a existência de apenas dois valores pode influenciar o seu processamento, dependendo do algoritmo utilizado. Já os atributos ordinais são aqueles cujos rótulos possuem uma ordem, por exemplo: ruim, razoável, bom e excelente.

No conjunto de entrada Série, os atributos são todos numéricos. No conjunto Características, os atributos são numéricos ou binários. E no conjunto Genérico há atributos nominais, numéricos e binários. Os atributos do conjunto de saída são nominais.

3.5. Métricas

Uma métrica muito utilizada é a taxa de acerto, apresentada na Equação 3.1, a qual divide a quantidade de acertos ($QtdAcerto$), pela quantidade total de inspeções ($QtdTotal$). Porém, quando as classes de saída são desbalanceadas, existem métricas mais apropriadas, como é explicado ainda neste tópico.

$$tx_{acerto} = \frac{QtdAcerto}{QtdTotal}. \tag{3.1}$$

A Figura 3.10 mostra a distribuição das classes “Normal” e “Fraude”, extraída das inspeções da AES Eletropaulo entre 2005 e 2007. Como se pode perceber, essa distribuição é bastante desbalanceada: 85% : 15%.

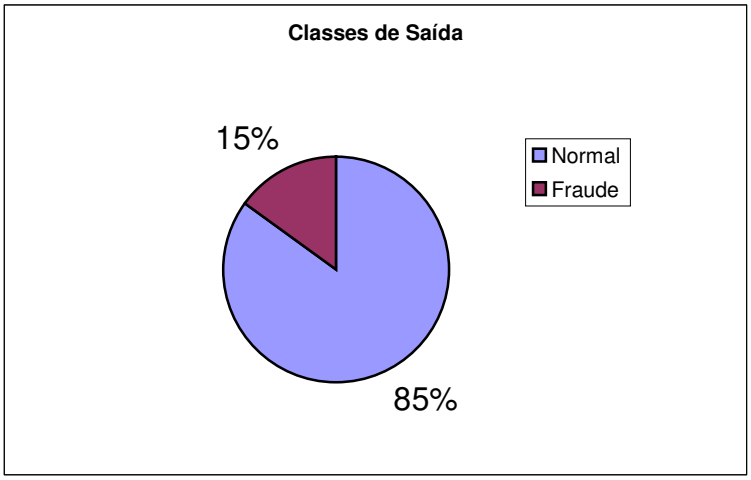

Figura 3.10: Distribuição da classe de saída relativa às inspeções entre 2005 e 2007.

Se for utilizada a taxa de acerto da Equação 3.1 como métrica, um classificador trivial (aquele que classifica todas as amostras como sendo da classe predominante) conseguiria uma taxa de acerto de 85%, bastando classificar todos os dados como normais, sem detectar nenhum suspeito.

É claro que isto não é interessante, pois a classificação dos dados como normais não forneceria qualquer elemento para identificação de fraudes, de modo que novas métricas foram definidas para poder medir o desempenho dos classificadores. Logo, para avaliar o desempenho dos classificadores, foram utilizadas métricas derivadas da "matriz de confusão" da Tabela 3.1 (PROVOST & KOHAVI, 1988).

A matriz de confusão a seguir exibe duas classes, *n* (Normal) e *f* (Fraude), a quantidade de consumidores normais classificados corretamente como normais ($Q_{n\to n}$) e como fraudulentos ($Q_{n\to f}$), e a quantidade de consumidores fraudulentos classificados como normais ($Q_{f\to n}$) e como fraudulentos ($Q_{f\to f}$).

Tabela 3.1: Definição da matriz de confusão

Matriz de Confusão		Classe Predita	
		n	f
Classe Real	n	$Q_{n \to n}$	$Q_{n \to f}$
	f	$Q_{f \to n}$	$Q_{f \to f}$

A taxa de acerto da matriz de confusão é calculada conforme a Equação 3.2:

$$tx_{acerto} = \frac{Q_{n \to n} + Q_{f \to f}}{Q_{n \to n} + Q_{n \to f} + Q_{f \to f} + Q_{f \to n}}. \tag{3.2}$$

As novas métricas utilizadas quando se trabalha com classes desbalanceadas são:

- especificidade
- confiabilidade

A especificidade é a razão entre o número de fraudadores corretamente classificados e o número total de fraudadores existentes, segundo a Equação 3.3:

$$especificidade = \frac{Q_{f \to f}}{Q_{f \to f} + Q_{f \to n}}. \tag{3.3}$$

Já a confiabilidade é a razão entre o número de suspeitos corretamente classificados e o número total de casos classificados como suspeitos, segundo a Equação 3.4:

$$confiabilidade = \frac{Q_{f \to f}}{Q_{f \to f} + Q_{n \to f}}. \tag{3.4}$$

A especificidade proporciona uma noção da cobertura do classificador, isto é, o percentual do conjunto de fraudadores que o classificador está conseguindo identificar. Por outro lado, a confiabilidade proporciona uma noção de precisão das inspeções, ou seja, o percentual de sucessos na identificação de reais fraudadores no total de inspeções recomendadas.

Idealmente, o classificador de melhor desempenho seria aquele que maximizasse essas duas métricas, obtendo 100% para as duas métricas, o que significa 100% de acerto também. Contudo, na prática o que geralmente ocorre é melhorar uma métrica e piorar a outra, como é explicado no exemplo a seguir.

Uma empresa tem 5 milhões de clientes e estima-se que 2,5% dos clientes são fraudadores, totalizando 125 mil. Suponha que a empresa que saiba que uma determinada região (com 1000 clientes) tem alto índice de perda comercial (medido pela diferença entre consumo fornecido e faturado) e que ela inspecione todos os clientes dessa região e detecte 500 fraudadores. Tem-se, portanto, uma confiabilidade de 50% (500 / 1000) e uma especificidade de 0,40% (500 / 125 mil). A confiabilidade é alta, significando que a empresa fez poucas inspeções desnecessárias. Porém, a especificidade é muito baixa, detectando apenas 500 fraudadores dos 125 mil estimados.

Diante disso, a empresa precisa fazer mais inspeções para detectar mais fraudadores. Contudo, ela não tem nenhum outro indício de onde estão os fraudadores, precisando escolher aleatoriamente os novos lugares a serem inspecionados. Se fosse possível à empresa inspecionar todos os clientes, ela obteria uma especificidade de 100% (pois detectaria todos os fraudadores). No entanto, a confiabilidade seria de apenas 2,5% (125 mil / 5 milhões).

Quanto mais se inspeciona, é natural que a especificidade aumente, pois ela nunca diminui. Todavia, torna-se mais difícil acertar as inspeções, fazendo com que a confiabilidade diminua. Portanto, ao fazer as inspeções é necessário monitorar a confiabilidade para que ela não diminua a ponto da receita recuperada ser menor do que o custo das inspeções.

A utilização de ferramentas inteligentes em Perdas Comerciais visa detectar os locais mais promissores para serem inspecionados, de modo que a confiabilidade não caia tanto à medida que as inspeções sejam feitas.

Utilizando-se o classificador trivial (todas as amostras são rotuladas com a classe dominante) em cima dos dados da Figura 3.10, nenhum suspeito é identificado. Assim, a especificidade e a confiabilidade são iguais a zero, pois $Q_{f \to f}$ é zero, embora a taxa de acerto seja 85%.

A taxa de erro reportada pela AES Eletropaulo é a própria confiabilidade (número de fraudadores detectados / número de inspeções). Entretanto, não é possível saber com exatidão a especificidade, pois não se sabe quantos clientes fraudulentos realmente existem. Apenas se conhece a quantidade de energia não faturada.

A Figura 3.11 mostra o processo de análise de dados. *S1* é o conjunto de todos os clientes da AES Eletropaulo, sendo que não é possível saber a quantidade de clientes fraudulentos (F) e clientes Normais (N) em *S1*. *S2* é conjunto de inspeções gerado pelos classificadores da AES Eletropaulo contendo os clientes suspeitos de fraude, sendo que a confiabilidade é determinada dividindo o tamanho do sub-conjunto F de *S2* pelo tamanho de *S2* (ou a área de F de *S2* dividida pela área de *S2*). No entanto, não é possível determinar a especificidade de *S2*, como já dito no parágrafo anterior, pois é necessário saber o tamanho do sub-conjunto F de *S1* (e não de *S2*). Suponha que os conjuntos *S3* e *S4* sejam gerados por dois classificadores distintos sobre o conjunto de inspeções *S2*. Neste caso, é possível calcular, além da confiabilidade, a especificidade, pois é conhecido o tamanho do conjunto F de *S2*.

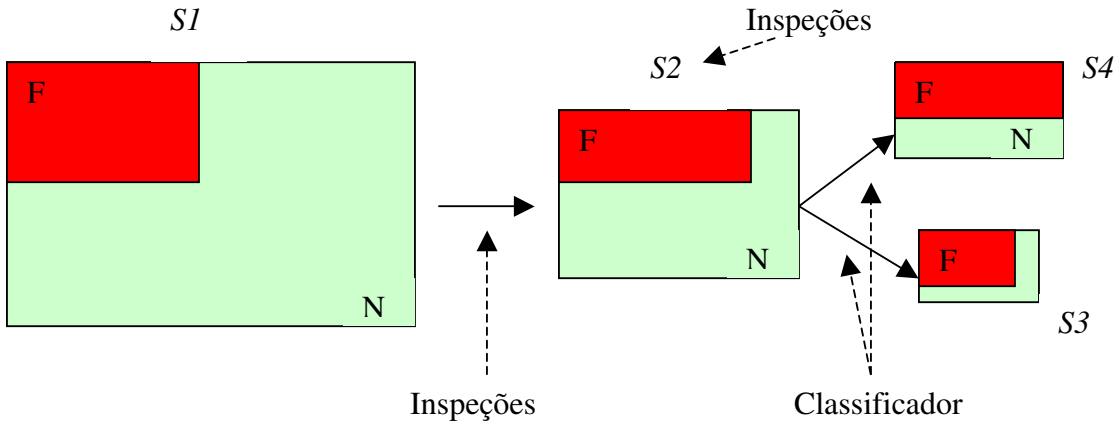

Figura 3.11: Processo de análise dos dados. Rótulo dos clientes: F (fraude) / N (normal). Conjuntos: S1 (clientes), S2 (inspeções), S3 e S4 (saídas do classificador).

Este trabalho consiste em utilizar o conjunto de inspeções *S2* como entrada e sintetizar classificadores para gerar conjuntos como *S3* e *S4*, sendo que a especificidade e a confiabilidade são utilizadas para comparar o desempenho das ferramentas de classificação.

Conforme já explicado neste tópico, é observado na Figura 3.11 que a confiabilidade do conjunto *S2* é maior do que de *S1* e as dos conjuntos *S3* e *S4* são maiores do que *S2*. Isto é o esperado, pois em caso contrário, as ferramentas de classificação não estariam funcionando adequadamente. Observe também que a confiabilidade de *S3* é maior que a de *S4*. No entanto, a especificidade é menor, pois em *S3* há menos fraudes do que em *S4*. Assim sendo, não é possível estabelecer uma figura de mérito entre os desempenhos dos classificadores que geraram *S3* e *S4*, pois um conseguiu uma confiabilidade melhor e o outro uma especificidade melhor.

