DETECÇÃO DE FRAUDE OU ERRO DE MEDIÇÃO EM GRANDES CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO ROUGH SETS BASEADO EM DADOS HISTÓRICOS E EM DADOS EM TEMPO REAL

CRISTIAN MARA MAZZINI MEDEIROS PATRÍCIO

CAMPO GRANDE

2005

UNIVERSIDADE FEDERAL DO MATO GROSSO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

DETECÇÃO DE FRAUDE OU ERRO DE MEDIÇÃO EM GRANDES CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO ROUGH SETS BASEADO EM DADOS HISTÓRICOS E EM DADOS EM TEMPO REAL

Dissertação submetida à
Universidade Federal de Mato Grosso do Sul
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

CRISTIAN MARA MAZZINI MEDEIROS PATRÍCIO

Campo Grande, Julho de 2005.

DETECÇÃO DE FRAUDE OU ERRO DE MEDIÇÃO EM GRANDES CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO ROUGH SETS BASEADO EM DADOS HISTÓRICOS E EM DADOS EM TEMPO REAL

Cristian Mara Mazzini Medeiros Patrício

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Energia Elétrica*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campo Grande.'

_____

João Onofre Pereira Pinto, Ph.D.
Orientador

_____

Kathya Silvia Collazos Linares, Dra.
Co-Orientadora

_____

João Onofre Pereira Pinto, Ph.D.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

_____

João Onofre Pereira Pinto, Ph.D.
Presidente

_____

José Demisio Simões da Silva, Dr.

_____

Celso Correia de Souza, Dr.

_____

Evandro Manzina, Dr.

Ao meu esposo Engenheiro Paulo Patrício da Silva
À minha filha Fernanda Mazzini Patrício
Aos meus pais,
Lídio Souza Medeiros e Arize Mazzini Medeiros

Agradecimentos

A Deus, porque até aqui me ajudou. Tenho sempre em meu coração a certeza de que “Tudo posso Naquele que me fortalece (Filipenses 4.13)”, e a conclusão de mais esta etapa em minha vida é prova fiel de que Deus é justo e honra os seus filhos.

Ao Professor. João Onofre Pereira Pinto, pela forma cordial com que me acolheu no Programa de Mestrado do Departamento de Engenharia Elétrica da UFMS. Por sua capacidade de realização, criatividade, sugestões. Atribui-me a honra de sua orientação neste mestrado. Incansavelmente compreendeu todas as minhas dificuldades impostas pelo cotidiano e grandiosamente apoiou-me quando mais precisei.

Á Professora Kathya Silvia Collazos Linares, pelo carinho com que se dispôs a ajudar-me no desenvolvimento deste trabalho.

Ao Professor Celso Correia de Souza, amigo que sempre me incentivou e ajudou-me a caminhar adiante em busca de meus objetivos. Com muita sabedoria, aconselhou-me inúmera vez em horas de difíceis tomadas de decisões.

Aos colegas e amigos de mestrado José Edison Cabral Júnior e Luigi Galotto Junior pela disposição de ajudar sempre.

Ao meu esposo Engenheiro Paulo Patrício da Silva. Acompanhou-me mais de perto neste desafio. Ajudou-me a superar obstáculos que pareciam ser intransponíveis. Demonstrou-me infinitamente mais o seu amor.

A minha filha Fernanda Mazzini Patrício, pela alegria, carinho, amor e compreensão, mesmo sendo tão pequenina.

Aos meus pais, pelos joelhos dobrados a Deus, suplicando força para eu vencer. Estiveram sempre de prontidão para ajudar-me.

Resumo da Dissertação apresentada à UFMS como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

DETECÇÃO DE FRAUDE OU ERRO DE MEDIÇÃO EM GRANDES CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO ROUGH SETS BASEADO EM DADOS HISTÓRICOS E EM DADOS EM TEMPO REAL

Cristian Mara Mazzini Medeiros Patrício

Julho / 2005

Orientador: João Onofre Pereira Pinto, Ph.D.
Área de Concentração: Energia Elétrica.
Palavras-chave: Fraude ou erro de Medição, Energia Elétrica, Rough Sets.
Número de Páginas: 120

O objetivo geral deste trabalho é apresentar uma metodologia que define perfis de comportamentos diários de unidades consumidoras de energia elétrica ligadas em alta tensão, com a finalidade de detectar fraudes ou erros de medição. A partir desta metodologia construiu-se um sistema baseado em regras, utilizando informações estáticas dos consumidores, ou seja, informações que não variam no tempo e, dados dinâmicos obtidos em tempo real. O desenvolvimento deste sistema seguiu os princípios gerais de descoberta do conhecimento a partir de bancos de dados, utilizando na mineração das informações a teoria de Rough Sets para seleção de atributos relevantes e geração de regras. Classificou-se o comportamento das unidades, através dos perfis diários ou semanais, em normais e anormais. Os clientes classificados como anormais são selecionados para inspeção técnica. Os resultados são considerados satisfatórios, uma vez que a taxa de acerto na identificação de fraude obtida pelo sistema, utilizando-se análise semanal na unidade consumidora, a partir da pré-seleção dos consumidores com suspeita de fraude foi de 64,70%. Pode-se considerar, portanto, que a metodologia desenvolvida mostrou-se capaz de ajudar a solucionar problemas das perdas comerciais relacionadas à fraude ou erro de medição das concessionárias de energia elétrica.

Abstract of Dissertation presented to UFMS as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

DETECTION OF FRAUD OR MEASUREMENT ERROR IN BIG ELECTRIC POWER´S CONSUMERS USING ROUGH SETS BASED IN HISTORICAL DATA AND IN DATA BASED IN REAL TIME

Cristian Mara Mazzini Medeiros Patrício

July / 2005

Advisor: João Onofre Pereira Pinto, Ph.D.
Area of Concentration: Electrical Energy
Key-word: Fraud or Measurement Error, Electric Power, Rough Sets.
Page number: 120

The general purpose of this work is to present a methodology that defines daily behavior profiles of consumers unities of electric energy, fed through high voltage distribution lines, with the purpose of detecting frauds or measurement errors. This methodology allowed the development of a rule based system that uses consumer's static information, i.e., information that does not vary in time and dynamic data, obtained in real time. The development of this system follows the general principles of knowledge discovery in data bases, using the theory of Rough Sets for mining of information, selection of relevant attributes and generation of rules. The system classifies the behavior of the unities, through daily or weekly profiles, in normal or abnormal. The consumer classified as abnormal undergo in-site inspection. The results were considered satisfactory since the rate of correct identification was 64.70%. Therefore, the developed methodology is shown to be capable aid to resolve the problems of commercial losses related to fraud or error of measurement from the utility companies.

Sumário

Abreviaturas / Siglas

A	Sistema de informação
ANEEL	Agência Nacional de Energia Elétrica
ATR	Conjunto de atributos
B	Subconjunto de atributos pertencentes a C
B^*	Conjunto de atributos mínimos – Reduto
$\underline{B}(X)$	Aproximação Inferior de X em relação a B
$\overline{B}(X)$	Aproximação Superior de X em relação a B
C	Conjunto de atributos condicionais
Cl	Classes formadas por um conjunto de atributos condicionais
CM	Constante do medidor
Cons	Consumo registrado na memória de massa
Cons_Int	Consumo integralizado
CT	Classe tarifária
D	Conjunto de atributos de decisão
DCFP	Demanda contratada no período Fora da Ponta
DCP	Demanda contratada no período da Ponta
Dem	Demanda registrada na memória de massa
DS	Dia da semana analisado
ELETROBRÁS	Centrais Elétricas do Brasil S/A
$F_A(B)$	Função de discernibilidade em relação a B
Fat_Car	Fator de carga
$IND_A(B)$	Relação de equivalência em relação a B

DCBD	Descoberta de Conhecimento em Base de Dados
kV	quilovolt
kVar	quilovolt-ampère-reativo
kVarh	quilovolt-ampère-reativo-hora
kW	quilowatts
kWh	quilowatts-hora
$M_D(B)$	Matriz de discernibilidade em relação a B
Max_Dem	Máxima demanda registrada na memória de massa
Med_Dem	Média das demanda registrada na memória de massa
MME	Ministério de Minas e Energia
MT	Modalidade tarifária
NM	Número do medidor
NR	Número do registro da informação de Memória de Massa
NRI	Número de registros iguais
R	Sistema de informação
RDC	Resumo Diário Completo
RDFP	Resumo Diário Fora Ponta
RDP	Resumo Diário Ponta
RI	Conjunto de classes com φ inferior
RR	Conjunto de classes com φ superior
R	Conjunto de regras extraído do banco de dados reduzido
R_δ	Conjunto de R acrescido do valor de δ
$RED_A(B)$	Conjunto de redutos em relação a B
$RR_{\min}$	Conjunto de Regras gerais minimizadas
$R_\delta valida$	Conjunto de regras válidas definidas por δ
$RF(X)$	Região de Fronteira de X

$RR_{validas}$	Regras gerais minimizadas válidas
$\Re$	Sistema de informação munido de D
SH	Segmentos horários Ponta e Fora Ponta
SR	Segmentos horários Indutivo e Capacitivo
U	Conjunto de objetos ou registros
$U/IND_A(B)$	Conjunto quociente de U pela relação $IND_A(B)$
X	Conjunto de objetos ou registros com respeito a B
$\overline{X}$	Média
δ	Consistência
δ^*	Consistência medida
σ	Desvio padrão
$\alpha_B(\underline{B}(X))$	Coeficiente da Qualidade da Aproximação Inferior de X com respeito a B
$\alpha_B(\overline{B}(X))$	Coeficiente da Qualidade da Aproximação Superior de X com respeito a B
$\alpha_B(X)$	Coeficiente de Imprecisão de X com respeito a B
φ	Suporte
Δ	Intervalo de variação

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Capítulo 1
Introdução

1.1. Contextualização

As fraudes ocorrem nas mais diversas áreas e segmentos, como por exemplo, em cartões de crédito, em telefonia celular, em seguradoras, no setor bancário, em energia elétrica, entre outras. A detecção de fraudes no Brasil e no mundo tem sido tema de pesquisa e desenvolvimento, sendo assim, uma área de grandes investimentos das empresas.

Muitos métodos foram propostos para os sistemas de detecção de fraudes, baseados principalmente em técnicas da inteligência artificial para extrair informações contundentes de bancos de dados, normalmente, existente nas empresas interessadas no processo. A seguir são relacionados alguns destes métodos.

Ghosh e Reilly [1], em um dos trabalhos publicados em 1994, desenvolveram um sistema para um banco americano, o *Mellon Bank*, utilizando redes neurais para detectar fraude na sua carteira de cartões de crédito. A rede utilizada foi do tipo base radial com três camadas e alimentação direta (*feed-forward*). Foi selecionada uma grande quantidade de transações com cartão de crédito, com todas as atividades observadas em um grupo de usuários de cartão, em um período de dois meses subseqüentes. Para o treinamento desta rede foi utilizado um conjunto de dados de transações de contas de cartões sem fraude e com transações fraudulentas, enquanto que para a simulação da rede nenhuma transação fraudulenta foi colocada inicialmente. A rede criada foi capaz de detectar um número significativamente grande de fraudes, reduzindo a fator de 20% os alarmes falsos que faziam os sistemas de detecção de fraude baseados em regras.

Para diagnosticar fraude em cartões de crédito, Brause et al [2] desenvolveram uma técnica baseada em estatística. Esta tarefa de detecção necessita de diagnóstico muito especial devido a uma proporção muito pequena de fraude nas transações, uma vez que somente uma transação financeira é fraudadora em mil (1:1000). Com isso, nenhum sucesso de predição menor que 99,9% é aceitável, segundo eles. Para tanto, apresentaram uma solução através de combinação probabilística e rede neural adaptativa para analisar banco de dados de transações de cartões de crédito. Levando-se em conta as características dos fraudadores, estes foram agrupados de acordo com quatro regras geradas através do método probabilístico de mineração de dados. Para cada classe formada através destes agrupamentos, desenvolveu-se uma rede neural do tipo *base-radial* especializada em suas características isoladamente. Estas redes foram agrupadas para formar uma decisão comum – se fraudador ou não. Esta técnica de mineração de dados combinando estatística e rede neural artificial proporcionou uma alta taxa de cobertura de fraude com baixo índice de alarmes falsos.

Estima-se que operadoras de telefonia celular podem perder mais de um milhão de dólares ao dia devido ao uso fraudulento de telefones celulares. Taniguchi et al [3], em um trabalho publicado em 1998 apresentaram três propostas de detecção de fraude para esta área. Primeiramente desenvolveram um método utilizando rede neural *feed-forward* de múltiplas camadas, usada para mapear uma função não-linear descriminativa entre duas classes: com fraude e sem fraude. Os dados utilizados para o treinamento e simulação desta rede foram retirados das faturas dos assinantes, uma vez que estas foram consideradas capazes de descrever o comportamento dos clientes. O método detectou mais de 85% de fraude sem causar alarmes falsos. Os dois últimos métodos propostos utilizaram-se de dados capazes de caracterizar o comportamento diário de cada assinante. Para modelar o segundo método, estimação de densidade de probabilidade, usou-se um modelo de mistura Gaussiana [4], apresentando um desempenho considerado aceitável, detectando mais de 70% de casos de fraudes sem causar alarmes falsos. Como não há regras determinísticas que permita identificar um assinante fraudulento, o último método apresentado na proposta destes autores foi de uma rede Bayesiana que apresenta uma estrutura para lidar com incertezas probabilísticas e criar um “aparato” adequado para resolver o problema de detecção de fraude. O desempenho deste último método foi similar ao segundo.

A fraude em telefonia móvel também foi alvo de estudos para Burge et al [5]. Em um dos tópicos destacados no trabalho desenvolvido em 1997, eles analisaram o perfil de usuários do sistema, investigando numerosas seqüências de faturas para identificar o comportamento do assinante. Propuseram detectar fraude baseando-se em regras e baseando-se em redes neurais, supervisionada e não-supervisionada. As redes utilizadas foram do tipo *multi-layer perceptron*, usando função sigmóide e com duas camadas ocultas. Utilizou-se o mesmo banco de dados para avaliar o desempenho destas propostas e os resultados foram considerados satisfatórios, porém não foi informado o percentual de acertos ou de falhas na detecção de fraudes na utilização de telefonia móvel.

O sistema elétrico, particularmente as unidades consumidoras de energia, também enfrenta o problema de fraude. De acordo com o estudo apresentado por Cabral et al [6,7], é possível identificar consumidores fraudadores de energia elétrica ligados em baixa tensão englobando as diversas classes de consumidores: pequeno, médio e grande porte quanto ao consumo de energia. Como é necessário que haja inspeções técnicas em campo, este trabalho ajuda a selecionar as unidades consumidoras a serem inspecionadas, de tal forma que o resultado, quanto à detecção de fraude, nesses consumidores passe a ser melhor, uma vez que, para encontrar um fraudador sem ter uma diretriz de busca, é um processo lento e oneroso para a empresa concessionária de energia. O trabalho propõe uma metodologia baseada em Rough Sets para detecção de fraudes nesse segmento utilizando dados históricos. A aplicação da metodologia identifica padrões de comportamento fraudulentos em bancos de dados de empresas de energia elétrica e, a partir destes padrões, derivam-se regras de classificação que, em futuros processos de inspeção, indicarão quais clientes apresentam perfis fraudulentos. Com inspeções guiadas por comportamentos suspeitos, aumenta-se a taxa de acerto e a quantidade de fraudes detectadas, diminuindo as perdas com fraudes nas empresas de distribuição de energia elétrica.

No trabalho de Reis et al [8] é apresentado um sistema de pré-seleção de consumidores de energia elétrica para inspeção, com o objetivo de detectar fraudes e erros de medição baseado em uma árvore de decisão CART [9]. A partir do banco de dados de uma empresa de distribuição de energia elétrica, foram selecionados 5 atributos (dentre os 52 disponíveis) e 40.000 clientes (de um total de 600.000). O

treinamento do sistema ocorreu com 50% dos clientes, selecionados aleatoriamente e, o teste foi realizado com os clientes remanescentes. O resultado alcançado pelo sistema foi de uma taxa de acerto de 40% para fraudadores. A taxa alcançada pela empresa em questão, utilizando seus próprios métodos, chega a ser até 35% menor que a taxa encontrada pelo sistema desenvolvido.

1.2. Fraude no Setor Elétrico

Os sinônimos de fraude são dolo, burla, engano, logro, contrabando. Pode-se definir a fraude no setor elétrico como subtração ilícita de energia sem o respectivo pagamento por parte de quem dela usufrui.

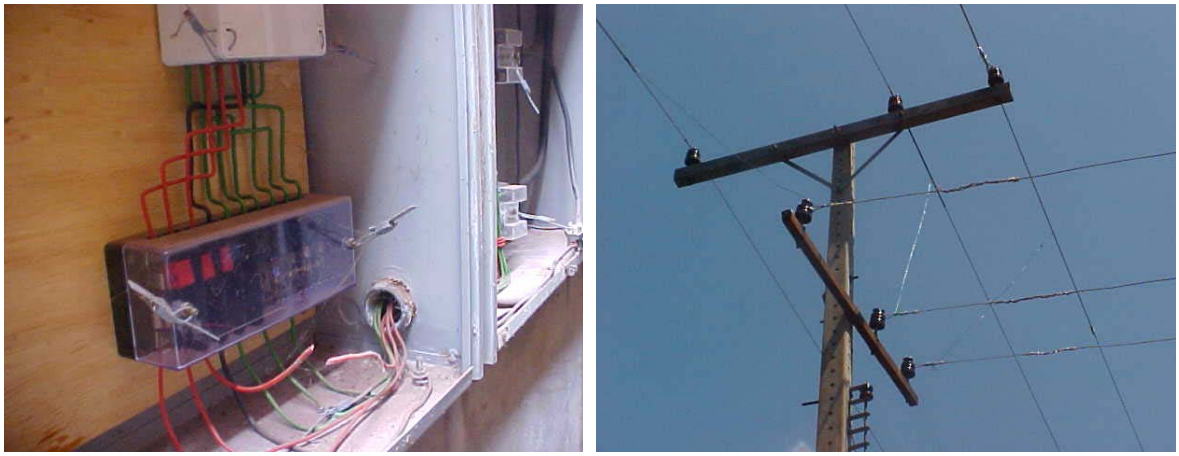
O combate à fraude contra o consumo de energia elétrica é necessário em virtude do número crescente de casos de subtração de energia elétrica ocorrentes em todo o país, e que são tratados pelos concessionários das mais variadas formas, quer com relação à repressão, quer com relação à prevenção. A sistematização das normas e dos procedimentos poderá dar aos concessionários maiores e melhores condições de reduzir, minimizar ou eliminar esse problema, que vem se mostrando de difícil solução, dada a quantidade de maneiras e modos através dos quais vem sendo subtraída ilicitamente a energia.

Segue algumas formas de subtração de energia elétrica em grandes clientes, ligados em alta tensão:

- Alteração nas rotinas internas do hardware do medidor, reprogramando as configurações responsáveis pela contagem de pulsos de energia ativa e reativa (1+1 = 0.5).
- Alteração do valor da constante do medidor através de reprogramação de configurações.
- Substituição de componentes eletrônicos das placas de hardware do medidor;
- Inversão das ligações do medidor;
- Troca do transformador de corrente – TC, alterando a placa de identificação do mesmo;
- Instalação de jumper na chave de aferição provocando um desvio de corrente;

- Descalibração do medidor;
- Ligação direta à rede secundária;
- Corte de fios de ligação entre a chave de aferição e o medidor.

A Figura 1.1 ilustra dois casos reais de fraude: (a) Um dos fios de ligação cortado – um tipo de fraude mais usual, próximo ao medidor e (b) Ligação direta na alta 13,8KV – um caso mais difícil devido ao perigo e por ser de fácil identificação.



(a) (b)

Figura 1.1 – (a) Fio cortado, (b) Ligação direta na alta 13,8 KV.

Em conseqüência aos procedimentos de subtração ilícita de energia, há uma diminuição no consumo mensal medido pela unidade consumidora, provocando um faturamento inferior ao correto para o cliente. Este problema está presente em todos os tipos de consumidores. Entretanto, os clientes atendidos em alta tensão refletem maiores prejuízos às concessionárias devido ao alto consumo de energia e às tarifas diferenciadas para demanda e consumo.

No intuito de contribuir para a solução deste problema identificando fraudes ou erros de medição em unidades consumidoras de energia elétrica, propõe-se neste trabalho, a utilização de informações específicas embasadas na regulamentação do setor, obtidas através de dados históricos e dados em tempo real.

A concessionária de energia elétrica, através da Resolução 456 da ANEEL [10], Art. 21, deve organizar e manter atualizados cadastros relativos às unidades consumidoras, onde conste, obrigatoriamente, para cada uma delas, no mínimo as seguintes informações:

I. Identificação do consumidor;
II. Número ou código de referência da unidade consumidora;
III. Endereço da unidade consumidora, incluindo o nome do município;
IV. Classe e subclasse se houver, da unidade consumidora;
V. Data de início de fornecimento;
VI. Tensão nominal de fornecimento;
VII. Potencia disponibilizada e, quando for o caso, a carga instalada declarada ou prevista no projeto de instalações elétricas;
VIII. Valores de demanda de potência e consumo de energia elétrica ativa, expressos em contrato, quando for o caso;
IX. Informações relativas aos sistemas de medição de demanda de potência e de consumo de energia elétrica ativa e reativa, de fator de potência e, na falta destas medições, o critério de faturamento;
X. Histórico de leitura e de faturamento referentes aos últimos 60 (sessenta) ciclos consecutivos e completos, arquivados em meio magnético, inclusive com as alíquotas referentes a impostos incidentes sobre o faturamento realizado;
XI. Código referente à tarifa aplicável;
XII. Código referente ao pagamento de juros do Empréstimo Compulsório/ ELETROBRÁS.

Para constatar fraude em um consumidor de energia é imprescindível que haja uma inspeção técnica *in loco* (Resolução 456 da ANEEL, Art. 72), e esta identifique a fraude fisicamente ou prove a redução de consumo registrado na unidade, através de medição paralela realizada pela concessionária. Portanto, o processo de inspeção é necessário e dispendioso para qualquer empresa.

1.3. Justificativa

A fraude de energia elétrica é um dos dois tipos de perdas de energia que sempre estiveram presentes no sistema elétrico brasileiro: perdas técnicas decorrentes do processo de transmissão e distribuição de energia e, as perdas comerciais ocasionadas por ligações clandestinas, fraudes, auto-religação, erros não intencionais e outros. O problema “perdas”, no entanto, foi negligenciado no passado,

no modelo verticalizado, uma vez que os prejuízos financeiros causados por tais problemas, ou eram repassadas para as tarifas, ficando o consumidor com a responsabilidade de assumir ônus da ineficiência do sistema, ou o tesouro assumia o prejuízo das empresas estatais, passando para o contribuinte esta responsabilidade.

Em meados dos anos 90, com o advento das privatizações do setor elétrico, as tarifas de energia elétrica passaram a ser estabelecidas nos Contratos de Concessão. Surge então, em 1997, a Agência Nacional de Energia Elétrica - ANEEL, autarquia em regime especial, vinculada ao Ministério de Minas e Energia - MME, tendo como atribuições: regular e fiscalizar a geração, a transmissão, a distribuição e a comercialização da energia elétrica, atendendo reclamações de agentes e consumidores com equilíbrio entre as partes e em benefício da sociedade. A partir de então, através das diversas medidas adotadas pelo órgão regulador, o enfoque ao tratamento das perdas mudou, uma vez que elas passaram a influenciar o lucro das companhias de distribuição.

A ANEEL vem diminuindo, gradativamente, os valores percentuais das perdas de energia repassáveis às tarifas dos consumidores, obrigando as concessionárias a adotar medidas efetivas para coibir a fraude de energia e a inadimplência dos usuários do sistema.