A confiabilidade e a especificidade são métricas comumente encontradas em trabalhos a respeito de fraude, seja no setor elétrico (COMETTI & VAREJÃO, 2005) como na detecção de spams em sistemas computacionais de mensagens eletrônicas (TRETYAKOV, 2004), em que a confiabilidade é calculada pela quantidade de spams detectados dividida pela quantidade de mensagens filtradas, e a especificidade é calculada pela quantidade de spams detectados dividida pela quantidade total de spams.

CAPÍTULO 4

Teste e Análise das Ferramentas Computacionais

Este capítulo descreve e analisa os resultados dos testes realizados com as ferramentas computacionais C4.5, RNA, Naive Bayes e SVM para os diferentes conjuntos de entradas.

4.1. Especificação dos Testes

Os conjuntos de entrada que foram utilizados são:

- Série;
- Características;
- Genérico;
- Os três conjuntos anteriores simultaneamente.

Caso não haja nenhuma observação no próprio teste, são válidas as seguintes especificações:

- o treinamento foi feito com os dados referentes às inspeções de julho de 2006 a dezembro de 2006;
- os testes foram executados com os dados referentes às inspeções de janeiro de 2007 a outubro de 2007;
- apenas os dados referentes às inspeções residenciais foram considerados;
- só as inspeções com código de retorno Fraude ou Normal foram utilizadas.

A Tabela 4.1 mostra a quantidade de inspeções separadas em função dos seus rótulos e os índices de confiabilidade, apurados para cada período, empregando a Equação 3.4.

Tabela 4.1: Quantidade de inspeções utilizadas nos testes e a confiabilidade calculada para cada período.

Qtd Inspeções	jul-dez 2006	jan - out 2007
Normal	125.224	126.172
Fraudes	16.565	20.344
Confiabilidade	11,7	13,9
Total	**141.789**	**146.516**

4.2. C4.5

4.2.1.Introdução

O C4.5 (QUINLAN, 1993) é um algoritmo de indução de árvores de decisão, cujo predecessor é o ID3 (QUINLAN, 1979). Embora já se tenha lançado o C5.0, o C4.5 foi utilizado nos testes porque o seu código fonte está disponível, enquanto que o C5.0 é um software comercial. O C4.5 é um dos algoritmos mais utilizados para a construção de árvores de decisão.

A Figura 4.1 mostra uma árvore de decisão, que é uma estrutura que contém:

- folha(s), indicando uma classe;
- nó(s) de decisão, que define(m) algum teste sobre o valor de um atributo específico, com um ramo e sub-árvore para cada um dos valores possíveis do teste.

A árvore de decisão pode ser usada para classificar uma amostra iniciando-se pela raiz da árvore e movendo-se através dela até que uma folha seja encontrada. A cada nó de decisão, a saída do teste de decisão é determinada e inicia-se o processo pela raiz da sub-árvore correspondente a essa saída.

Um mesmo conjunto de dados pode gerar várias árvores de decisão distintas. Assim, usando o exemplo da Figura 4.1, o nó raiz poderia ser "atividade" em vez de "local", fazendo com que o nó "local" passe a ocupar uma outra posição na árvore. Essa troca de nós faz com que a árvore se torne maior ou menor, isto é, pode ser necessário percorrer um caminho maior ou menor para se chegar a uma decisão. Na construção da árvore de decisão, procura-se associar a cada nó de decisão o atributo mais informativo entre aqueles ainda não utilizados no caminho desde a raiz da árvore. No entanto, cada algoritmo tem a sua própria metodologia para distinguir o atributo mais informativo, fazendo com que a topologia da árvore e a qualidade da árvore variem em função do algoritmo utilizado.

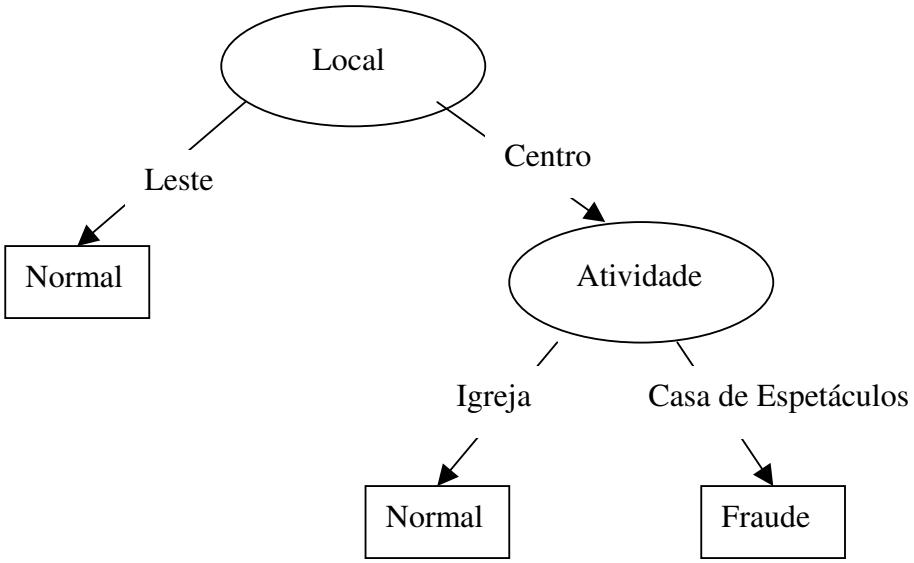

Figura 4.1: Exemplo fictício de árvore de decisão, tomando atributos presentes na base de dados da Eletropaulo.

Assim, existem várias propostas para se construir a árvore de decisão, entre elas, o C4.5. Ele é do tipo guloso (isto é, executa sempre o melhor passo avaliado localmente, sem se preocupar se este passo, junto à seqüência completa de passos, vai produzir a melhor solução ao final) e do tipo "dividir para conquistar" (partindo da raiz, criam-se sub-árvores até chegar nas folhas, o que implica em uma divisão hierárquica do problema de decisão original em múltiplos sub-problemas de decisão, os quais tendem a ser mais simples).

O C4.5 suporta tanto dados numéricos como dados categóricos, simultaneamente. Assim, este algoritmo não impõe nenhuma limitação em relação ao tipo de dado de entrada,

nem em relação à quantidade de dados. Além disso, o C4.5 suporta também dados faltantes e realiza a operação de poda. A operação de poda impede o crescimento da árvore quando o teste de significância estatística não encontra associação entre qualquer atributo e a classe de um nó em particular. O objetivo da poda é evitar o sobre-ajuste.

A metodologia utilizada pelo C4.5 para escolher o atributo a ser testado em cada nó de decisão está baseada no conceito de entropia. A Equação 4.1 mostra como é feito o cálculo da entropia, sendo que *S* é o conjunto de amostras, *freq*(C_j, S) é o número de amostras em *S* que pertencem à classe C_j e $|S|$ é o número de amostras em *S*.

$$entropia(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} * \log_2(\frac{freq(C_f, S)}{|S|}) \tag{4.1}$$

O termo $freq(C_j, S)/|S|$ é a probabilidade de se obter a classe C_j ao se escolher aleatoriamente uma amostra de *S*. Já o segundo termo, $-\log_2(freq(C_j, S)/|S|)$, é uma medida, em bits, da quantidade de informação carregada pela mensagem. A constante *k* está associada ao número de classes.

Agora, suponha que exista um teste X que particione o conjunto de treinamento *T* em sub-conjuntos $\{T_1, T_2, ..., T_n\}$. Se o teste é para ser avaliado sem explorar as divisões subsequentes de T_i, a única informação de orientação disponível é a distribuição das classes em *T* e seus subconjuntos, e uma medida similar de entropia é dada pela Equação 4.2.

$$entropia_X(T) = \sum_{j=1}^{n} \frac{|T_j|}{|T|} * entropia(T_j) \tag{4.2}$$

A Equação 4.3 fornece o ganho conseguido ao particionar T pelo teste X. Assim, para cada nós da árvore de decisão, o C4.5 escolhe o teste X que maximize o *ganho*(*X*).

$$ganho(X) = entropia(T) - entropia_X(T) \tag{4.3}$$

Mais detalhes sobre este conceito podem ser obtidos em QUINLAN (1993).

4.2.2. Testes de Desempenho

O objetivo destes primeiros testes é verificar o comportamento do C4.5 em função de cada um dos conjuntos de dados de entrada.

Cada coluna da Tabela 4.2 indica um conjunto de dados de entrada e as linhas contêm a especificidade (Equação 3.3), a confiabilidade (Equação 3.4), a quantidade de consumidores fraudulentos classificados corretamente como fraudulentos ($Q_{f \to f}$) e a quantidade de consumidores normais classificados erroneamente como fraudulentos ($Q_{n \to f}$).

Em relação aos testes em que os índices de desempenho foram obtidos utilizando os mesmos dados de treinamento referentes às inspeções de 2006, a taxa de acerto não variou muito em função do tipo de entrada, ficando em torno de 90%. Porém, a especificidade (67,5%) quando se utilizou o conjunto Genérico foi muito superior em relação aos outros conjuntos de entrada, enquanto que a confiabilidade do conjunto Genérico (83,0%) também foi superior.

Analisando as quantidades de inspeções, a união das altas taxas de confiabilidade e de especificidade do conjunto de entrada Genérico leva a detectar 8.727 (33.893 menos 25.166) mais fraudes e errando 124 (7.057 menos 6.933) suspeitos a menos em relação ao conjunto Série.

Tabela 4.2: Métricas extraídas de diferentes conjuntos de entrada. Treinado e testado com inspeções de 2006.

Conjunto de entrada	Série	Características	Genérico
Tx. Acerto	91,2	89,1	93,6
Especif.	50,1	31,6	67,5
Confiabilidade	78,1	75,3	83,0
$Q_{f \to f}$	25.166	15.846	33.893
$Q_{n \to f}$	7.057	5.192	6.933

Em comparação com o conjunto Características, a utilização do conjunto Genérico faz dobrar o número de fraudes detectadas (33.893 versus 15.846) e com um aumento de 1.741 inspeções sem sucesso.

Fazendo a mesma análise com os dados da Tabela 4.3, em que os índices de desempenho foram gerados utilizando as inspeções do ano de 2007, a superioridade do conjunto Genérico é ainda maior, conseguindo uma especificidade de quase 58% e uma confiabilidade de 73%, enquanto as outras entradas possuem uma especificidade em torno de 13% e confiabilidade menor do que 36%. Isto resulta em uma quantidade menor de inspeções sem sucesso (4.506 inspeções erradas versus 7.110 e 5.298) e mais do que o quádruplo do número de fraudes detectadas (12.432 versus 2.864 e 2.884).

A utilização do conjunto Série mostrou melhor desempenho do que o conjunto Características quando os dados foram testados com os mesmos dados de entrada, porém um desempenho inferior quando comparada com os testes utilizando os dados do ano de 2007. Isto pode sugerir que a utilização do conjunto Série pode ter levado a um sobre-ajuste das regras, conforme explicado na subseção 2.3.3.

Tabela 4.3: Métricas extraídas de diferentes conjuntos de entrada. Treinamento: inspeções de 2006, testes: inspeções de 2007.