Considerando que as fraudes e os problemas na medição, ocorridos em unidades consumidoras de energia, têm um considerável fator participativo quanto às perdas comerciais das distribuidoras, e que, as concessionárias, levadas pelas regulamentações do setor elétrico, devem dispensar grande atenção para o combate aos atos ilícitos de furto de energia e ao problema na medição, e que dentro das fraudes, os clientes ligados em alta tensão trazem grandes prejuízos financeiros às concessionárias, quando fraudadores, apresenta-se neste trabalho um sistema computacional para identificação de fraudes ou erros na medição em unidades consumidoras ligadas à alta tensão, em tempo real, usando Rough Sets, com o objetivo de reduzir tais prejuízos.

1.4. Objetivos da Pesquisa

1.4.1. Objetivo Geral

Como o intuito do trabalho é detectar fraudes ou erros de medição através de dados, em unidades consumidoras de energia elétrica atendidas em alta tensão, nesta dissertação desenvolve-se uma metodologia e um sistema computacional para detectar tais possibilidades, baseados em dados históricos e dados em tempo real, utilizando a teoria de Rough Sets. O sistema classifica os consumidores em normais ou anormais, selecionando os anormais para serem inspecionados por uma equipe técnica. Logo, este trabalho, contribui para um aumento na taxa de acerto na identificação de fraudadores.

1.4.2. Objetivos Específicos

Este trabalho foi desenvolvido considerando os seguintes objetivos específicos:

- Determinar atributos relevantes dos grandes consumidores de energia elétrica;
- Analisar bancos de dados de grandes consumidores para estabelecer padrões sobre dados normais e anormais a partir dos atributos levantados;
- Aplicar, com desenvoltura, a Técnica de Rough Sets na redução de atributos em sistemas de informação;
- Definir um padrão de comportamento para cada consumidor analisado.

1.4.3. Metodologia

Este trabalho foi desenvolvido considerando a seguinte metodologia:

- Estudo e análise da Descoberta de Conhecimento em Base de Dados – DCBD (*Knowledge Discovery in Database*);
- Estudo e análise da teoria de Rough Sets para a classificação de dados;
- Apresentação da metodologia para a detecção de fraude ou erro de medição em grandes clientes consumidores de energia elétrica;
- Aplicação da metodologia proposta no desenvolvimento de um sistema

computacional com o objetivo de classificar o comportamento das unidades consumidoras de energia elétrica, através de perfis diários, em normais ou anormais, sendo a última classificação correspondendo a possível fraude ou erro de medição.

1.5. Organização dos Capítulos

O texto está dividido em capítulos, que descrevem em detalhes todos os resultados obtidos na busca do atendimento dos objetivos enunciados na seção anterior. Sendo assim, será descrita a seguir a forma de organização em termos dos objetivos propostos:

- O Capítulo 1 contém a Introdução, no qual é apresentada a motivação da pesquisa, informações de fraudes em diversos segmentos, fraude especifica do setor elétrico e uma visão geral do conteúdo da dissertação.
- No Capítulo 2 é tratado a Descoberta do Conhecimento em Base de Dados. Neste capítulo encontram-se os conceitos, as definições, a estrutura e as técnicas de DCBD, de forma resumida.
- No Capítulo 3 é abordado o aspecto teórico de Rough Sets. Aqui se apresenta o embasamento teórico de suporte para a análise realizada sobre o método de Rough Sets.
- No Capítulo 4 é tratado a Classificação de Dados usando Rough Sets. Este capítulo apresenta possibilidade de generalização da informação, geração de regras e redução de regras através dos parâmetros suporte e consistência.
- No Capítulo 5 é apresentada a Metodologia para Detecção de Fraude ou Erro de Medição em Grandes Clientes. Apresentam-se primeiramente as terminologias convencionadas pelo setor elétrico a fim de proporcionar ao leitor uma familiarização da linguagem utilizada na dissertação e, as principais informações dos grandes clientes consumidores de energia elétrica, ligados à alta tensão. Por fim, é apresentada a metodologia proposta com o objetivo de identificar fraudes ou erro de medição em unidades consumidoras de energia elétrica.
- No Capítulo 6 é realizado o Estudo de Caso, onde são apresentados os resultados obtidos através da implementação da metodologia proposta em um

sistema computacional desenvolvido. Apresentam-se os resultados seguidos de comentários, obtidos através da variação de dois parâmetros: valor nominal do atributo de decisão e valor da consistência admitida para as regras de classificação.

- O presente trabalho encerra-se com o Capítulo 7 contendo as conclusões da pesquisa realizada e proposta de trabalhos futuros. Enumera-se nesse capítulo as contribuições realizadas e as questões que não foram implementadas.

Capítulo 2
Descoberta do Conhecimento em Banco de Dados

2.1. Introdução

Nas duas últimas décadas houve um aumento significativo na quantidade de informações armazenadas em banco de dados no formato eletrônico, devido a grande capacidade computacional de armazenamento destes dados. Com tanta informação armazenada, o problema passou a ser o que fazer com estes dados valiosos[1].

As tecnologias atuais fazem à coleta de dados de maneira rápida e fácil, entretanto as análises de dados tendem a ser lentas e custosas devido a grande quantidade de informações armazenadas. Há uma hipótese de que parte das informações úteis podem estar escondidas na massa de dados analisada e, portanto, métodos semi-automáticos para localizar informações interessantes a partir dos dados podem ser úteis [13].

A Descoberta de Conhecimento em Banco de Dados (DCBD) é uma metodologia que tem por objetivo a obtenção de conhecimento extraindo informações potencialmente úteis e não triviais de grandes conjuntos de dados [17]. O conhecimento extraído pode ser expresso através de regras que descrevem as propriedades dos dados, os padrões mais freqüentes, agrupamentos de objetos na base de dados, ou ainda pela classificação dos dados [18].

Como a literatura sobre DCBD é extensa, neste capítulo, optou-se por fazer apenas uma breve introdução. Definições mais precisas e discussões mais profundas sobre este tema podem ser encontradas em [11, 12, 13, 14, 15, 16, 17, 18].

[1] Este capítulo está baseado nos trabalhos [11] e [12]

2.2. O Processo de Descoberta de Conhecimento em Base de Dados

O fato de introduzir uma grande quantidade de dados em um determinado programa ou sistema não implica em obtenção direta de conhecimentos úteis extraídos desses dados. O objetivo da DCBD é justamente obter conhecimento útil a partir de grandes bancos de dados. Para tanto, o usuário de um sistema que utiliza um método de DCBD seleciona subconjuntos de dados, escolhe classes de padrões apropriados e utiliza critérios para definir padrões de interesse. Portanto, os sistemas que utilizam o método de DCBD devem ser vistos como sistemas que utilizam uma ferramenta interativa e não como um sistema de análise automático.

No processo de DCBD, pode-se obter vários resultados possíveis:

- Confirmação do óbvio, quando a informação obtida após o processo continua sendo a mesma que se tinha antes do processo;
- Novo conhecimento, quando é descoberto o conhecimento associado ao conjunto de dados. Isto é, quando se obtém informações inspiradas ou que não se mostraram evidentes a partir da observação dos dados;
- Nenhuma relação significativa, quando a análise dos dados não leva a nenhum resultado significativo. Isto ocorre porque os atributos dos dados são insuficientes para determinar os padrões existentes nos dados ou não têm nenhuma correlação com os mesmos.

A DCBD é um processo que contém diversas etapas. Um ciclo completo deste processo inspirado em Fayyad [18], está representado pela Figura 2.1.

Na primeira etapa ocorre a consolidação dos dados, seleção e pré-processamento. Nesta etapa os dados são gravados nos chamados “armazéns de dados”. A segunda etapa refere-se à mineração de dados que é a maior componente da DCBD, pois examina uma grande quantidade dados, buscando descobrir as relações não explícitas entre esses dados, com a finalidade de obter modelos que representem o sistema analisado. A terceira e última etapa do ciclo da DCBD, refere-se à interpretação e avaliação que, através dos modelos obtidos na etapa anterior, torna possível interpretar e avaliar o conhecimento descoberto para, finalmente,

disponibilizar o resultado. Estas etapas são descritas com mais detalhes nos próximos itens.

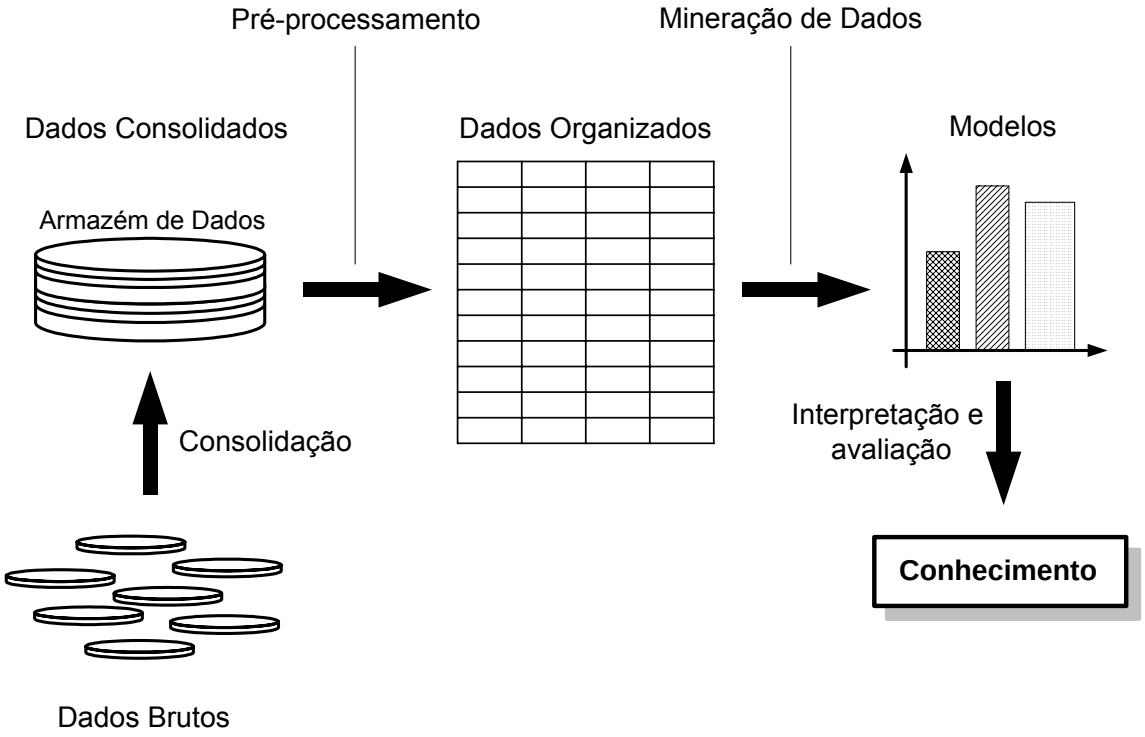


Figura 2.1 – Processo completo da descoberta de conhecimento em base de dados.

2.2.1. Etapa de Armazenamento de Dados

A primeira etapa para descoberta de conhecimento em grande quantidade de dados envolve as seguintes operações: consolidação dos dados, seleção e pré-processamento.

Consolidação dos Dados

Consolidar os dados é estruturá-los de forma conveniente para serem explorados e armazenados. Portanto, inicia-se esta etapa com a preparação do conjunto de dados referente ao assunto que se deseja obter conhecimento. Na Figura 2.2 pode-se observar a etapa de consolidação dos dados.

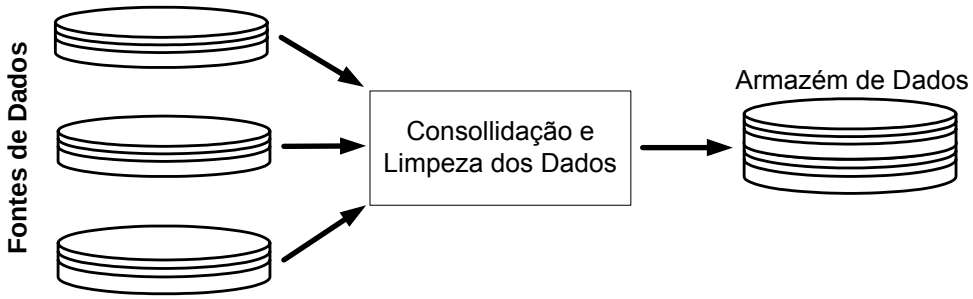


Figura 2.2 – Fonte de dados para o armazém de dados.

As fases ocorridas na consolidação dos dados podem ser entendidas da seguinte forma:

- Fonte de Dados: selecionam-se nesta fase os dados a serem utilizados e podem provir de diversos sistemas de gerenciamento de dados [19];
- Consolidação e Limpeza dos Dados: identificam-se nesta fase as inconsistências dos dados, perdas, redundância, etc. Após a identificação destes problemas, realizam-se as limpezas dos erros nos dados;
- Armazém de Dados: armazenamento dos dados consolidados, ou seja, este é o resultado da consolidação.

Essa etapa pode tomar mais de 80% do tempo previsto em todo processo da DCBD, no entanto é importante tomar o tempo que seja necessário, uma vez que a qualidade dos resultados está diretamente relacionada com a qualidade dos dados [19].

Seleção e Pré-Processamento

O ponto de partida da DCBD, representado na Figura 2.1 por dados brutos, refere-se a um conjunto de dados. Após a consolidação dos dados, em que estes passam por uma estruturação conveniente para serem explorados e armazenados, inicia-se a seleção e pré-processamento, que completam as seguintes tarefas:

- Criar um conjunto de dados alvo a partir de todos os dados disponíveis. Pode-se usar o método de amostragem para criar este conjunto.
- Retirar atributos que são desnecessários sem perder a informação. Todos os

atributos que possam ter alguma influência sobre o resultado final devem ser considerados. Se alguns desses atributos não forem relevantes para o processo, os algoritmos de mineração de dados descobrirão e se terá certeza, de modo científico e não indutivo, de que o atributo não é relevante.

- Finalmente realiza-se o pré-processamento. Nesta operação eliminam-se registros incompletos que não tenham sido localizados e nem eliminados na limpeza dos dados. Pode-se calcular nesta fase alguns atributos do tipo média, desvio-padrão, dentre outros.

2.2.2. Etapa de Mineração dos Dados

A etapa de Mineração de Dados é uma parte da descoberta do conhecimento e depende fundamentalmente do método utilizado para o tratamento dos dados. Esse é o passo onde os padrões freqüentes e de interesse são descobertos a partir dos dados [20].

Os dois principais objetivos de mineração de dados são predição e descrição.

- Predição: envolve o uso de algumas variáveis nas bases de dados para predizer valores futuros ou desconhecidos de outras variáveis de interesse.
- Descrição: busca obter padrões que descrevam os dados e aprender uma hipótese generalizada, um modelo, a partir dos dados selecionados.

A etapa de Mineração de Dados caracteriza-se pela existência de um algoritmo que, diante da tarefa de DCBD especificada, será capaz de extrair eficientemente o caminho implícito e útil de um banco de dados. As tarefas da DCBD dependem do domínio da aplicação e do interesse do usuário. De modo geral, cada tarefa de DCBD extrai um tipo diferente de conhecimento do banco de dados, logo cada tarefa requer um algoritmo diferente para a extração de conhecimento. Algumas tarefas são:

Classificação – essa tarefa utiliza o aprendizado como um modelo preditivo, isto é, a base de dados é o meio que a Mineração de Dados tem para inferir um modelo. Na forma supervisionada classifica-se os dados dentro de uma ou várias classes pré-definidas pelo usuário a partir do banco de dados. Supõe-se

que a base de dados contém um ou mais atributos que denotam a classe de um registro. Esses atributos são chamados de “atributos de classificação”. Os atributos restantes são chamados “atributos de descrição”. Aprender as regras de classificação significa que o sistema tem que obter as regras que predizem as classes para as quais os registros irão pertencer a partir dos atributos de classificação, como mostrado na Figura 2.3. Primeiramente, o usuário define as condições para cada classe fracionando, posteriormente, à base de dados em subconjuntos de classes $C_1,\cdots,C_n$. Assim, o sistema de Mineração de Dados terá de construir as descrições $D_1,\cdots,D_n$ para estas classes fracionadas [17].

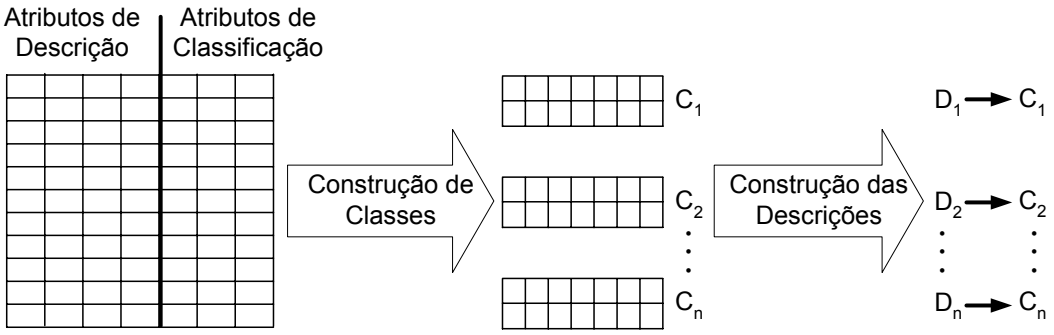


Figura 2.3 – Aprendizado das regras de classificação através de base de dados.

- Agrupamento – é uma tarefa na qual se procura identificar um conjunto finito de categorias ou grupos para descrever os dados. Nessa fase, o algoritmo deve criar as classes através da produção de partições de conjuntos de dados em conjuntos de registros. Essa partição é feita de modo que registros com valores semelhantes, ou seja, com propriedade de interesses comuns, seja reunidas em uma mesma classe. Uma vez que as classes estejam definidas, pode-se aplicar um algoritmo de classificação, produzindo, assim, regras para cada uma delas. A qualidade do resultado do agrupamento depende da medida utilizada para estabelecer a similaridade usada pelo método e sua implementação, além de sua habilidade para descobrir padrões escondidos.

- Dependência do Modelo – essa tarefa consiste em obter um modelo que

descreva dependências significativas entre os atributos. Essas dependências existem em dois níveis: estrutural e quantitativo [18]. O nível estrutural do modelo especifica quais atributos são localmente dependentes dos outros. O nível quantitativo do modelo especifica a força das dependências, utilizando alguma escala numérica [21]. Assim, essa tarefa busca descobrir automaticamente tais dependências.

- Regressão – essa tarefa é conceitualmente similar a tarefa de classificação. A grande diferença é que nessa tarefa o objetivo é contínuo, ou seja, pode ter qualquer valor real ou qualquer número inteiro num intervalo arbitrário, ao invés de um valor discreto [22]. Assim, regressão representa o aprendizado através de uma função que mapeia um conjunto de dados para uma variável denominada valor real de predição.

Portanto, os objetivos da predição ou descrição dependem do escopo do usuário. Depois do objetivo ser definido, acima do domínio de aplicação, é determinada à técnica de Mineração de Dados a ser utilizada.

A Tabela 2.1 mostra algumas tarefas da DCBD e as técnicas utilizadas na mineração de dados.

Tabela 2.1 – Tarefas da DCBD e suas técnicas de mineração de dados.

Tarefas da DCBD	Técnicas de Mineração de Dados
Classificação	Evolucionárias, Conexionistas, Árvores de decisão, Rough Sets
Agrupamento	Conexionistas, Nebulosas, Estatísticas
Dependências	Estatísticas
Regressão	Estatísticas

2.2.3. Etapa de Interpretação e Avaliação

Os padrões descobertos pelo processo de DCBD devem ser analisados pelo usuário, pois o mesmo informa quando os padrões são descobertos. O processo de DCBD não termina quando os padrões são descobertos. O usuário tem de ser capaz de compreender o que foi descoberto para ver os dados e conferir os padrões descobertos com o conhecimento de base. Para verificar se os resultados obtidos

possuem uma correspondência com os dados, deve-se passar por um processo de interpretação e avaliação, da seguinte forma:

- Interpretação: pode ser feita através de árvores indutivas e modelos de regras, que podem ser interpretados diretamente. O agrupamento dos resultados pode ser apresentado sob forma de gráficos e/ou tabelas.
- Avaliação: é feita através de validação estatística, sendo necessária também a revisão qualitativa pelo especialista da área.

Nesta etapa as ferramentas de visualização são muito úteis na análise de sensibilidade (relação entrada/saída), histograma de distribuição de freqüências.

2.2.4. Considerações Finais

Como foi apresentada, a DCBD tem o objetivo de extrair conhecimentos em base de dados, e uma das vantagens dessa metodologia é que não necessita de bases de dados perfeitas [13], ela precisa apenas de base de dados organizados. A DCBD é útil em muitas aplicações para solucionar problemas do mundo real. Aplicações dessa metodologia podem ser encontradas em diversas áreas [14, 15, 16] e ela utiliza como algoritmo de mineração diversos métodos, tais como, Estatísticas, Redes Neurais, Algoritmos Genéticos, Lógica Nebulosa, Rough Sets, etc. No próximo capítulo, apresenta-se a Teoria de Rough Sets, a qual será utilizada no algoritmo de mineração para a classificação de dados.

Capítulo 3
Rough Sets

3.1. Introdução

A quantidade de informações disponíveis cresceu exponencialmente nos últimos anos com o advento da evolução tecnológica, o que levou a problemas opostos de 20 anos atrás. Hoje, a quantidade de dados é tão grande que se torna mandatório a elaboração de modelos que auxiliem na tomada de decisões gerenciais.

Normalmente, num banco de dados, tem-se um número muito grande de atributos armazenados dos quais parte são relevantes e partes são irrelevantes, desnecessários, para a tomada de uma específica decisão.

Por exemplo, uma decisão gerencial precisa ser tomada em uma empresa, com base nas informações existentes em seu banco de dados. Dos dados ali armazenados, nem todos são necessários para a tomada da decisão, pois não alterarão o resultado. Estas informações que não modificam os resultados com relação à decisão são denominadas como atributos irrelevantes. Então, pode-se dizer que, para obter um modelo que auxilie em uma decisão gerencial específica, apenas uma quantidade reduzida de atributos são fundamentais na avaliação, ou seja, necessários para se chegar a uma decisão. Deve-se, portanto, separar os atributos fundamentais dos irrelevantes de forma a reduzir a quantidade de atributos a serem utilizados na tomada da decisão e, conseqüentemente, no modelo a ser definido neste processo. A Figura 3.1 representa esta redução.

A teoria de Rough Sets, que foi introduzida primeiramente por Zdzislaw Pawlak em 1982 [23], possui propriedades que permitem eliminar variáveis ou atributos irrelevantes através do processo de redução do sistema de informação, baseando-se na definição de redutos, os quais são subconjuntos de atributos capazes de manter as mesmas propriedades da representação de conhecimento quando esta é feita

utilizando todos os atributos. Este procedimento de eliminação de atributos irrelevantes é uma das características dessa teoria.

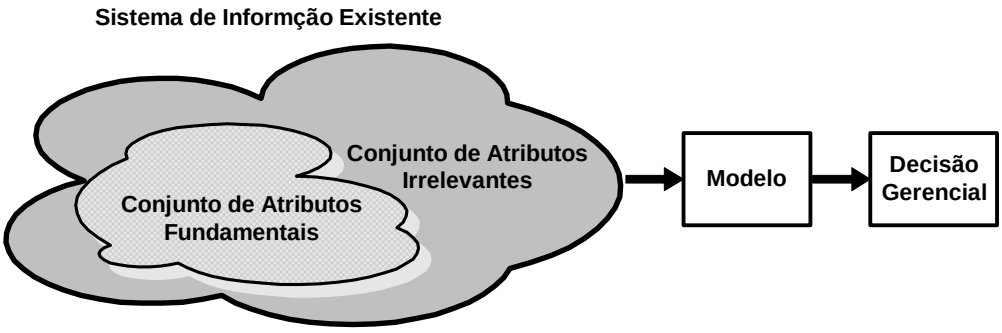


Figura 3.1 – Representação simplificada do processo de modelagem.

Rough Sets também proporciona grande habilidade na classificação de objetos. Os objetos contidos em um sistema, de acordo com suas características, são agrupados em classes. Os objetos agrupados em uma mesma classe são indiscerníveis entre si. Como na maioria dos bancos de dados há informações imprecisas, a teoria de Rough Sets é capaz de administrar estas imprecisões, informações ruidosas e incompletas presentes nestes sistemas. Assim, objetos que não podem ser especificados através dos dados disponíveis são classificados por esta teoria através de dois conceitos: as aproximações inferior e superior, que serão abordados mais adiante.