Conjunto de entrada	Série	Características	Genérico	Todos os Conjuntos
Tx. Acerto	86,1	87,1	92,7	92,1
Especif.	13,3	13,4	57,9	52,5
Conf.	28,7	35,2	73,4	71,9
No.Acertos	2.864	2.884	12.432	11.263
No.Erros	7.110	5.298	4.506	4.399

Quando foram utilizados os três conjuntos de entrada simultaneamente (Série, Características e Genérico) para teste, os resultados foram um pouco inferiores aos relativos à utilização isolada do conjunto Genérico. Ou seja, o conjunto Série mais o conjunto Características tiveram pouca importância em relação ao conjunto Genérico. Para comprovar essa hipótese, foram verificadas as regras geradas e observou-se que os atributos do conjunto Série só foram utilizados no final da árvore de decisão. Isto é um indicativo de um possível

sobre-ajuste que fez com que o desempenho do algoritmo caísse em relação ao conjunto Genérico isolado.

A Tabela 4.4 mostra os resultados obtidos por COMETTI & VAREJÃO (2005) com dados da ESCELSA utilizando características extraídas da série de Fourier das séries de consumo. A Tabela 4.4 não pode ser comparada com os resultados apresentados nas Tabelas 4.2 e 4.3, pois os dados de entrada são diferentes.

Tabela 4.4: Especificidade e confiabilidade obtidas com dados da ESCELSA.

Índices	ESCELSA
Especificidade	15
Confiabilidade	51

Entretanto, o que é mais relevante é que os testes que foram feitos com o conjunto Genérico superaram bastante os testes que foram feitos com os conjuntos derivados das séries de consumo.

Assim, pode-se concluir que, para o algoritmo C4.5, o conjunto Genérico apresentou um desempenho superior. Talvez isto não seja verdadeiro para os demais tipos de ferramenta de classificação, o que é verificado nos próximos testes.

4.2.3. Análise de Sensibilidade

O objetivo desta subseção é analisar a necessidade de conhecer quais os atributos relevantes em função da classe de saída e a importância da quantidade de amostras de dados de entrada.

a) Análise de sensibilidade em relação aos atributos de entrada das classes Anomalia e Normal.

Para verificar a importância dos atributos de entrada, o C4.5 foi executado utilizando para treinamento dados provenientes das classes Anomalia e Normal (a classe Fraude não é utilizada nesta subseção). Como a classe Anomalia está associada aos medidores que apresentam defeitos, é essencial utilizar ao menos um atributo que carregue alguma informação sobre as características intrínsecas do medidor. Esse atributo é o "prefixo do medidor", que está relacionado com o fabricante, modelo e lote de fabricação do medidor. Assim, foi utilizado o conjunto Genérico acrescentando-se ou não este novo atributo nos testes realizados nesta subseção.

É relevante observar que muitas vezes o defeito do medidor não depende apenas do projeto e da sua fabricação, mas também das condições de operação a que ele é submetido, como sobre-tensão na rede elétrica e a existência de insetos e água dentro da caixa do medidor. Além disso, cada modelo de medidor opera de maneira diferente quando submetidos a condições inadequadas de operação.

Nos testes relatados nas Tabelas de 4.5 a 4.9, o C4.5 foi treinado com os dados provenientes das inspeções do ano de 2006 e testados com os dados das inspeções de 2007.

Tabela 4.5: Métricas extraídas dos testes com o atributo "prefixo do medidor" presente e ausente.

	Atributo "prefixo do medidor"	
	Ausente	**Presente**
Tx. Acerto	84,2	85,0
Especif.	0,8	8,3
Conf.	34,8	67,2
No.Acertos	254	2.507
No.Erros	475	1.226

A diferença de desempenho entre a ausência e a presença do atributo "prefixo do medidor" foi grande, pois a especificidade sobe de 0,8% para 8,3% e a confiabilidade de 34,8% para 67,2% ao se inserir o referido atributo. Este resultado era previsível, pois a classe Anomalia tem forte dependência das características do medidor. Assim sendo, sua distinção é prejudicada na ausência de atributos relacionados ao medidor.

Foram feitos também testes usando as três classes de saída (Fraude, Anomalia e Normal) com a presença e a ausência do atributo "prefixo do medidor" com as inspeções relativas ao período já pré-definido.

A taxa de acerto original, mostrada nas Tabelas 4.6 e 4.8, é calculada por 1 menos a taxa de erro (a quantidade de clientes normais dividida pelo número de inspeções)

$$taxa\_de\_acerto\_original = 1 - \frac{\sum normais}{\sum inspeções} \qquad (4.4)$$

Pela Tabela 4.6, analisando os índices de especificidade e confiabilidade separadamente, observa-se que eles apresentam resultados semelhantes aos testes feitos para as classes Fraude/Normal e Anomalia/Normal, cuja especificidade e confiabilidade foram, respectivamente, 57,9% e 73,4% (conforme a Tabela 4.3) para as classes Fraude/Normal e 0,8% e 34,8% para as classes Anomalia/Normal (ver Tabela 4.5).

Tabela 4.6: Matriz de confusão dos testes com o atributo "prefixo do medidor" ausente.

Atributo "prefixo do medidor" **Ausente**						
	Classe Predita			**Índices**		**%**
Classe Real	Normal	Fraude	Anomalia	Taxa de Acerto		79,5
Normal	158.160	4.391	594		Especif.	57,7
Fraude	8.959	12.378	133	**Fraude**	Conf.	68,2
Anomalia	28.659	1.373	228		Especif.	0,8
Tx. Acerto Original (%)		24,1		**Anomalia**	Conf.	23,9

No entanto, se não considerarmos a existência das três classes e sumarizarmos os dados referentes às classes Fraude (F) e Anomalia (A) conforme a Tabela 4.7, observaremos que a especificidade é 27,3%. Isto mostra a importância de entender bem o problema e conhecer quais as classes existentes e os atributos relevantes para cada classe. Caso contrário, os dados de uma classe desconhecida podem atrapalhar a classificação da outra classe.

Tabela 4.7: Métricas agregadas extraídas para testes com o atributo "prefixo do medidor" ausente.

Atributo "prefixo do medidor" **Ausente**				
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	F+A	Tx. Acerto	80,2
Normal	158.160	4.985	Especif	27,3
F+A	37.618	14.112	Conf.	73,9

A Tabela 4.8 relata os testes com os mesmos dados de entrada do teste anterior, mas inserindo o atributo "prefixo do medidor". Observe novamente que a especificidade e confiabilidade utilizando dados das três classes de saída foram muito próximos dos índices obtidos dos testes que utilizaram o C4.5 para separar apenas suas classes Fraude/Normal (especificidade = 57,9% e confiabilidade = 73,4%, ver Tabela 4.3) e Anomalia/Normal (especificidade = 8,3% e confiabilidade = 67,2%, ver Tabela 4.5). Isto mostra que se pode utilizar o C4.5 para separar as três classes simultaneamente sem perda de qualidade.

Tabela 4.8: Matriz de confusão dos testes com o atributo "prefixo do medidor" presente.

Atributo "prefixo do medidor" **Presente**						
	Classe Predita			**Índices**		**%**
Classe Real	Normal	Fraude	Anomalia	Taxa de Acerto		80,4
Normal	157.175	4.811	1.159		Especif.	60,9
Fraude	8.165	13.084	221	**Fraudes**	Conf.	67,6
Anomalia	26.379	1.473	2.408		Especif.	8,0
Tx. Acerto Original (%)		24,1		**Anomalia**	Conf.	63,6

Observe pela Tabela 4.9 que, mesmo com a presença do atributo "prefixo do medidor", a especificidade fica comprometida se não separarmos as classes do problema.

Tabela 4.9: Métricas agregadas extraídas dos testes com o atributo "prefixo do medidor" presente.

Atributo "prefixo do medidor" **Presente**				
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	F+A	Tx. Acerto	81,1
Normal	157.175	5.970	Especif	33,2
F+A	34.544	17.186	Conf.	74,2

O teste mostrado pela Tabela 4.2 com o conjunto Genérico foi repetido sem o atributo "quantidade de carga declarada", o que provocou uma grande queda na especificidade e na confiabilidade, como mostra a Tabela 4.10. Este novo teste também comprova a importância de determinados atributos de entrada para distinguir a classe.

Tabela 4.10: Métricas agregadas extraídas de testes com o atributo "quantidade de carga declarada" ausente e presente.

Atributo "quantidade de carga declarada"		
	Presente	**Ausente**
Tx. Acerto	92,7	88,9
Especificidade	57,9	17,4
Confiabilidade	73,4	58,2

b) Análise de sensibilidade para a quantidade de dados de entrada

O objetivo desta subseção é entender qual a importância da quantidade de dados de entrada no desempenho do classificador.

Conforme mostra a Tabela 4.11, foram executados 9 testes variando o período de treinamento, indicado em azul na tabela, e de obtenção dos índices de desempenho (taxa de acerto, especificidade e confiabilidade), hachurado na tabela. O treinamento dos testes abrangeu três períodos sobrepostos (de setembro/2006 a dezembro/2006, de maio/2006 a dezembro/2006 e de janeiro/2006 a dezembro/2006) e a obtenção dos índices de desempenho abrangeu outros três períodos (setembro/2006 a dezembro/2006, apenas o mês de janeiro/2007 e de janeiro/2007 a outubro/2007).

Tabela 4.11: Períodos de treinamento e testes

Teste	ano 2006			ano 2007	
	jan-abril	maio-ago	set-dez	Jan	fevereiro-outubro
I			Obtenção de índices		
II		Treinamento	Obtenção de índices		
III	Treinamento	Treinamento	Obtenção de índices		
IV			Treinamento	Obtenção de índices	
V		Treinamento	Treinamento	Obtenção de índices	
VI	Treinamento	Treinamento	Treinamento	Obtenção de índices	
VII			Treinamento	Obtenção de índices	Obtenção de índices
VIII		Treinamento	Treinamento	Obtenção de índices	Obtenção de índices
IX	Treinamento	Treinamento	Treinamento	Obtenção de índices	Obtenção de índices

Observe que no teste I, o treinamento ocorreu com os dados de setembro/2006 a dezembro/2006 e os índices foram obtidos utilizando os mesmos dados de treinamento. No testes II e III, o treinamento ocorreu, respectivamente, com os dados de maio/2006 a dezembro/2006 e com os dados de janeiro/2006 a dezembro/2006 e foram utilizados os dados de setembro/2006 a dezembro/2006 para os cálculos dos índices de desempenho. Ou seja, nos testes I, II e III, os dados utilizados na obtenção dos índices de desempenho já estavam contidos nos dados de treinamento.

Nos testes IV, V e VI, os índices de desempenho foram calculados sobre os dados de janeiro de 2007 e os dados de treinamento foram, respectivamente, de setembro/2006 a dezembro/2006, de maio/2006 a dezembro/2006 e de janeiro/2006 a dezembro/2006.

Os dados de treinamento do teste VII foram os mesmos do teste IV, do teste VIII foram os mesmo do teste V e do teste IX foram os mesmos do teste VI. Porém, os índices de desempenho para os testes VII, VIII e IX foram obtidos utilizando os dados de janeiro/2007 a outubro/2007.

As taxas de acerto mantiveram-se praticamente constante nos testes relatados nas Tabelas 4.12 e 4.13 em que a quantidade de dados de treinamento foi variada. No entanto, a especificidade aumentou e a confiabilidade diminuiu à medida que aumentava a quantidade de

dados de treinamento. É interessante observar que este fato também foi observado nos testes I, II e III, onde os dados utilizados para os cálculos da especificidade e confiabilidade já estavam presentes durante o treinamento, conforme é mostrado na Tabela 4.14.