A fundamentação matemática desta teoria permite a descoberta de padrões ocultos na base de dados. Sua utilidade no campo de mineração de dados, em processamento de sinais de som e imagem, clusterização, entre outros, pode ser comprovada pelo crescente número de aplicações divulgadas com estes conteúdos [23, 24].

Neste capítulo, o assunto é abordado da seguinte forma: após uma breve introdução sobre a teoria de Rough Sets, seus conceitos e descrições é apresentado, seguidos de exemplos para facilitar o entendimento. Primeiramente, é apresentado um sistema de informação, seguido da definição de indiscernibilidade, aproximações dos conjuntos, bem como a qualidade dessas aproximações. A seção seguinte trata a definição da redução da informação que será é subdividida em três partes: matriz de discernibilidade, função de discernibilidade e redução. Após a redução da informação

é tratada a geração de regras e, por fim, são apresentadas as considerações finais sobre a utilização de Rough Sets em classificação de banco de dados.

3.2. Sistema de Informação e de Decisão

A forma mais comum para representação dos dados na abordagem de Rough Sets é através de um sistema de informação que contém um conjunto de objetos, sendo que cada objeto tem uma quantidade de atributos. Esses atributos são os mesmos para cada um dos objetos, mas seus valores nominais podem diferir. Portanto, um sistema de informação pode ser representado, por exemplo, pela Tabela 3.1 que refere-se a um conjunto de crianças e um conjunto de características de alguns brinquedos.

Tabela 3.1 – Sistema de informação.

U	*Atributos Condicionais (C)*				
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*
6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*
9	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*

Assim, um sistema de informação pode ser indicado por:

$$A = (U, C) \tag{3.1}$$

sendo U o conjunto de objetos e C o conjunto de atributos disponíveis na base de dados.

Em muitos casos é importante a classificação dos objetos considerando um atributo que informa a decisão a ser tomada. Pode-se assumir, portanto, que um sistema de informação munido de um atributo de decisão é denominado sistema de decisão, denotado por $\Re$:

$$\Re = (U, C \cup D) \tag{3.2}$$

sendo D o conjunto de atributos de decisão.

Um sistema de decisão $\Re$ pode ser representado pela Tabela 3.2, o qual é capaz de fornecer a atitude das crianças com relação aos brinquedos [25].

Tabela 3.2 – Sistema de decisão.

U	*Atributos Condicionais (C)*					*Atributo de Decisão (D)*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	*Negativa*
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*	*Neutra*
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*	*Positiva*
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	*Negativa*
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*	*Neutra*
6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Positiva*
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*	*Positiva*
9	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Neutra*
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*	*Neutra*

Para o sistema de decisão descrito na Tabela 3.2, tem-se:

a) conjunto de objetos ou registros: $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$;
b) conjunto de atributos condicionais: $C = \{\text{Cor, Tamanho, Tato, Textura, Material}\}$;
c) conjunto do atributo de decisão: $D = \{\text{Atitude}\}$;

Assim, como pode haver mais de uma característica ou valor nominal para cada atributo condicional, pode haver mais de uma opção de característica para o atributo de decisão.

Os valores nominais dos atributos considerados no sistema de decisão da Tabela 3.2, estão representados na Tabela 3.3.

Tabela 3.3 – Valores nominais dos atributos.

	Atributo	**Valores Nominais**
Atributos Condicionais	Cor	*Azul, Vermelho, Amarelo e Verde.*
	Tamanho	*Grande, Médio e Pequeno.*
	Tato	*Duro, Moderado e Macio.*
	Textura	*Liso, Áspero e Indefinido.*
	Material	*Plástico, Madeira, Pelúcia e Metal.*
Atributo de Decisão	Atitude	*Neutra, Negativa e Positiva.*

3.3. Indiscernibilidade

Considerando os Atributos Condicionais (C), para todo subconjunto de atributos $B \subseteq C$ do sistema de informação A, equação (3.1), uma relação de equivalência $IND_A(B)$ é associada no sistema, chamada relação de indiscernibilidade, definida como:

$$IND_A(B) = \{(x,y) \in U^2 \mid \forall c \in B, c(x) = c(y)\} \tag{3.3}$$

Assim, x e y são indiscerníveis entre si para todo atributo de B.

Se a relação de indiscernibilidade existe entre dois objetos, significa que todos os valores de seus atributos são idênticos com respeito ao subconjunto de atributos B considerado, ou seja, não podem ser diferenciados entre si. O conjunto de todas as classes de equivalência determinadas por $IND_A(B)$ é representado por $U/IND_A(B)$, denominado conjunto quociente de U pela relação $IND_A(B)$.

Para a base de dados da Tabela 3.1, alguns dos possíveis subconjuntos dos Atributos Condicionais são: {Cor}, {Tamanho}, {Tato}, {Textura}, {Material}, {Cor, Tamanho}, {Cor, Tato}, {Cor, Textura}, {Cor, Material}, {Tamanho, Tato}, {Tamanho, Textura}, {Tamanho, Material}, {Tamanho, Textura, Material}, entre outras possíveis combinações, num total de 31 (trinta e um) subconjuntos não vazios. Assim, $U/IND_A(B)$ para alguns dos possíveis subconjuntos $B \subseteq C$ são:

$U/IND_A(\{\text{Cor}\})$ = {{1, 4}, {2}, {3, 5, 7, 8}, {6, 9, 10}}

$U/IND_A(\{\text{Tamanho}\})$ = {{1, 6, 9}, {2, 4, 10}, {3, 5, 7, 8}}

$U/IND_A(\{\text{Tato}\})$ = {{1, 6, 7, 8, 9}, {2, 4, 10}, {3, 5}}

$U/IND_A(\{\text{Textura}\})$ = {{1, 5, 7, 8}, {2, 6, 9, 10}, {3, 4}}

$U/IND_A(\{\text{Material}\})$ = {{1, 4, 5, 8, 10}, {2, 6, 9}, {3}, {7}}

$U/IND_A(\{\text{Cor, Tamanho}\})$ = {{1}, {2}, {3, 5, 7, 8}, {4}, {6, 9}, {10}}

$U/IND_A(\{\text{Tamanho, Textura, Material}\})$ = {{1}, {2}, {3}, {4}, {5, 8}, {6, 9}, {7},{10}}

$U/IND_A(\{\text{Cor, Tamanho, Tato, Textura, Material}\})$ = {{1}, {2}, {3}, {4}, {5}, {6, 9}, {7}, {8}, {10}}

Para cada subconjunto de $B \subseteq C$ os objetos são agrupados e os grupos consistem de objetos que são indiscerníveis entre si. De acordo com a teoria de Rough Sets, cada um desses grupos é uma classe de equivalência (Cl).

Considerando o subconjunto $B = \{\text{Tamanho}\}$, os objetos 1, 6 e 9 estão na mesma classe de equivalência e são indiscerníveis, assim como os objetos 2, 4 e 10 e os objetos 3, 5, 7 e 8. Assim, as classes de equivalências para o subconjunto {Tamanho} estão representadas na Tabela 3.4. A classe Cl_1 originou-se dos objetos 1, 6 e 9, a classe Cl_2 dos objetos 2, 4 e 10 e a classe Cl_3 dos objetos 3, 5, 7 e 8.

Tabela 3.4 – Classes para $B = \{\text{Tamanho}\}$.

	Atributo Condicional
Classes	**Tamanho**
Cl_1	*Grande*
Cl_2	*Médio*
Cl_3	*Pequeno*

É importante relembrar que não se analisa o Atributo de Decisão para estes agrupamentos. Os objetos de uma mesma classe podem possuir diferentes valores para o Atributo de Decisão. Por exemplo, na classe Cl_1 o objeto 1 tem como Atributo de Decisão, a Atitude igual à Negativa, assim como para o objeto 6, a Atitude é Positiva e para o objeto 9, a Atitude é Neutra.

3.4. Aproximação de Conjuntos

A Tabela 3.5 apresenta o agrupamento dos objetos do sistema de informação (Tabela 3.2), conforme as características do atributo de decisão.

Inicialmente, seja a seguinte questão: quais as características dos atributos condicionais que definem as atitudes das crianças com relação aos brinquedos como sendo Negativa, Neutra ou Positiva? Pode-se notar que não há uma resposta única para esta pergunta, pois as crianças 6 e 9 apresentam as mesmas características nos Atributos Condicionais, mas diferenciam-se no Atributo de Decisão.

Pode-se dizer **com certeza**, conforme Tabela 3.5, que qualquer criança com características iguais as das crianças 1 ou 4 terão atitude *Negativa*, assim como qualquer criança com características iguais as crianças 2, 5 ou 10 terão atitudes

Neutra e qualquer criança com características iguais as das crianças 3, 7 ou 8 terão atitudes *Positiva*. Nada se pode afirmar para crianças com características iguais as das crianças 6 e 9. São nesses casos que a noção de Rough Sets emerge.

Tabela 3.5 – Atributo de decisão agrupado por seus valores nominais.

U	*Atributos Condicionais (C)*					*Atributo de Decisão (D)*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	*Negativa*
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	*Negativa*
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*	*Neutra*
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*	*Neutra*
9	***Verde***	***Grande***	***Duro***	***Liso***	***Madeira***	***Neutra***
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*	*Neutra*
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*	*Positiva*
6	***Verde***	***Grande***	***Duro***	***Liso***	***Madeira***	***Positiva***
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*	*Positiva*

Seja $A = (U,C)$ um sistema de informação, $B \subseteq C$ e $X \subseteq U$, onde X é o conjunto de objetos ou registros com respeito a B, isto é, X é obtido através das informações dos atributos de B. Assim, define-se Aproximação Inferior de X em relação a B, denotado por $\underline{B}(X)$ e Aproximação Superior de X em relação a B, denotado por $\overline{B}(X)$, como:

$$\underline{B}(X) = \{x \in U \mid U/IND_A(B) \subseteq X\} \tag{3.4}$$

e

$$\overline{B}(X) = \{x \in U \mid U/IND_A(B) \cap X \neq \emptyset\} \tag{3.5}$$

Os objetos ou registros da Aproximação Inferior $\underline{B}(X)$ são classificados **com certeza** como membros de X, utilizando o conjunto de atributos B, enquanto que os objetos da Aproximação Superior $\overline{B}(X)$ **podem** ser classificados como possíveis membros de X, utilizando o mesmo conjunto B. Portanto, obtém-se uma Região de Fronteira de X pela diferença de $\overline{B}(X)$ para $\underline{B}(X)$, denotado por $RF(X)$, ou seja,

$$RF(X) = \overline{B}(X) - \underline{B}(X) \tag{3.6}$$

que consiste de objetos impossíveis de serem classificados em X. Ainda, é possível definir como Fora da Região de X o conjunto $U - \overline{B}(X)$, ou seja, consiste de objetos que não pertencem a X, considerando o mesmo conjunto B.

Um conjunto X é definido como *rough* (impreciso) se sua Região de Fronteira é diferente do conjunto vazio ($RF(X) \neq \emptyset$), e é definido como *crisp* (preciso) se o conjunto for vazio ($RF(X) = \emptyset$).

Seja $B =$ {Cor, Tamanho, Tato, Textura, Material}, $U/IND_A(B)$ = {{1}, {2}, {3}, {4}, {5}, {6, 9}, {7}, {8}, {10}} e $X = \{3, 6, 7, 8\}$, o conjunto formado por crianças que correspondem no atributo de decisão, Atitude = Positiva, tem-se as seguintes aproximações, da Tabela 3.2:

Aproximação Inferior $= \underline{B}(X) = \{\{3\}, \{7\}, \{8\}\}$

Aproximação Superior $= \overline{B}(X) = \{\{3\}, \{6, 9\}, \{7\}, \{8\}\}$

Região de Fronteira $= \overline{B}(X) - \underline{B}(X) = \{\{3\}, \{6, 9\}, \{7\}, \{8\}\} - \{\{3\}, \{7\}, \{8\}\} = \{\{6, 9\}\}$

Fora da Região $= U - \overline{B}(X) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6, 9\}, \{7\}, \{8\}, \{10\}\} - \{\{3\}, \{6, 9\}, \{7\}, \{8\}\}$
$= \{\{1\}, \{2\}, \{4\}, \{5\}, \{10\}\}$

A Figura 3.2 exibe as aproximações definidas pelo exemplo acima em forma de conjuntos.

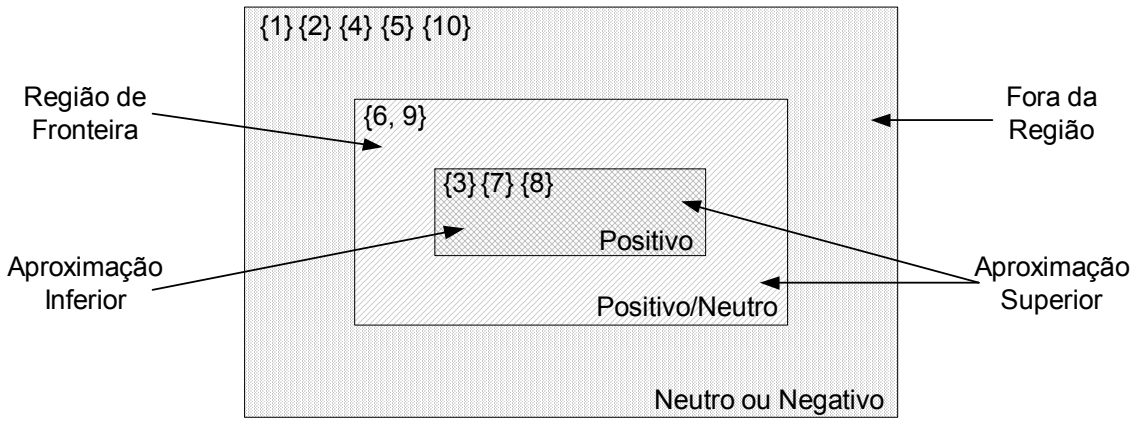


Figura 3.2 – Aproximações em forma de conjuntos.

3.4.1. Qualidade das Aproximações

A qualidade das aproximações obtidas pelas definições dadas previamente pode ser caracterizada numericamente a partir dos próprios elementos que a

definem. O coeficiente para medir essas qualidades é representado por $\alpha_B(X)$, sendo X o conjunto de objetos ou registros com respeito à B, e podem ser realizadas de três formas:

1-Coeficiente de Imprecisão $\alpha_B(X)$, que pode ser entendido como a qualidade da aproximação de X, dado por:

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \tag{3.7}$$

em que $|\underline{B}(X)|$ e $|\overline{B}(X)|$ denotam a cardinalidade das Aproximações Inferior e Superior, respectivamente, e são conjuntos não-vazios ($\overline{B}(X) \neq \emptyset$). Obviamente, $0 \leq \alpha_B \leq 1$.

Se $\alpha_B(X) = 1$, X é *crisp* (preciso) em relação ao conjunto de atributos B.

Se $\alpha_B(X) < 1$, X é *rough* (impreciso) em relação ao conjunto de atributos B.

2-Coeficiente da Qualidade da Aproximação Superior $\alpha_B(\overline{B}(X))$, que pode ser interpretado como sendo o percentual de todos os objetos possivelmente classificados como pertencentes a X, dado por:

$$\alpha_B(\overline{B}(X)) = \frac{|\overline{B}(X)|}{|U|} \tag{3.8}$$

sendo $|U|$ a cardinalidade do conjunto de objetos do Sistema de Informação e, $U \neq \emptyset$.

3-Coeficiente da Qualidade da Aproximação Inferior $\alpha_B(\underline{B}(X))$, que pode ser interpretado como sendo o percentual de todos os objetos certamente classificados como pertencentes a X, dado por:

$$\alpha_B(\underline{B}(X)) = \frac{|\underline{B}(X)|}{|U|} \tag{3.9}$$

Considerando $X = \{3, 6, 7, 8\}$, do exemplo anterior, tem-se:

$\alpha_B(X) = \frac{|\{3, 7, 8\}|}{|\{3, 6, 7, 8, 9\}|} = \frac{3}{5} = 0.6$, ou seja, 60% de X é preciso, com respeito a B.

$\alpha_B(\overline{B}(X)) = \frac{|\{3, 6, 7, 8, 9\}|}{|\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}|} = \frac{5}{10} = 0.5$, ou seja, 50% de U possivelmente pertence à X.

$\alpha_B(\underline{B}(X)) = \frac{|\{3, 7, 8\}|}{|\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}|} = \frac{3}{10} = 0.3$, ou seja, 30% de U certamente pertence à X.

3.5. Redução do Sistema de Informação

A disposição dos dados num sistema de informação não deve possuir redundâncias, pois isso aumenta a complexidade computacional de tal forma que a extração de regras torna-se uma tarefa difícil, desperdiçando tempo e recursos computacionais. Com o intuito de reduzir o Sistema de Informação, tratando essas redundâncias, é que surge o processo de redução de atributos, sem alterar a relação de indiscernibilidade, denominado redução da informação. Essa redução será realizada pela função de discernibilidade a partir da matriz de discernibilidade, que serão estudados à frente.

3.5.1. Matriz de Discernibilidade

Considerando o conjunto de atributos $B = \{\text{Cor, Tamanho, Tato, Textura, Material}\}$ para o sistema de informação A da Tabela 3.1, então o conjunto de todos as classes de equivalência determinadas por B sobre A é dado por $U/IND_A(B) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6, 9\}, \{7\}, \{8\}, \{10\}\}$, que estão representadas na Tabela 3.6.

Tabela 3.6 – Classes de equivalências determinadas por B sobre A.

	Atributos Condicionais (C)				
Classes	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**
Cl_1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*
Cl_2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*
Cl_3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*
Cl_4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*
Cl_5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*
Cl_6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*
Cl_7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*
Cl_8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*
Cl_9	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*

A matriz de discernibilidade de um sistema de informação A, denotada por $M_D(B)$, é uma matriz simétrica $n \times n$ com

$$m_D(i, j) = \{b \in B \mid b(Cl(i)) \neq b(Cl(j))\}, \text{ para } i, j = 1, 2, \cdots, n \qquad (3.10)$$

em que $1 \leq i, j \leq n$, $n = |U/IND_A(B)|$ e, Cl a classe de equivalência formada por objetos indiscerníveis entre si para o subconjunto B. Logo, os elementos da matriz de discernibilidade $m_D(i, j)$ é o conjunto de atributos condicionais de B que diferenciam os objetos das classes (conjunto das classes de equivalência) com relação aos seus valores nominais.

Considerando Cor = Co; Tamanho = Ta; Tato = To; Textura = Te; Material = Ma, com a finalidade da construção da matriz de discernibilidade $M_D(B)$, tem-se na Tabela 3.7 a sua representação.

Tabela 3.7 – Matriz de discernibilidade.

	Cl_1	**Cl_2**	**Cl_3**	**Cl_4**	**Cl_5**	**Cl_6**	**Cl_7**	**Cl_8**	**Cl_9**
Cl_1	Ø								
Cl_2	Co,Ta,To, Te,Ma	Ø							
Cl_3	Co,Ta,To, Te,Ma	Co,Ta,To, Te,Ma	Ø						
Cl_4	Ta,To,Te	Co,Te,Ma	Co,Ta,To, Ma	Ø					
Cl_5	Co,Ta,To	Co,Ta,To, Te,Ma	Te,Ma	Co,Ta,To, Te	Ø				
Cl_6	Co,Te,Ma	Co,Ta,To	Co,Ta,To, Te,Ma	Co,Ta,To, Te, Ma	Co,Ta,To, Te, Ma	Ø			
Cl_7	Co,Ta,Ma	Co,Ta,To, Te,Ma	To,Te,Ma	Co,Ta,To, Te, Ma	To,Ma	Co,Ta,Te, Ma	Ø		
Cl_8	Co,Ta	Co,Ta,To, Te,Ma	To,Te,Ma	Co,Ta,To, Te	To	Co,Ta,Te, Ma	Ma	Ø	
Cl_9	Co,Ta,To, Te	Co,Ma	Co,Ta,To, Te,Ma	Co,Te	Co,Ta,To, Te	Ta,To,Ma	Co,Ta,To, Te,Ma	Co,Ta,To, Te	Ø

A Matriz de discernibilidade é uma matriz simétrica.

3.5.2. Função de Discernibilidade

A função de discernibilidade, denotada por $F_A(B)$, é uma função booleana com m variáveis que determina o conjunto mínimo de atributos necessários para diferenciar qualquer classe de equivalência das demais, definida como:

$$F_A(b_1^*,\ b_2^*,\ \cdots,\ b_m^*) = \wedge\left\{\vee\, m_D^*(i,j) \mid i,j = 1,2,\cdots,n,\quad m_D(i,j) \neq \emptyset\right\} \tag{3.11}$$

Sendo $m_D^*(i,j) = \left\{b^* \mid b \in m_D(i,j)\right\}$.

Utilizando o método de simplificação de expressões booleanas na função $F_A(B)$, obtém-se o conjunto de todos os implicantes primos dessa função, que determina todas as reduções de A. A simplificação é um processo de manipulação algébrica das funções lógicas com a finalidade de reduzir o número de variáveis e operações necessárias para a sua realização.

A função de discernibilidade $F_A(B)$ é obtida da seguinte forma: para os atributos contidos dentro de cada célula da matriz de discernibilidade, aplica-se o operador "soma", "or" ou "$\vee$" e, entre as células dessa matriz, utiliza-se o operador "produto", "and" ou "$\wedge$", resultando em uma expressão booleana de "Produto – da - Soma".

A $F_A(B)$ da Tabela 3.7 é representada por:

$F_A(B)$= (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Ta v To v Te) $\wedge$ (Co v Ta v To) $\wedge$ (Co v Te v Ma) $\wedge$ (Co v Ta v Ma) $\wedge$ (Co v Ta) $\wedge$ (Co v Ta v To v Te) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ma) $\wedge$ (Co v Ta v To v Ma) $\wedge$ (Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (To v Te v Ma) $\wedge$ (To v Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te) $\wedge$ (Co v Te) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (To v Ma) $\wedge$ (To) $\wedge$ (Co v Ta v To v Te) $\wedge$ (Co v Ta v Te v Ma) $\wedge$ (Co v Ta v Te v Ma) $\wedge$ (Ta v To v Ma) $\wedge$ (Ma) $\wedge$ (Co v Ta v To v Te v Ma) $\wedge$ (Co v Ta v To v Te)

Simplificando esta expressão, utilizando teoremas, propriedades e postulados da Álgebra Booleana, obtém-se a seguinte expressão minimizada:

$$F_A(B) = ((\text{Co} \vee \text{Ta} \vee \text{Te}) \wedge \text{To} \wedge \text{Ma}),$$

que ainda pode ser escrita na forma de “Soma – do – Produto”, ou seja,

$$F_A(B) = ((\text{Co} \wedge \text{To} \wedge \text{Ma}) \vee (\text{Ta} \wedge \text{To} \wedge \text{Te} \wedge \text{Ma}))$$

A função de discernibilidade simplificada é dada pelo conjunto mínimo de atributos necessários para discernir as classes formadas por todos as classes de equivalência da relação $IND_A(B)$

3.5.3. Redução

Reduto é um conjunto mínimo de atributos necessários para manter as mesmas propriedades de conhecimento de um sistema de informação, quando este é construida utilizando todos os atributos. Ou seja, o reduto é capaz de classificar objetos, ou classes, sem alterar a representação do conhecimento.

Um Reduto de B sobre um sistema de informação A é um conjunto de atributos B^*, em que $B^* \subseteq B$, sendo todos os atributos $c \in (B - B^*)$ dispensáveis. Com isso, $U/IND_A(B) = U/IND_A(B^*)$.

O conjunto formado pelo termo mínimo da função de discernibilidade $F_A(B)$ determina os redutos de B.