Para saber a razão disso, é necessário entender com maior grau de detalhes o C4.5. Porém, uma possível especulação a respeito é que aumentando a quantidade de dados de treinamento é natural que a quantidade de padrões repetidos também aumenta, fazendo com que um determinado padrão deixe de ser considerado espúrio e passe a ser representado por uma regra específica.

Uma outra observação que se extrai da comparação entre as Tabelas 4.12 e 4.13 é que a especificidade não se altera muito quando os testes foram realizados com os dados de janeiro/2007 e de janeiro/2007 a outubro/2007. Isto é uma indicação de que um único mês pode exprimir uma boa representação do período.

Tabela 4.12: Testes com variação da quantidade de inspeções durante o treinamento e testados com as inspeções de janeiro de 2007.

Conjunto Genérico				
Treinamento	2006	set-dez	maio-dez	jan-dez
Teste	janeiro de 2007			
Taxa de Acerto		92,6	92,6	92,8
Especificidade		47,9	54,4	57,6
Confiabilidade		80,7	75,6	75,6

Tabela 4.13: Testes com variação da quantidade de inspeções durante o treinamento e testados com as inspeções de janeiro/2007 a outubro / 2007.

Conjunto Genérico				
Treinamento	2006	set-dez	maio-dez	jan-dez
Teste	jan./2007 a out./2007			
Taxa de Acerto		92,5	92,6	92,7
Especificidade		49,8	55,3	57,9
Confiabilidade		78,1	74,6	73,4

Tabela 4.14: Testes com variação da quantidade de inspeções durante o treinamento e teste com parte (set./2006 a dez./2006) das inspeções do treinamento.

Conjunto Genérico				
Treinamento	2006	set-dez	maio-dez	jan-dez
Teste	set./2006 a dez./2006			
Taxa de Acerto		95,5	95,3	95,1
Especificidade		55,1	62,9	64,0
Confiabilidade		83,0	75,1	72,5

4.3. Redes Neurais Artificiais (RNAs)

As Redes Neurais Artificiais (RNAs) são reconhecidamente competentes em tarefas de classificação, particularmente quando operam em espaços contínuos. No entanto, os conjuntos de dados de entrada têm atributos tanto no espaço contínuo como no discreto (atributos categóricos). Assim, o objetivo deste capítulo é verificar como as Redes Neurais Artificiais se comportam com os diferentes tipos de conjuntos de dados de entrada.

4.3.1. Introdução

As RNAs têm sido motivadas desde o seu surgimento pela capacidade de reconhecimento do cérebro humano, a qual é totalmente diferente dos computadores digitais convencionais. O cérebro é altamente complexo, não-linear, e processa informações em paralelo. Ele tem a capacidade de organizar suas estruturas, conhecidas como neurônios, para que desempenhem certas tarefas muitas vezes mais rápidas do que o mais rápido computador digital existente atualmente. O sistema de visão humano, por exemplo, faz a representação do ambiente ao redor e fornece informações para as pessoas interagirem com ele. Essas tarefas o cérebro realiza em aproximadamente 100 – 200 ms, sendo que o computador convencional leva dias para executar tarefas de muito menor complexidade (HAYKIN, 1999).

A origem da neurocomputação é geralmente atribuída a MCCULLOCH & PITTS, cujo artigo de 1943 descreve um modelo lógico de redes neurais que unificava seus estudos de

neurofisiologia e lógica matemática e que era capaz, em princípio, de computar qualquer função computável. Atualmente, as redes neurais estão presentes em diversas áreas de aplicação, como: controle, reconhecimento de padrões, robótica e mineração de dados (HAYKIN, 1999).

Um modelo de rede neural é apresentado na Figura 4.2, a qual é formada por um conjunto de neurônios interligados, sendo que cada neurônio tem como função de transferência uma soma ponderada das entradas e executa uma função não-linear sobre a saída (HAYKIN, 1999).

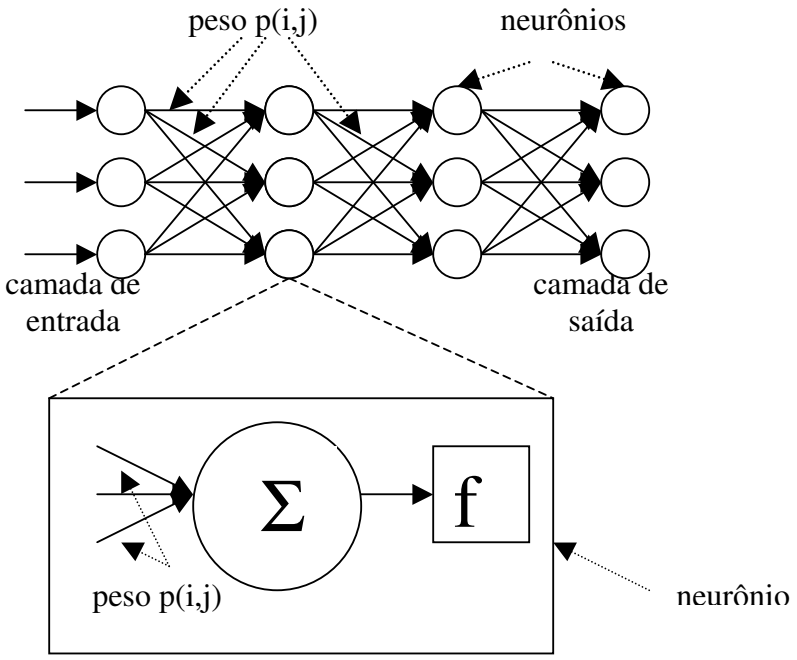

Figura 4.2: Modelo de uma rede neural artificial

Essa configuração relativamente simples proporciona algumas propriedades importantes como:

- arquitetura altamente distribuída e processamento paralelo;
- capacidade de aprendizado;
- poder de generalização (as redes neurais podem operar bem junto a dados não apresentadas durante o treinamento).

Alguns aspectos importantes de uma RNA são: definir a topologia da rede, a quantidade de camadas, o número de neurônios de cada camada, a função de ativação do neurônio e o algoritmo de treinamento.

Em RNAs, existem três fases: treinamento, validação e teste. A fase de treinamento é quando ocorre o aprendizado. Esta é uma fase iterativa cujo objetivo é ajustar os pesos em cada conexão para que a rede responda corretamente à saída desejada. O objetivo da fase de validação é dizer quando o treinamento deve ser parado para evitar que ocorra sobre-ajuste. O tempo de treinamento e a sua precisão dependem da topologia da rede, dos dados disponíveis para treinamento e do algoritmo de treinamento utilizado. Na fase teste são medidos os índices de desempenho da rede, como taxa de acerto, especificidade e confiabilidade.

O tipo de aprendizado utilizado neste trabalho é o supervisionado. Para a sua utilização, é necessária que a rede seja provida com os dados de entrada e suas respectivas saídas desejadas para que os pesos das conexões sejam ajustados.

Outro tipo de aprendizado é o não-supervisionado. Ele não é utilizado neste trabalho, mas é mencionado como ele pode ser utilizado em trabalhos futuros. Neste caso, não se sabe a saída e espera-se que a rede neural capture alguma propriedade estatística presente nos dados de entrada.

4.3.2. Testes de Desempenho

A plataforma de testes utilizada para implementar a ferramenta de classificação baseada em uma RNA foi o Matlab 6.5, o qual proporciona muitas facilidades tanto para as fases de desenvolvimento como depuração e manutenção de código. Em contrapartida, por se tratar de um interpretador de código fonte, a execução do software é lenta e tem-se pouco controle do uso de memória. O tempo de processamento não foi um limitante para os estudos realizados. No entanto, foi necessário trabalhar com acesso a arquivos para contornar as limitações de memória.

Os dados de saída da RNA foram especificados como "−1" para a classe Fraude e "+1" para a classe Normal. Assim, quanto mais próximo o valor da variável de saída estiver de "−1"

e “+1”, mais confiável é a decisão. Diante disso, foi criada uma faixa de indecisão entre −0,25 e +0,25, onde os resultados não seriam classificados.

A Figura 4.3 mostra o modelo da RNA utilizada nos testes. Embora não esteja indicado na figura, todos os neurônios recebem uma entrada de valor constante e igual a 1, denominada de entrada de polarização. Ela é composta por 30 neurônios na camada intermediária e um neurônio na camada de saída, todos os neurônios usam a função tangente hiperbólica e o treinamento foi feito utilizando um algoritmo de otimização de segunda ordem baseado no gradiente conjugado estendido (VON ZUBEN, 1996; HAYKIN, 1999), com 1.000 iterações e taxa de ajuste de 25%. Optou-se por um número elevado de neurônios na camada intermediária (30 neurônios) em virtude de se empregar um conjunto de validação para evitar sobre-treinamento.

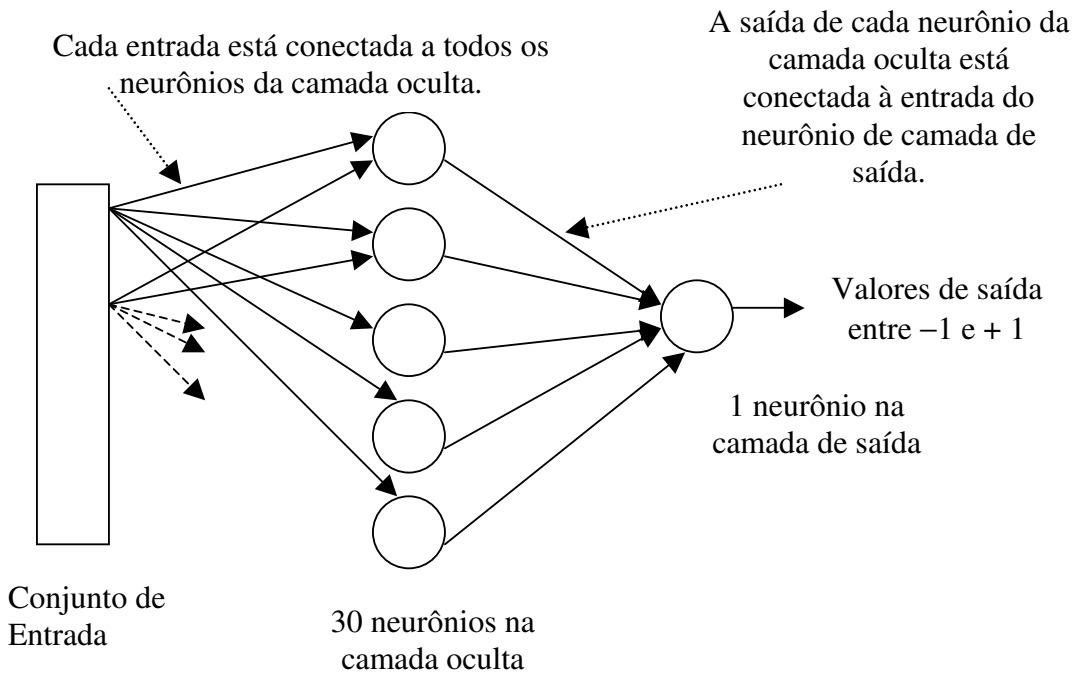

Figura 4.3: Modelo da RNA utilizada nos testes.

Foram realizados testes com e sem a faixa de indecisão e o campo “Indecisão” da tabela indica o número de inspeções não classificadas porque estão na faixa de indecisão. O valor desse campo é zero quando a faixa de indecisão não foi utilizada. Sem a faixa de indecisão,

quando a saída da RNA é maior do que zero a classe é Normal. Logo, a classe é Fraude quando a saída da RNA é menor ou igual a zero.

a) Dados de Entrada: Série

Utilizando-se as séries de consumo como vetor de entrada, a RNA foi treinada com os dados referentes às inspeções de 2006. A Tabela 4.15 mostra a matriz de confusão baseada nos mesmos dados de treinamento, enquanto que a Tabela 4.16 mostra a matriz de confusão baseada nos dados de inspeção de 2007. Embora a taxa de acerto tenha sido pouco alterada, os índices de especificidade e confiabilidade caíram bastante. Isto pode ser um efeito de sobre-ajuste, ou então o padrão dos dados de 2007 é diferente dos de 2006 quando se usa o conjunto Série como entrada.

A taxa (tx.) de acerto original mostrada nas Tabelas 4.15 e 4.16 é dada pela Equação 4.4.

Tabela 4.15: Matriz de confusão com os mesmos dados de treinamento e teste

Treinam.	2006		**Teste**	2006
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	89,8
Normal	123.135	2.089	Especif	25,6
Fraude	12.327	4.238	Conf.	67,0
Tx. Acerto Original		11,7		

Tabela 4.16: Matriz de confusão com dados de treinamento diferentes dos dados de teste

Treinam.	2006		**Teste**	2007
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	85,3
Normal	124.600	1.572	Especif	1,7
Fraude	20.002	342	Conf.	17,9
Tx. Acerto Original		13,9		

Na Tabela 4.17, os testes foram feitos com os mesmos dados de treinamento e teste da Tabela 4.16, mas com a faixa de incerteza habilitada, isto é, foram desconsideradas as amostras cujas saídas da RNA estavam entre $-0{,}25$ e $+0{,}25$. Com isso, a especificidade caiu, como era

esperado. Porém, a confiabilidade também caiu, possivelmente porque uma parte das amostras que estava dentro da faixa de incerteza tinha sido rotulada corretamente, mesmo estando na faixa de incerteza.

Tabela 4.17: Matriz de confusão com dados de treinamento diferentes dos dados de teste

Treinam.	2006		**Teste**	2007
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	86,0
Normal	122.026	564	Especif	0,6
Fraude	19.331	117	Conf.	17,2
Incerteza	4.478			
Tx. Acerto Original		13,9		

Baseando-se na Tabela 4.16, pode-se verificar que o desempenho da RNA tendo como entrada o conjunto Série não foi satisfatório, pois, das 20.344 fraudes existentes, foram selecionadas apenas 342, além de 1.572 erroneamente. Além disso, a confiabilidade de 17,9% é próxima da taxa de acerto original da AES Eletropaulo de 13,9%.

b) Dados de Entrada: Características

Fazendo os mesmos testes que foram feitos para o conjunto Série, porém usando o conjunto Características, observa-se, comparando as Tabelas 4.15 e 4.18, em que o treinamento e a teste foram feitos com os mesmos dados, que os resultados pioraram, principalmente em relação ao índice de especificidade.

Tabela 4.18: Matriz de confusão com dados de treinamento iguais aos de teste.

Treinam.	2006		**Teste**	2006
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	88,9
Normal	123.801	1.423	Especif	13,3
Fraude	14.369	2.196	Conf.	60,7
Tx. Acerto Original		11,7		

No entanto, ao testar com os dados de 2007, houve uma melhora significativa com a especificidade subindo de 1,7% para 7,3% e a confiabilidade de 17,9% para 49,8% (compare as Tabelas 4.16 e 4.19). Mas, mesmo assim, a especificidade de 7,3% é muito baixa, pois apenas 1.487 fraudes foram detectadas das 20.344 fraudes existentes no conjunto.

Tabela 4.19: Matriz de confusão com dados de treinamento diferentes dos dados de teste.

Treinam.	2006		**Teste**	2007
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	86,1
Normal	124.676	1.496	Especif	7,3
Fraude	18.857	1.487	Conf.	49,8
Tx. Acerto Original		13,9		

A Tabela 4.20 mostra aquilo que era esperado: aumento da confiabilidade com a diminuição da especificidade, devido à presença da faixa de incerteza.

Tabela 4.20: Matriz de confusão com dados de treinamento diferentes dos dados de teste e com faixa de incerteza.

Treinam.	2006		**Teste**	2007
Classe Real	**Classe Predita**		**Índices**	**%**
	Normal	Fraude	Tx. Acerto	88,6
Normal	120.128	241	Especif	1,7
Fraude	15.323	336	Conf.	58,2
Incerteza	10.488			
Tx. Acerto Original		13,9		

c) Dados de Entrada: conjunto Genérico

O conjunto Genérico é composto tanto de atributos categóricos como contínuos. O tratamento dos valores contínuos já é bem conhecido. A seguir, são discutidas duas alternativas para o tratamento dos valores categóricos para os dados deste trabalho.

A primeira alternativa é simplesmente mapear cada valor categórico para um valor numérico, como mostra a Tabela 4.21, a qual usa a variável categórica “região” como

exemplo. Assim, esse código serve de entrada para a RNA e é possível com uma única entrada representar as 6 regiões.

Tabela 4.21: Mapeamento do atributo categórico região em códigos numéricos.

região	código
Oeste	1
SP Sul	2
Anhembi	3
Centro	4
Leste	5
ABC	6

Como a RNA utilizada tem a propriedade de ser um aproximador universal (HAYKIN, 1999), ela seria capaz de trabalhar com esse tipo de entrada. No entanto, para facilitar o trabalho da RNA, diminuindo assim o número de iterações e de neurônios necessários, costuma-se normalizar a entrada usando o desvio padrão.

Porém, a normalização não faz muito sentindo para esse exemplo, pois a distribuição estatística das variáveis não seria normal.

Uma representação alternativa seria representar cada valor de cada variável categórica por uma variável binária. Assim, os 6 diferentes valores para esta variável (Oeste, Sul, Anhembi, Centro, Leste, ABC) seriam representados em 6 variáveis binárias distintas (uma para o valor Oeste, outra para o valor Sul, ...) como entrada da RNA. Com esse tipo de representação, se a região é Centro, as variáveis de entrada Oeste, Sul, Anhembi, Leste e ABC teriam valores iguais a 0, enquanto somente a variável Centro teria valor de entrada igual a 1. Isto significa que esse tipo de representação leva à criação de matrizes com bastante atributos, porém esparsas.

A Tabela 4.22 mostra os resultados dos testes com o conjunto Genérico, sendo que os atributos categóricos foram convertidos tanto para binários (a) como para uma simples representação em código numérico (b), sendo que o treinamento foi feito com os dados das inspeções de 2006.

Tabela 4.22: Métricas dos testes feitos utilizando o conjunto Genérico e convertendo os atributos categóricos para binários (a) e representação em código numérico (b).

Índices	Entradas - Representação Binária	
	2006	2007
Tx. Acerto	92,9	69,4
Especif.	52,8	39,2
Conf.	79,6	19,7

a)

Índices	Entradas - Representação Numérica	
	2006	2007
Tx. Acerto	91,9	70,6
Especif.	43,0	74,3
Conf.	78,1	28,6

b)

Observa-se que quando os índices de desempenho foram obtidos utilizando os mesmos dados de treinamento (Tabela 4.22 a), os testes com os atributos categóricos convertidos em atributos binários apresentaram melhores resultados (mas, não tão superiores). No entanto, quando os índices de desempenho foram extraídos com os dados do ano de 2007 (diferentes dos de treinamento), o desempenho da RNA que utilizou o mapeamento para valores numéricos foi muito melhor. Portanto, notou-se que a capacidade de generalização da RNA caiu com a transformação para atributos binários.

Os mesmos testes foram aplicados com a utilização dos três conjuntos de entrada simultaneamente e convertendo os atributos categóricos para binários (Tabela 4.23 a) e representado-os no formato numérico (Tabela 4.23 b). Os resultados foram similares aos obtidos com os testes que utilizaram apenas o conjunto Genérico como entrada: a conversão para binários obteve melhores resultados quando os índices foram extraídos dos dados de treinamento e o simples mapeamento em números conseguiu um melhor desempenho (mas, não tão grande como nos testes da Tabela 4.22) para cálculo dos índices com dados diferentes dos dados de treinamento. Assim, para os demais testes com atributos categóricos, é utilizado o simples mapeamento em números.

Tabela 4.23: Métricas dos testes feitos utilizando os 3 conjuntos de entrada e convertendo os atributos categóricos para binários (a) e representação em código numérico (b).

Índices	Entradas - Representação Binária	
	2006	2007
Tx. Acerto	94,5	82,0
Especif.	63,0	43,0
Conf.	86,6	37,2

a)

Índices	Entradas – Representação Numérica	
	2006	2007
Tx. Acerto	93,4	85,2
Especif.	54,4	46,1
Conf.	83,6	46,6

b)

O próximo teste foi realizado habilitando a faixa de confiança (Tabela 4.24). Como era esperado, houve aumento da confiabilidade e queda na especificidade.

Tabela 4.24: Métricas dos testes feitos utilizando o conjunto Genérico com os atributos categóricos mapeados em valores numéricos com os dados de treinamento referentes ao ano de 2006 e de teste aos anos de 2006, de 2007 sem faixa de confiança e de 2007 com faixa de confiança.

Índices	2006	2007 sem faixa de confiança	2007 com faixa de confiança
Tx. Acerto	91,9	70,6	81,1
Especif.	43,0	74,3	41,1
Conf.	78,1	28,6	42,1
Incertezas			70012

Foram feitos testes para checar o comportamento da especificidade e confiabilidade em função da quantidade de iterações (Figura 4.4). A entrada foi o conjunto Genérico com os atributos mapeados em valores numéricos e tanto o treinamento como o teste foram feitos com dados de 2006, sem faixa de confiança. Foi observado que a especificidade aumenta e a confiabilidade cai. Isto ocorre porque o objetivo da RNA é aumentar a taxa de acerto (que pela Equação 3.1 significa classificar corretamente tanto os consumidores normais como os fraudulentos) com o aumento do número de iterações, isto faz com aumente o número de fraudes detectadas, pagando o preço de classificar erroneamente normais como fraudadores. O importante é que mais consumidores fraudulentos sejam classificados corretamente do que consumidores normais sejam classificados erroneamente. Isto faz com que a confiabilidade

caia, pois mais consumidores normais são classificados fraudulentos e a especificidade aumenta porque mais consumidores fraudulentos são classificados corretamente.