Considerando o exemplo de $B = \{\text{Cor}, \text{Tamanho}, \text{Tato}, \text{Textura}, \text{Material}\}$ e a função de discernibilidade $F_A(B)$ para este conjunto de atributos, previamente definida como $F_A(B) = ((\text{Co} \wedge \text{To} \wedge \text{Ma}) \vee (\text{Ta} \wedge \text{To} \wedge \text{Te} \wedge \text{Ma}))$, o conjunto de redutos desta função é:

$$RED_A(B) = \{\{\text{Cor}, \text{Tato}, \text{Material}\}, \{\text{Tamanho}, \text{Tato}, \text{Textura}, \text{Material}\}\}.$$

Portanto, a $U/IND_A(B^*)$ para cada reduto é:

$$U/IND_A(\text{Cor}, \text{Tato}, \text{Material}) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6, 9\}, \{7\}, \{8\}, \{10\}\}$$

$U/IND_A(\text{Tamanho, Tato, Textura, Material}) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6, 9\}, \{7\}, \{8\}, \{10\}\}$

Assim, $U/IND_A(B)$ = $U/IND_A(B^*)$, portanto, os redutos encontrados através do procedimento da redução da informação são redutos para o sistema de informação da Tabela 3.1.

Podem existir mais de um reduto para um mesmo conjunto de atributos.

O sistema de informação original considerado em nosso exemplo da Tabela 3.1 pode ser representado por um de seus redutos. Considerando o Reduto $B^* = \{\text{Cor, Tato, Material}\}$, temos então a redução do sistema de informação original, representado na Tabela 3.8.

Tabela 3.8 – Redução do sistema de informação original.

U	*Atributos Condicionais*		
Criança	**Cor**	**Tato**	**Material**
1	*Azul*	*Duro*	*Plástico*
2	*Vermelho*	*Moderado*	*Madeira*
3	*Amarelo*	*Macio*	*Pelúcia*
4	*Azul*	*Moderado*	*Plástico*
5	*Amarelo*	*Macio*	*Plástico*
6	*Verde*	*Duro*	*Madeira*
7	*Amarelo*	*Duro*	*Metal*
8	*Amarelo*	*Duro*	*Plástico*
9	*Verde*	*Duro*	*Madeira*
10	*Verde*	*Moderado*	*Plástico*

Os subconjuntos de atributos obtidos através da redução do sistema de informação são capazes de manter as mesmas propriedades da representação de conhecimento quando esta é feita utilizando todos os atributos.

3.6. Geração de Regras

Os atributos encontrados através do método de redução do sistema de informação podem ser descritos na forma de regras, ou seja, as regras de classificação são extraídas do banco de dados reduzido. Para transformar um reduto em regras, deve-se somente unir o atributo de decisão.

Para exemplificar as regras geradas pelos métodos abordados, seja o Reduto $B^* = \{\text{Cor, Tato, Material}\}$ do sistema de decisão $\Re$ da Tabela 3.2. As regras de

decisão que descrevem a dependência de $\{\text{Atitude}\}$ com relação a B^*, podem ser representadas na forma de “Se ... então ...” :

R_1: SE	Cor = *Azul*	E	Tato = *Duro*	E	Material = *Plástico*	ENTÃO	Atitude = *Negativa*
R_2: SE	Cor = *Vermelho*	E	Tato = *Moderado*	E	Material = *Madeira*	ENTÃO	Atitude = *Neutra*
R_3: SE	Cor = *Amarelo*	E	Tato = *Macio*	E	Material = *Pelúcia*	ENTÃO	Atitude = *Positiva*
R_4: SE	Cor = *Azul*	E	Tato = *Moderado*	E	Material = *Plástico*	ENTÃO	Atitude = *Negativa*
R_5: SE	Cor = *Amarelo*	E	Tato = *Macio*	E	Material = *Plástico*	ENTÃO	Atitude = *Neutra*
R_6: SE	Cor = *Verde*	E	Tato = *Duro*	E	Material = *Madeira*	ENTÃO	Atitude = *Positiva*
R_7: SE	Cor = *Amarelo*	E	Tato = *Duro*	E	Material = *Metal*	ENTÃO	Atitude = *Positiva*
R_8: SE	Cor = *Amarelo*	E	Tato = *Duro*	E	Material = *Plástico*	ENTÃO	Atitude = *Positiva*
R_9: SE	Cor = *Verde*	E	Tato = *Duro*	E	Material = *Madeira*	ENTÃO	Atitude = *Neutra*
R_{10}: SE	Cor = *Verde*	E	Tato = *Moderado*	E	Material = *Plástico*	ENTÃO	Atitude = *Neutra*

As regras R_6 e R_9 têm as mesmas características para os Atributos Condicionais Cor, Tato e Material, mas diferentes decisões. Portanto, aplicando esta regra, não se pode afirmar que a decisão será correta. Regras desse tipo são chamadas de não-determinísticas (*inconsistentes*), enquanto que as regras R_1, R_2, R_3, R_4, R_5, R_7, R_8 e R_{10} são chamadas de determinística (*consistentes*). Desta forma, apenas as regras consistentes são consideradas como válidas para classificar o Sistema de Decisão.

O ganho desta técnica para o exemplo considerado na Tabela 3.2 foi a redução de operadores “E”, que passou de quatro para dois. O número de regras de decisão apresentado acima pode ser ainda reduzido e será objeto de estudo para o próximo capítulo.

3.7. Considerações Finais

Neste capítulo, apresentou-se a teoria de Rough Sets baseado no método para aprendizagem e obtenção de conhecimento em banco de dados (DCBD). Definimos primeiramente sistema de informação e de decisão, a indiscernibilidade entre objetos, as aproximações dos conjuntos bem como a qualidade de cada aproximação, a redução do sistema de informação e a geração de regras. Podemos observar que a aproximação de conjuntos está diretamente ligada “a indiscernibilidade ou relação de não-discernibilidade”. A redução do sistema de informação e conseqüentemente a geração de regras são o objetivo desta teoria uma vez que o sistema de informação pode ser desnecessariamente grande. Utilizando a teoria de Rough Sets no exemplo proposto na Tabela 3.1, conclui-se que houve redução de atributos irrelevantes, conforme a Tabela 3.8. Esta redução implica diretamente na diminuição do custo computacional do sistema. As regras de classificação geradas são determinísticas, uma vez que precisam ser consistentes com todos os dados do sistema de decisão.

No próximo capítulo será apresentado um método para obtenção de regras de classificação baseado na teoria de Rough Sets que não irá gerar necessariamente todas as regras de classificação consistentes com o banco de dados. Este método permitirá ao usuário definir um nível mínimo de consistência para a regra de classificação, e a partir de então, gerar regras que atendam a este requisito.

Capítulo 4
Classificação de Dados Usando Rough Sets

4.1. Introdução

A teoria de Rough Sets tem se revelado como sendo uma poderosa ferramenta para mineração de dados e para a DCBD. Diversos métodos baseados nessa teoria foram desenvolvidos para conseguir regras de classificação [26, 27]. Alguns destes métodos são determinísticos, isto é, deles derivam-se regras de classificação totalmente consistentes com os dados do banco de dados. Enquanto esta característica determinística é desejável em algumas aplicações, ela apresenta dois problemas como método de classificação geral:

- Primeiramente, ela não pode trabalhar com ruído de dados de forma efetiva. Ela não considera uma classificação que é 99% consistente com o banco de dados e considera o 1% restante como ruído de dados. Isto pode fazer com que algumas valiosas informações deixem de ser descobertas.
- Em segundo lugar algumas regras de classificação satisfarão um pequeno número de objetos no banco de dados.

A parte mais difícil na classificação de banco de dados é encontrar um número mínimo de regras que caracterizem cada um dos objetos de um sistema de decisão. Apresentou-se no Capítulo 3 a teoria de Rough Sets, capaz de gerar regras de classificação determinística. Será apresentado, neste capítulo, um método baseado em Rough Sets, capaz de gerar regras de classificação não-determinística. O método proposto neste capítulo permite ao usuário especificar o mínimo de consistência que a regra de classificação terá que satisfazer e somente gerar regras que atendam a este requisito. Isto quer dizer que neste método as regras são requeridas para dar suficiente suporte e serem consistentes somente o necessário ao banco de dados.

4.2. Classificação Rough

Nesta seção é acrescentado um parâmetro ao método Rough Sets, apresentado no capítulo anterior, de modo a permitir que as regras de classificação não-determinísticas sejam obtidas. Para melhor entendimento, será utilizado o sistema de decisão representado por R, conforme Tabela 4.1.

Tabela 4.1 – Sistema de decisão – R.

U	*Atributos Condicionais* (*C*)					*Atributo de Decisão* (*D*)	*Numero de Registros Iguais*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	*Negativa*	1
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*	*Neutra*	1
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*	*Positiva*	1
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	*Negativa*	1
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*	*Neutra*	1
6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Positiva*	1
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*	1
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*	*Positiva*	1
9	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Neutra*	1
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*	*Neutra*	1

Todos os atributos de R são bem conhecidos com exceção da última coluna que é usada para indicar o Número de Objetos Iguais – NRI, ocorridos no sistema de decisão. No caso da Tabela 4.1, não existe registros iguais para este sistema, portanto todas as linhas da coluna NRI apresentam valor unitário. Supondo que os registros 6 e 9 fossem completamente iguais, ou seja, os mesmos valores nominais para os atributos condicionais e de decisão, então estes seriam agrupados e a Tabela 4.1 passaria a ser representada pela Tabela 4.2.

Tabela 4.2 – Sistema de decisão agrupado.

U	*Atributos Condicionais* (*C*)					*Atributo de Decisão* (*D*)	*Numero de Registros Iguais*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	*Negativa*	1
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*	*Neutra*	1
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*	*Positiva*	1
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	*Negativa*	1
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*	*Neutra*	1
(6,9)	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Positiva*	2
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*	1
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*	*Positiva*	1
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*	*Neutra*	1

A coluna NRI inserida no sistema é de fundamental importância para este método de geração de regras não-determinísticas, pois ela é usada para obter informações das regras quanto a dois parâmetros inseridos no processo de

classificação: suporte e consistência.

Dada uma regra $r \rightarrow$ "SE (*atributos condicionais*) ENTÃO (*atributo de decisão*)", a partir dessa regra pode-se definir o suporte e a consistência da seguinte maneira:

- *Suporte* – φ. O suporte da regra r é definido como o número de objetos em R que tenham os mesmos valores nominais dos *atributos condicionais* e do *atributo de decisão* de r.
- *Consistência* – δ. A consistência da regra r é definida pela razão entre o número de objetos em R que tenham os mesmos valores nominais dos *atributos condicionais* e do *atributo de decisão* de , ou seja, o suporte de r, e o número de objetos em R que tenham os mesmos valores nominais dos *atributos condicionais*.

Por exemplo: considerando a regra $r \rightarrow$ "SE ($Cor = Verde$ e $Tamanho = Grande$ e $Tato = Duro$ e $Textura = Liso$ e $Material = Madeira$) ENTÃO ($Atitude = Positiva$)", para o sistema de decisão da Tabela 4.1, o suporte e consistência para esta regra será:

- *Suporte*: $\varphi = 1$ (referente ao objeto 6)
- *Consistência*: $\delta = \frac{\varphi}{2} = 0.5$. (referente aos objetos 6 e 9).

Inserindo os parâmetros suporte e consistência na Tabela 4.1, teremos como resultado a Tabela 4.3.

Tabela 4.3 – Sistema de decisão com os parâmetros φ e δ.

U	*Atributos Condicionais (C)*					*Atributo de Decisão (D)*	*Numero de Registros Iguais*	*Suporte*	*Consistência*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**	φ	δ
1	Azul	Grande	Duro	Indefinido	Plástico	Negativa	1	1	1
2	Vermelho	Médio	Moderado	Liso	Madeira	Neutra	1	1	1
3	Amarelo	Pequeno	Macio	Áspero	Pelúcia	Positiva	1	1	1
4	Azul	Médio	Moderado	Áspero	Plástico	Negativa	1	1	1
5	Amarelo	Pequeno	Macio	Indefinido	Plástico	Neutra	1	1	1
6	Verde	Grande	Duro	Liso	Madeira	Positiva	1	1	0.5
7	Amarelo	Pequeno	Duro	Indefinido	Metal	Positiva	1	1	1
8	Amarelo	Pequeno	Duro	Indefinido	Plástico	Positiva	1	1	1
9	Verde	Grande	Duro	Liso	Madeira	Positiva	1	1	0.5
10	Verde	Médio	Moderado	Liso	Plástico	Neutra	1	1	1

4.2.1. Generalização da Informação

O propósito de generalização da informação é reduzir o número de classes em um sistema de decisão de forma que a análise deste sistema fique mais fácil. Como o interesse é identificar, no banco de dados, classes que tenham suporte suficiente, é importante agrupar as classes com suporte insuficiente em uma única classe e, retirá-las do sistema de decisão a ser analisado resultando em uma redução geral de número de classes neste sistema. Assim, para este procedimento utiliza-se o valor do suporte. Neste caso, o φ é utilizado em R para analisar a quantidade de objetos ou registros que classificam-se em cada valor nominal do conjunto do atributo de decisão $D = \{Neutra, Positiva, Negativa\}$.