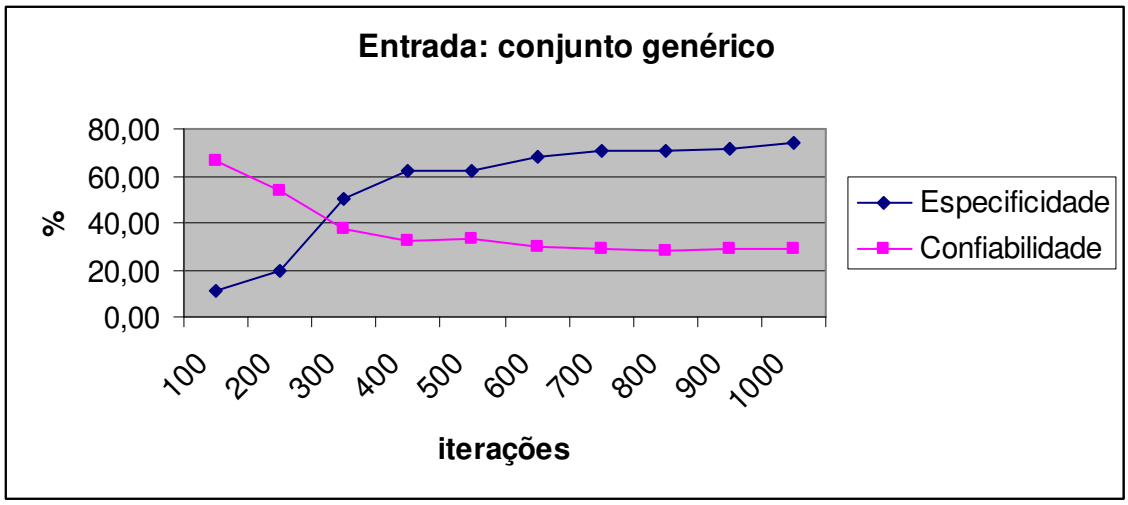

Figura 4.4: Variação da especificidade e confiabilidade em função do número de iterações.

d) Dados de Entrada: Série, Características e Genérico.

Foram realizados testes utilizando os três conjuntos de entrada simultaneamente, sendo que os atributos categóricos foram simplesmente mapeados para numéricos. O treinamento foi feito com os dados de 2006 e os índices foram extraídos utilizando os dados de 2006 e 2007 (com e sem faixa de incerteza). Parte desses testes já foi executada anteriormente e os resultados foram apresentados na Tabela 4.24. Os resultados apresentados na Tabela 4.25 são similares aos da Tabela 4.24: queda da especificidade e confiabilidade quando a RNA é testada com dados diferentes dos de treinamento. Repare que a inserção da faixa de incertezas aumenta muito a confiabilidade, porém provoca uma queda na especificidade.

Tabela 4.25: Métricas dos testes feitos utilizando os três conjuntos de entrada simultaneamente. Os testes foram executados com dados de 2006 e 2007 (com e sem faixa de incerteza).

Índices	2006	2007 sem faixa de confiança	2007 com faixa de confiança
Tx. Acerto	93,4	85,2	91,3
Especif.	54,4	46,1	26,3
Conf.	83,6	46,6	56,9
Incertezas			37.067

Comparando os resultados da Tabela 4.25 com os da Tabela 4.24, em que a RNA foi treinada apenas com o conjunto Genérico, percebe-se uma melhoria nos índices extraídos a partir dos dados de treinamento. Porém, devido ao princípio de dominância de Pareto (Steuer, 1986), não se pode afirmar que seu desempenho melhorou quando a RNA foi testada com os dados do ano de 2007. A existência de mais entradas aumenta a probabilidade de distinguir duas ou mais amostras. No entanto, se as entradas a mais não contiverem informação útil para distinguir a classe, elas acabam provocando um sobre-ajuste da rede, o que explica a melhoria da rede quando os dados de testes foram os mesmos de treinamento.

4.4. Support Vector Machine (SVM)

4.4.1. Introdução

"Support Vector Machine" (SVM) é uma técnica de aprendizado de máquina fundamentada na teoria do aprendizado estatístico e que vem sendo desenvolvida ao longo das últimas três décadas (VAPNIK & CHERVONENKIS, 1974; VAPNIK, 1982; VAPNIK, 1995).

Conforme é mostrado na Figura 4.5 (a qual está em duas dimensões apenas para efeito de visualização), o uso de SVM como classificador consiste em aplicar um operador não-linear no espaço de entrada, cujos dados não são separados linearmente, para mapeá-los para um espaço de dimensão maior, onde seja possível separá-los por hiperplanos lineares. O

hiperplano ótimo é aquele que maximiza a margem entre as duas classes e tende a ser aquele que melhor generaliza quando se retorna o mapeamento para o espaço original dos dados.

Os vetores-suporte são as amostras que estão mais próximas do hiperplano e, assim, definem completamente a sua equação. A complexidade do mapeamento depende do número de vetores-suporte e não da dimensão do espaço de entrada. Além disso, encontrar o hiperplano ótimo envolve um problema de otimização convexa, com solução única.

Um dos desafios no uso de SVM é a escolha da função de kernel que será usada na definição do operador não-linear que faz o mapeamento para o espaço de maior dimensão, também denominado espaço de características (LIMA, 2004).

Foram realizados testes para verificar o comportamento do SVM em relação aos dados relativos às inspeções da AES Eletropaulo.

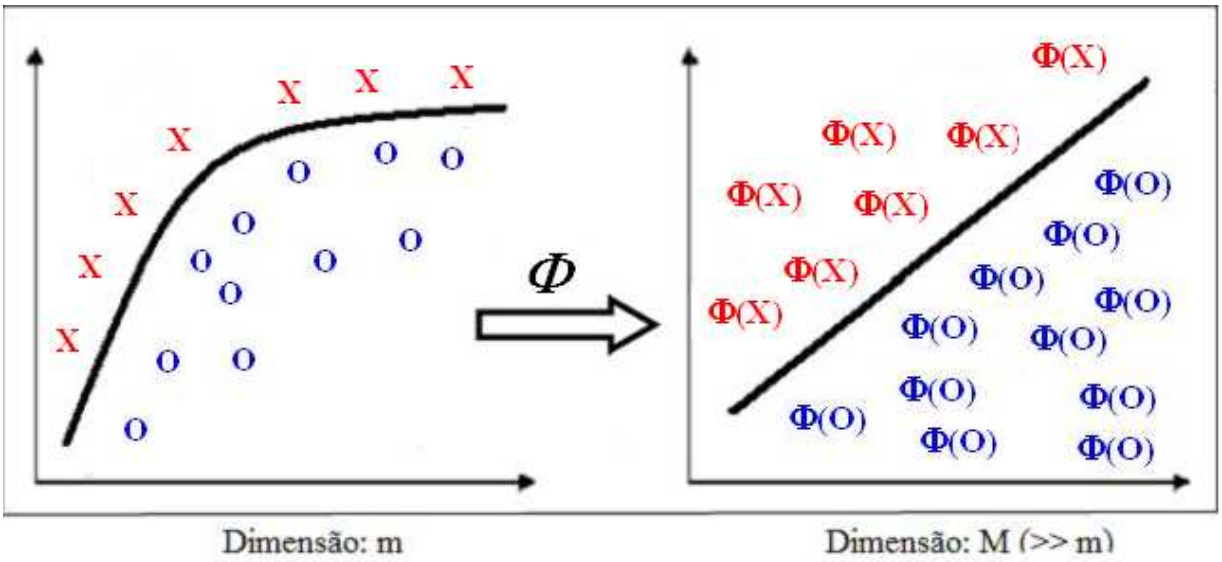

Figura 4.5: Exemplo de mapeamento para o espaço de características, onde é possível a separação linear das duas classes, originalmente não separáveis, por um hiperplano. Foi utilizado M=2 por questões de visualização, pois M >> m em aplicações práticas.

4.4.2. Testes de Desempenho

Os testes executados com SVM são apresentados na Tabela 4.26. Os testes com o conjunto Série apresentaram o pior desempenho em relação às três métricas utilizadas, os testes com o conjunto Características apresentaram uma melhor especificidade (32,8%) e os

testes com o conjunto Genérico apresentaram uma confiabilidade melhor. Os testes utilizando os três conjuntos simultaneamente obtiveram, para os três índices, um resultando intermediário em relação aos conjuntos Características e Genérico.

O SVM tem muitos parâmetros para serem configurados, sendo a função kernel o principal parâmetro. Para os testes, foi utilizada uma função kernel linear. Assim sendo, é possível conseguir um melhor desempenho alterando os parâmetros.

Tabela 4.26: Execução do SVM utilizando diferentes conjuntos

Conjunto de entrada	Série	Características	Genérico	Todos Conjuntos
Tx. Acerto	80,7	81,2	85,6	85,1
Especif.	11,7	32,8	26,0	29,9
Conf.	18,9	32,6	46,7	44,3
No.Acertos	2.387	6.681	5.284	6.078
No.Erros	10.270	13.831	6.026	7.633

4.5. Naive Bayes

4.5.1. Introdução

"Naive Bayesian Classifier" (DUDA & HART, 1973; LANGLEY et al., 1992), também chamado de "Simple Bayes Classifier" (DOMINGOS & PAZZANI, 1997) baseia-se no Teorema de Bayes que, dada uma partição $\{C_1, C_2, ..., C_n\}$ do espaço amostral Ω e um evento qualquer A, expressa a probabilidade condicional $P(C_i \mid A)$ dado que ocorreu o evento A (BUSSAB & MORETTIN, 1987), conforme a Equação 4.5.

$$P(C_i \mid A) = \frac{P(C_i) * P(A \mid C_i)}{\sum_{j=1}^{n} P(C_j) * P(A \mid C_j)}, j = 1,...,n \qquad (4.5)$$

Esse classificador é chamado de *naive* (ou ingênuo, em português) porque ele supõe que os atributos são independentes, isto é, $C_i \cap C_j = \phi$ e $P(A_i, A_j) = P(A_i) * P(A_j)$.

O algoritmo desenvolvido calcula a relevância do atributo A para a classe C segundo a Equação 4.6, onde A^+ significa a presença do atributo e A^- a sua ausência.

$$\text{Relevância}(C \mid A) = \frac{P(C \mid A^+)}{P(C \mid A^-)} \qquad (4.6)$$

Assim, dado um conjunto de atributos do cliente, se a Relevância da classe fraude for λ vezes maior do que a Relevância da classe normal, o cliente é classificado como fraudulento; caso contrário, como normal. Fazendo o valor de λ maior do que um, faz com que o algoritmo seja mais restritivo em classificar um cliente como fraudulento. Com isso, a confiabilidade aumenta.

Como já foi mencionado, esse algoritmo pressupõe que os atributos sejam independentes. Caso contrário, os atributos dependentes terão um peso maior no resultado, provocando uma distorção na análise. Na próxima subseção, são apresentados os resultados para os testes com quatro conjuntos de entrada.

4.5.2. Teste de Desempenho

O algoritmo Naive Bayes foi testado com quatro conjuntos de entrada e λ igual a 1 e os resultados estão na Tabela 4.27. O conjunto Características obteve a melhor especificidade (61,7%) e a pior confiabilidade, enquanto que o conjunto Genérico obteve a melhor confiabilidade (28,9%). Os testes com todos os conjuntos superaram o conjunto Série tanto em relação a especificidade (54,7% × 49,2%) como na confiabilidade (21,7% × 21,5%). Porém a taxa acerto do conjunto Série foi um pouco superior (67,9% × 66,3%). Quando foram utilizados todos os conjuntos simultaneamente como entrada do algoritmo, obteve-se um desempenho intermediário entre as entradas dos conjuntos Genérico e Características, comparando tanto a taxa de acerto como a especificidade e a confiabilidade. Diante disso, não se pode dizer que um conjunto se sobressaiu em relação ao outro, pois se o que se busca é a melhor confiabilidade, escolhe-se o conjunto Genérico, se for a melhor especificidade, utiliza-

se o conjunto Características, se for um valor intermediário de confiabilidade e especificidade, deve-se optar por todos os conjuntos simultaneamente.

Tabela 4.27: Execução do Naive Bayes utilizando diferentes conjuntos de entrada

Conjunto de entrada	Série	Características	Genérico	Todos Conjuntos
Tx. Acerto	67,9	51,4	75,6	66,3
Especif.	49,2	61,7	52,2	54,7
Conf.	21,5	16,5	28,9	21,7
No.Acertos	10.009	12.558	10.629	11.127
No.Erros	36.640	63.351	26.089	40.155

Na Tabela 4.28, há uma comparação entre os resultados obtidos por esse trabalho em relação aos de COMETTI & VAREJÃO (2005) para o Naive Bayes. Os resultados obtidos foram similares, pois a especificidade obtida com os dados da ESCELSA foi 58% e com os da AES Eletropaulo foram 49% para o conjunto Série e 62% para o conjunto Características, e as confiabilidades obtidas foram de 20% para a ESCELSA e 21% (conjuntos Série) e 17% (conjunto Características) para AES Eletropaulo. Ou seja, não se pode dizer que os resultados da ESCELSA foram melhores ou piores do que os da AES Eletropaulo, eles apresentaram a especificidade e a confiabilidade intermediárias em relação aos conjuntos Série e Características. Na comparação não foi utilizado o conjunto Genérico, porque COMETTI & VAREJÃO (2005) trabalharam somente com séries temporais e suas características.

Tabela 4.28: Comparação entre as especificidades e as confiabilidades obtidas com o Naive Bayes com os dados da AES Eletropaulo e ESCELSA.

	ESCELSA	AES Eletropaulo	
Índices		**Série**	**Características**
Especif.	58	49	62
Conf.	20	21	17

4.6. Testes de Desempenho sem a utilização de atributos de consumo

Nos testes anteriores, sempre foi utilizado algum atributo da série de consumo, pois mesmo o conjunto Genérico possui os atributos "carga declarada", "consumo médio" e "último consumo". Diante disso, foram feitos testes sem esses atributos de consumo e os resultados são apresentados na Tabela 4.29.

Comparando esses resultados com os da Tabela 4.30, onde os testes foram feitos com a inserção dos atributos "carga declarada", "consumo médio" e "último consumo", observou-se que a especificidade caiu bastante para o C4.5 na ausência dos atributos de consumo.

No caso do C4.5, comparando as Tabelas 4.30 e 4.31, observa-se que o desempenho dele com os atributos de consumo foi superior mesmo quando o teste ocorreu com os mesmos dados de treinamento (Tabela 4.31), porém sem os valores de consumo.

Também foi observado que o SVM teve o mesmo desempenho com ou sem os atributos de consumo. No caso da RNA, o seu desempenho foi inferior sem os atributos de consumo, porém não tão ruim como o C4.5. Já para o Naive Bayes, a comparação torna-se mais difícil porque quando há aumento da especificidade este está associado a uma queda da confiabilidade.

Tabela 4.29: Métricas dos algoritmos tendo como conjunto de entrada atributos sem valor e sem características de consumo. Dados de teste diferentes dos de treinamento.

Atributos **sem** valores de consumo				
Treino	2006		Teste	2007
Índices	C4.5	Naive B	SVM	RNA
Tx. Acerto	86,6	85,0	85,6	79,2
Especif.	7,7	26,4	26,7	29,9
Conf.	66,2	43,3	46,6	24,9

Tabela 4.30: Métricas dos algoritmos tendo como conjunto de entrada atributos com valores e características de consumo.

Atributos + **Carga Declar + Média/Último Consumo**				
Treino	2006		Teste	2007
Índices	C4.5	Naive B	SVM	RNA
Tx. Acerto	91,4	75,6	85,6	81,1
Especif.	56,0	52,2	26,0	41,1
Conf.	75,8	28,9	46,7	42,1

Tabela 4.31: Métricas dos algoritmos tendo como conjunto de entrada atributos sem valor e sem características de consumo. Dados de teste iguais aos de treinamento.

Atributos **sem** valores de consumo				
Treino	2006		Teste	2006
Índices	C4.5	Naive B	SVM	RNA
Tx. Acerto	91,5	87,8	88,2	95,8
Especif.	14,8	28,1	30,1	50,5
Conf.	72,3	32,2	34,7	87,1

4.7. Comparação entre as ferramentas de classificação e os conjuntos de entrada

Para a comparação das ferramentas de classificação, será usado o princípio de dominância de Pareto, pois envolve duas medidas diferentes: especificidade e confiabilidade.

Pelo princípio de Pareto (Steuer, 1986), uma solução domina a outra se a solução dominante tem ao menos um índice superior e nenhum inferior aos índices da outra solução.

Solução dominada é aquela que tem ao menos um índice inferior a outra solução e nenhum índice superior. Solução não-dominada é aquela que não é dominada por nenhuma outra solução. Porém, ela não é necessariamente uma solução dominante, ou seja, ela não necessariamente domina alguma outra solução.

A aplicação do princípio de Pareto pode ser feita tanto pelo gráfico da Figura 4.6 como pela Tabela 4.32. No entanto, através de um gráfico a visualização é mais simples, pois, no caso deste trabalho, como se deseja uma solução com alta especificidade e confiabilidade, basta procurar o ponto mais alto e mais à direita. É claro que pode ser que o ponto mais alto não seja o mais à direita. Neste caso não existe um único ponto dominante. Porém, se no gráfico for observada a existência de um ponto mais alto e mais à direita em relação ao ponto sob análise, este último certamente é um ponto dominado.

A Tabela 4.32 mostra a especificidade e a confiabilidade em função da ferramenta de classificação e do conjunto de entrada. Nessa tabela, consta também a quantidade de inspeções (No. Insp.) selecionadas pelo critério e a quantidade de fraudes detectadas (Fraude Detect.) que são mostradas na tabela apenas para ajudar a interpretação dos resultados, pois elas são obtidas da especificidade e da confiabilidade.

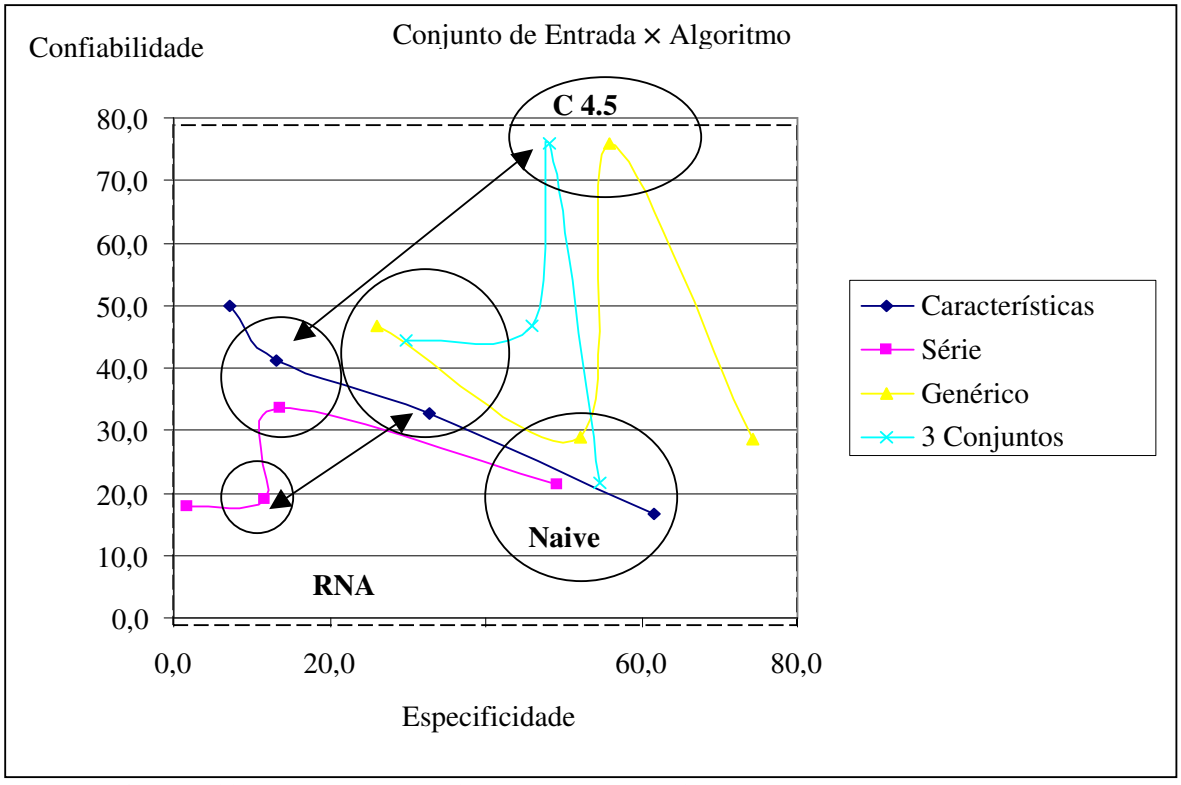

Figura 4.6: Especificidade versus confiabilidade para vários conjuntos de entrada e algoritmos (ferramentas de classificação).

As células coloridas da Tabela 4.32 indicam os algoritmos dominantes locais, isto é, quando a comparação é feita apenas para o mesmo conjunto de entrada, enquanto as posições em negrito e em vermelho indicam os dominantes globais.

A análise local é feita olhando a tabela verticalmente, ou seja, para um mesmo conjunto de entrada. Fazendo isto, é possível detectar qual, ou quais, os algoritmos que são dominantes para um mesmo conjunto de entrada.

Tabela 4.32: Métricas dos testes feitos variando tanto algoritmo como as entradas.

Algoritmo	Conjunto de Entrada / Indices	Série	Características	Genérico	Três Conjuntos
RNA	No. Insp	1.914	2.983	52.916	20.112
	Fraude Detect.	342	1.487	15.112	9.369
	Especificidade	1,7	7,3	**74,3**	46,1
	Confiabilidade	17,9	49,8	**28,6**	46,6
Naive Bayes	No. Insp	46.649	75.909	36.718	51.282
	Fraude Detect.	10.009	12.558	10.629	11.127
	Especificidade	49,2	61,7	52,2	54,7
	Confiabilidade	21,5	16,5	28,9	21,7
SVM	No. Insp	12.657	20.512	11.310	13.711
	Fraude Detect.	2.387	6.681	5.284	6.078
	Especificidade	11,7	32,8	26,0	29,9
	Confiabilidade	18,9	32,6	46,7	44,3
C4.5	No. Insp	8.216	6.501	15.031	12.891
	Fraude Detect.	2.756	2.681	11.397	9.792
	Especificidade	13,5	13,2	**56,0**	48,1
	Confiabilidade	33,5	41,2	**75,8**	76,0

Para o conjunto Série de consumo, os algoritmos dominantes são Naive Bayes e C4.5, pois tanto a especificidade como a confiabilidade de ambos são maiores que a especificidade e confiabilidade da RNA e do SVM. No entanto, o Naive Bayes não domina o C4.5 porque a confiabilidade do C4.5 é maior do que a do Naive Bayes. O C4.5 também não domina o Naive Bayes porque a especificidade do Naive Bayes é maior do que a do C4.5.

Em relação ao conjunto Características, não houve um algoritmo que dominasse os demais, pois, como mostra o gráfico, quanto mais à direita, mais baixo está o ponto. Ou seja, todos os pontos são não-dominados.

Analisando o conjunto Genérico, observa-se no gráfico que o C4.5 domina os algoritmos Naive Bayes e SVM, porém ele não domina a RNA, que por sua vez também não domina o C4.5, pois a RNA está posicionada na parte mais baixa do gráfico. Ou seja, neste caso há dois algoritmos não-dominados: C4.5 e RNA.

Quando se analisa a entrada formada pelos 3 conjuntos, tem-se que os algoritmos dominantes são C4.5 (o qual está mais alto e mais à direita do que a RNA e o SVM) e o Naive Bayes que está mais à direita (e abaixo) em relação ao C4.5.

Para checar o critério dominante global, deve-se analisar apenas as células coloridas da tabela (os não-dominados locais), ou procurar os pontos mais à direita e mais alto no gráfico. Assim, a entrada genérica junto com a RNA e junto com o C4.5 são as soluções dominantes globais.

Portanto, o melhor conjunto de entrada em relação aos dados do problema é o Genérico, pois é com esse conjunto que se conseguiu os melhores resultados. Além disso, utilizando os dados do C4.5 relativo ao conjunto Genérico, temos que a especificidade (56,0%) e a confiabilidade (75,8%) são bem superiores aos respectivos índices dos conjuntos Série e Características.

Em relação à ferramenta de classificação, a melhor é o C4.5, pois ele é o único em que todas as soluções locais são não-dominadas, além de ter uma solução dominante global.

O Naive Bayes é a ferramenta mais simples utilizada nos testes. Assim, pode parecer estranho ele não ser dominado em 75% dos testes. Porém, o que ocorreu é que ele não foi dominado porque sua especificidade foi sempre muito alta. No entanto, a sua confiabilidade é baixa, ao redor de 21% para os casos em que ele não foi dominado.

Em relação ao SVM, é necessário especificar melhor seus parâmetros para se conhecer a sua potencialidade. Logo, mais testes são necessários para saber se sua aplicação é útil para os dados da AES Eletropaulo.

A RNA teve uma solução não-dominada global porque a sua especificidade é muito alta para o conjunto Genérico, com uma confiabilidade (28,6%) não tão alta. No entanto, a RNA merece atenção devido à subseção 4.6 do Capítulo 4, onde testes foram apresentados com e sem atributos de consumo e em que o C4.5 não conseguiu bons resultados.

CAPÍTULO 5

5.1. Considerações Finais

O objetivo desta dissertação foi melhorar o índice de acertos das inspeções realizadas pela AES Eletropaulo, trabalhando com os dados disponibilizados pela empresa. Desde as primeiras reuniões, a empresa afirmou que um aumento de 2% da taxa de acerto já representaria um alto retorno financeiro para a empresa e para a sociedade. Além disso, a empresa também não possuía nenhuma ferramenta automatizada para gerar as inspeções.

A AES Eletropaulo disponibilizou dois diferentes tipos de dados reais, séries temporais de consumo elétrico e atributos específicos de cada cliente, como localização e tipo de atividade. A partir deles, foram gerados quatro conjuntos de dados de entrada:

- séries temporais de consumo elétrico;
- características extraídas dessas séries temporais;
- seleção de atributos, misturando tanto dados de séries de consumo, como características e atributos específicos de cada cliente;
- combinação dos três conjuntos anteriores.

Em cima desses quatro conjuntos, foram desenvolvidas quatro ferramentas de classificação de naturezas distintas (C4.5, SVM, RNA e Classificador Naive Bayes) com o objetivo de determinar qual ou quais destas combinações conseguiriam aumentar o número de acertos das inspeções realizadas pela AES Eletropaulo.

A conclusão a que se chegou é que uma ferramenta possui uma confiabilidade maior, enquanto outra tem uma especificidade maior e outra tem valores intermediários dessas

métricas. Assim, a utilização de todas as ferramentas é justificável, dependendo da intenção da empresa no momento da sua utilização. Isto é, se a empresa deseja realizar poucas inspeções, escolhem-se as ferramentas com maior confiabilidade. Se a intenção é inspecionar mais clientes, escolhem-se as ferramentas com maior especificidade.

Como durante a realização deste trabalho foi percebido que o principal objetivo da empresa é aumentar a receita e não necessariamente aumentar a taxa de acerto das inspeções, é de extrema utilidade ter ferramentas que tenham uma especificidade maior, mesmo com uma confiabilidade menor. Isto é verdade enquanto a soma do retorno financeiro sobre as fraudes descobertas é maior do que o custo total das inspeções.

Testes alternativos foram feitos para observar o desempenho das ferramentas de classificação utilizando variações dos padrões de entrada, seja em volume, em quantidade de atributos ou no formato da representação apresentada ao algoritmo.

5.2.Trabalhos Futuros

O próximo passo é validar esses resultados em campo. No entanto, não se pode afirmar se será possível conseguir bons resultados, pois, embora tenha se trabalhado com toda a base de dados real da AES Eletropaulo com mais de 6,2 milhões de consumidores e 1,2 milhão de inspeções realizadas em campo, essas inspeções têm um viés, pois foram selecionadas de acordo com vários critérios com o objetivo de dar o maior retorno financeiro para a empresa. Assim sendo, não se sabe se as regras determinadas pelas amostras são válidas para todo o conjunto. Mas, mesmo assim, as ferramentas podem ser utilizadas para filtrar as inspeções geradas pelo processo corrente para aumentar a taxa de acerto.

Também como trabalho futuro, novos tipos de dados deverão ser inseridos nos modelos (como dados georeferenciáveis, dados censitários do IBGE e dados técnicos relativos à manutenção dos medidores) com o intuito de melhorar a classificação, principalmente de anomalias.

Técnicas de agrupamento, como mapas auto-organizáveis (HAYKIN, 1999), poderão também ser utilizadas para segmentar os clientes da AES Eletropaulo com o intuito de verificar

se todos os grupos estão sendo inspecionados de forma proporcional, para que se possa tomar ações corretivas caso isso não aconteça.

É possível utilizar também um comitê de máquinas para combinar as 4 ferramentas descritas na classificação dos dados. Segundo HAYKIN (1999), comitês de máquinas visam fundir o conhecimento adquirido por cada uma das ferramentas especialistas para se chegar a uma decisão coletiva que é supostamente superior a de cada ferramenta tomada individualmente. Assim, o resultado final pode ser uma combinação ponderada da saída de cada classificador, sendo que o peso de cada classificador depende do quão bem o classificador tem se comportado para aquela região específica do espaço de dados da amostra.

Uma alternativa que também melhora o desempenho das ferramentas é utilizar a seleção de característica (BLUM & LANGLEY, 1997, KOHAVI & JOHN, 1997), que seleciona os atributos mais importantes para o bom desempenho dos classificadores, em vez de disponibilizar todos os atributos. Nada impede também que a seleção de variáveis seja utilizada junto com um comitê de máquinas.

Enfim, se as ferramentas discutidas neste trabalho apresentarem, em campo, apenas 10% do desempenho conseguidos nos testes feitos, a empresa terá um alto retorno financeiro. Isto é plausível, pois os testes foram feitos com dados reais. Além disso, esta dissertação propõe, como trabalhos futuros, alternativas para melhorar ainda mais o desempenho dos classificadores.

Referências Bibliográficas

BLUM, A., LANGLEY, P., "Selection of Relevant Features and Examples in Machine Learing", Artificial Intelligence, vol. 97, pp. 245-271, 1997.

BUSSAB, W., MORETTIN, P. A., "Estatística Básica", Atual Editora Ltda., 1987.

CARVALHO, A.M.B.R., CHIOSSI, T.C.S., "Introdução à Engenharia de Software", Editora da Unicamp, ISBN 85-268-0532-0, 2001.

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 de Consumidores", Relatório Final do Projeto de P&D, Ciclos 2003/2004, ESCELSA/ANEEL, 2005.

DOMINGOS, P. & PAZZANI, M., "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", Machine Learning, Vol. 29, , Números 2-3, pp. 103-130, 1997.

DUDA, R. & HART, P., "Pattern Classification and Scene Analysis", New York, NY Wiley, 1973.

FRANCISCO, E. R., "Relação entre o Consumo de Energia Elétrica, a Renda e a Caracterização Econômica de Famílias de Baixa Renda do Município de São Paulo", Dissertação (mestrado) - Escola de Administração de Empresas de São Paulo, Orientador: Prof. Dr. Francisco Aranha, 2006.

FRANCISCO, E.R., FAGUNDES, B.E., "Geostatiscal Study For Fraud and Energy Losses", Proceedings of ESRI International User Conference, San Diego, CA, Junho/2007

HAYKIN, S., "Neural Networks: a Comprehensive Foundation", Prentice Hall, ISBN 0-02-352761-7, 1999.

KLÖSGEN, W. & ZYTKOW J.M., "Handbook of Data Mining and Knowledge Discovery", Oxford University Press, 2002.

KOHAVI, R. & JOHN, G., "Wrappers For Feature Selection", Artificial Intelligence, Vol. 97, Números 1-2, pp. 273-324, 1997.

JIANG, R., TAGARIS H., LACHSZ A, JEFFREY M., "Wavelet Based Feature Extraction and Multiple Classifiers For Electricity Fraud Detection". IEEE Transmission and Distribution Conference and Exhibition 2002: Asia Pacific, IEEE/PES, Vol. 3, pp. 6-10, 2002.

LANGLEY, P., IBA, W. & THOMPSON, K., "An Analysis of Bayesian Classifiers." Proceedings of the Tenth National Conference on Artificial Intelligence, Menlo Park, CA, AAAI Press e MIT Press, pp. 223-228, 1992.

LIMA, C.A.M., "Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores-Suporte", Tese (Doutorado) - Universidade Estadual de Campinas, Orientador: Prof. Dr. Fernando José Von Zuben, 2004.

MCCULLOCH, W.S. & PITTS, W., "A Logical Calculus of the Ideas Immanent in Nervous Activity", Bulletin of Mathematical Biophysics, Vol. 5, pp. 115-133, 1943.

PROVOST, F. & KOHAVI, R., "Machine Learning", Springer Netherlands, Vol. 30, Números 2-3, pp. 127-274, 1998.

QUINLAN, J.R., "Discovering Rules by Induction from Large Collections of Examples", In D. Michie (ed.), Expert Systems in the Micro Electronic Age. Edinburgh, UK, Edinburgh University Press, 1979.

QUINLAN, J.R., "C4.5 Programs for Machine Learning", San Mateo, CA, Morgan Kaufmann Publishers, ISBN 1-55860-238-0, 1993.

STEUER, R. E., "Multiple Criteria Optimization: Theory, Computation and Application", ISBN 0-471-88846-X, 1986.

TRETYAKOV K., "Machine Learning Techniques in Spam Filtering", Data Mining Problem-oriented Seminar, MTAT.03.177, pp. 60- 79, 2004.

VAPNIK, V.N. & CHERVONENKIS, A.Y., "Theory of Pattern Recognition". Nauka, Moskow, Russian, 1974.

VAPNIK, V.N., "ESTIMATION OF DEPENDENCES BASED ON EMPIRICAL DATA". SPRINGER-VERLAG, BERLIM, 1982.

VAPNIK V.N., "THE NATURE OF STATISTICAL LEARNING THEORY", SPRINGER-VERLAG, 1995.

VON ZUBEN, F.J., "MODELOS PARAMÉTRICOS E NÃO-PARAMÉTRICOS DE REDES NEURAIS ARTIFICIAIS E APLICAÇÕES", TESE (DOUTORADO) - UNIVERSIDADE ESTADUAL DE CAMPINAS, ORIENTADOR: MÁRCIO LUIZ DE ANDRADE NETTO, 1996.