No sistema de decisão R da Tabela 4.1, existem dois registros com valor nominal do atributo de decisão Atitude = *Negativa*, quatro registro com Atitude = *Positiva* e quatro com Atitude = *Neutra*. Considerando a análise de suporte através do atributo de decisão, pode-se reestruturar a Tabela 4.1, conforme a Tabela 4.4.

Tabela 4.4 – Valor do suporte (φ) para cada valor nominal do atributo de decisão.

U	*Atributo de Decisão (D)*	*Numero de Registros Iguais*	*Suporte*
Criança	**Atitude**	**NRI**	φ
(1, 4)	Negativa	2	2
(2, 3, 5, 10)	Neutra	4	4
(6, 7, 8, 9)	Positiva	4	4

Assumindo, para este sistema, o valor de $\varphi \leq 2$ previamente definido por um especialista, e atendendo o propósito de generalização, precisa-se sinalizar os registros com suporte insuficiente para que estes sejam colocados em uma mesma classe (suporte insuficiente). O algoritmo de que realiza este processo está descrito no Quadro 4.1 que tem por objetivo substituir o valor nominal do atributo de decisão com $\varphi \leq 2$ do sistema de decisão por "-", resultando na Tabela 4.5.

Através da sinalização dos registros com suporte insuficiente, é possível particionar o sistema de decisão em dois subsistemas: conjunto de classes com suporte superior ao suporte previamente definido, denotado por RR e, conjunto de classes com suporte igual ou inferior ao suporte definido, denotado por RI. O subsistema de decisão RI contém todos os objetos com valores "-" para o atributo de

decisão (Tabela 4.6), e o subsistema de decisão RR contém o restante dos objetos (Tabela 4.7).

Quadro 4.1 – Algoritmo: sinalização de registros com suporte φ insuficiente.

```
Entrada: D , NRI e φ (suporte definido pelo usuário)
Saída:   R
Início
        Para classe d ∈ D
                Se NRI(d) ≤ φ
                        d = '–';
                Fim
        Fim
Fim
```

Tabela 4.5 – Generalização para Atitude = *Negativa*.

U	*Atributos Condicionais*					*Atributo de Decisão*	*Numero de Registros Iguais*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	"-"	1
2	*Vermelho*	*Médio*	*Macio*	*Liso*	*Madeira*	*Neutra*	1
3	*Amarelo*	*Pequeno*	*Duro*	*Áspero*	*Pelúcia*	*Positiva*	1
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	"-"	1
5	*Amarelo*	*Pequeno*	*Moderado*	*Indefinido*	*Plástico*	*Neutra*	1
6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Positiva*	1
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*	1
8	*Amarelo*	*Pequeno*	*Moderado*	*Indefinido*	*Plástico*	*Positiva*	1
9	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Neutra*	1
10	*Verde*	*Médio*	*Macio*	*Liso*	*Plástico*	*Neutra*	1

Tabela 4.6 – Subsistema de decisão RI.

U	*Atributos Condicionais*					*Atributo de Decisão*	*Numero de Registros Iguais*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**
1	*Azul*	*Grande*	*Duro*	*Indefinido*	*Plástico*	"-"	1
4	*Azul*	*Médio*	*Moderado*	*Áspero*	*Plástico*	"-"	1

Tabela 4.7 – Subsistema de decisão RR.

U	*Atributos Condicionais*					*Atributo de Decisão*	*Numero de Registros Iguais*
Criança	**Cor**	**Tamanho**	**Tato**	**Textura**	**Material**	**Atitude**	**NRI**
2	*Vermelho*	*Médio*	*Moderado*	*Liso*	*Madeira*	*Neutra*	1
3	*Amarelo*	*Pequeno*	*Macio*	*Áspero*	*Pelúcia*	*Positiva*	1
5	*Amarelo*	*Pequeno*	*Macio*	*Indefinido*	*Plástico*	*Neutra*	1
6	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Positiva*	1
7	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Metal*	*Positiva*	1
8	*Amarelo*	*Pequeno*	*Duro*	*Indefinido*	*Plástico*	*Positiva*	1
9	*Verde*	*Grande*	*Duro*	*Liso*	*Madeira*	*Neutra*	1
10	*Verde*	*Médio*	*Moderado*	*Liso*	*Plástico*	*Neutra*	1

É bem provável que a maioria das regras de classificação potencialmente úteis sejam

derivadas de RR. Esta é uma hipótese considerada razoável, pois de outra maneira, considerando RI na geração das regras de classificação, a qualidade e quantidade tornam-se consideravelmente pequenas devido à insuficiência de suporte, para serem úteis. Este particionamento é viável para banco de dados que contenham grandes números de registros e muitos deles apareçam raramente. O particionamento do sistema de decisão ajuda a economizar tempo computacional, pois resulta num sistema um pouco menor para obtenção de regras, quando esta é realizada sem a partição.

4.2.2. Redução do Sistema de Decisão

O segundo passo é idêntico ao método de redução do sistema de informação demonstrado no Capítulo 3, mas agora, a redução é obtida a partir de RR ao invés de aplicar a todo o sistema R, uma vez que a maioria das regras de classificação potencialmente úteis pertence a este subsistema, (RR), o qual satisfaz o requisito de suporte definido pelo especialista.

Considerando o exemplo de $B = \{\mathrm{Cor}, \mathrm{Tamanho}, \mathrm{Tato}, \mathrm{Textura}, \mathrm{Material}\}$ e aplicando-se os conceitos previamente definidos à Tabela 4.7, tem-se o conjunto dos atributos mínimos para a classificação do subsistema RR, formado por:

$$RED(B) = \{\{\mathrm{Tato}, \mathrm{Material}\}\}.$$

Considerando o reduto $B^* = \{\mathrm{Tato}, \mathrm{Material}\}$, tem-se então a redução do subsistema RR representada na Tabela 4.8.

Tabela 4.8 – Redução do subsistema de decisão RR.

U	*Atributos Condicionais*		*Atributo de Decisão*	*Número de Registros Iguais*	*Suporte*	*Consistência*
Criança	**Tato**	**Material**	**Atitude**	**NRI**	φ	δ
2	*Moderado*	*Madeira*	*Neutra*	1	1	1
3	*Macio*	*Pelúcia*	*Positiva*	1	1	1
5	*Macio*	*Plástico*	*Neutra*	1	1	1
6	*Duro*	*Madeira*	*Positiva*	1	1	0.5
7	*Duro*	*Metal*	*Positiva*	1	1	1
8	*Duro*	*Plástico*	*Positiva*	1	1	1
9	*Duro*	*Madeira*	*Neutra*	1	1	0.5
10	*Moderado*	*Plástico*	*Neutra*	1	1	1

Pode-se notar que a redução é simples, entretanto a atenção deve ser redobrada, pois as regras descobertas desta maneira não garantem total consistência com o sistema de decisão original (Tabela 4.1). Isso pode ocorrer porque as regras podem satisfazer um dado do subsistema de decisão RI com diferentes valores para o atributo de decisão.

Considerando o reduto B^* definido pelo sistema de decisão RR, o sistema RI poderá ser reorganizado conforme Tabela 4.9.

Tabela 4.9 – Redução do subsistema de decisão RI.

U	*Atributos Condicionais*		*Atributo de Decisão*	*Número de Registros Iguais*	*Suporte*	*Consistência*
Criança	**Tato**	**Material**	**Atitude**	**NRI**	φ	δ
1	*Duro*	*Plástico*	"-"	1	1	1
4	*Moderado*	*Plástico*	"-"	1	1	1

Na próxima seção é apresentado um procedimento de como generalizar o conjunto de regras de classificação desta redução que satisfaça os requisitos de suporte e consistência.

4.2.3. Geração de Regras

Para melhor esclarecimento e simplificação, considera-se que a geração de regras será realizada em duas partes:

- Geração de regras,
- Redução de regras.

Desenvolveu-se para este trabalho, um algoritmo que elimina algumas condições desnecessárias contidas nas regras reduzidas obtidas pela redução do subsistema de informação RR, obtendo, assim, um conjunto de regras gerais minimizadas, ou seja, regras com o máximo de condições desnecessárias removidas. Isso é feito suprimindo uma condição por vez para cada regra e verificando se a relação reduzida ainda permanece consistente. Se a regra continuar classificando a mesma quantidade de objetos no subsistema de informação RR, a condição é suprimida da regra. A Tabela 4.10 mostra um conjunto de regras maximamente generalizado, correspondente à redução em RR na Tabela 4.8, denotado por $RR_{\min}$.

Tabela 4.10 – Regras gerais minimizadas - $RR_{\min}$.

U	*Atributos Condicionais*		*Atributo de Decisão*	*Número de Registros Iguais*	*Suporte*	*Consistência*
Criança	**Tato**	**Material**	**Atitude**	**NRI**	φ	δ
2, 10	*Moderado*	"-"	*Neutra*	2	2	1
3	"-"	*Pelúcia*	*Positiva*	1	1	1
5	*Macio*	*Plástico*	*Neutra*	1	1	1
6	*Duro*	*Madeira*	*Positiva*	1	1	0.5
7	"-"	*Metal*	*Positiva*	1	1	1
8	*Duro*	*Plástico*	*Positiva*	1	1	1
9	*Duro*	*Madeira*	*Neutra*	1	1	0.5

O "-" na Tabela 4.10 indica que é uma condição não importante e conseqüentemente, desnecessária para a regra (condição é suprimida).

Realiza-se, então, após a definição de $RR_{\min}$, uma redução de regras para eliminar as regras de classificação que não satisfaçam os requisitos de suporte e consistência definidos pelo especialista, denotado por $RR_{validas}$. Esta redução de regras é descrita pelo algoritmo apresentado no Quadro 4.2.

Quadro 4.2 – Algoritmo: redução de regras.

```
Entrada: RR_min , RI, φ , δ
Saída:  RR_validas
Início
        RR_validas = [ ];
        [lri, cri] = size (RI);
        [lrr, crr] = size (RR_min );
        Para i = 1 : lrr
            Se RR_min (i, crr-1) > φ
                Para j = 1 : lri
                    c = [ ];
                    Se RR_min (i, 2: (crr-4)) = RI(j, 2:(cri-4))
                        c = [c ; RI(j,cri-1) ];
                    Fim
                    δ*(i, 1) = (RR_min (i, crr-1) / (RR_min (i, crr-1) + sum(c )))
                    Se δ*(i, 1) ≥ δ
                        RR_validas = [RR_validas ; RR_min (i, :) δ*(i, 1)];
                    Fim
                Fim
            Fim
        Fim
Fim
```

Para cada $RR_{\min}(i,:)$, primeiramente, checa-se se tem suporte suficiente para o sistema. Caso tenha, verifica-se, então, a sua consistência, avaliando $RR_{\min}(i,:)$ com

o subsistema de decisão RI, pois este contém objetos cobrindo classes com suporte insuficiente. A consistência é facilmente medida por δ^*, e c é o número de objetos de RI que satisfazem apenas os *atributos condicionais* da regra $RR_{\min}(i,:)$.

Levando-se em consideração os valores de suporte e consistência, previamente assumidos, $\varphi = 2$ e $\delta = 0.6$, obtém-se, da Tabela 4.10, o conjunto de regras válidas, conforme Tabela 4.11.

Tabela 4.11 – $RR_{validas}$ para $\varphi = 2$ e $\delta = 0.6$.

	Atributos Condicionais		*Atributo de Decisão*	*Numero de Registros Iguais*	*Suporte*	*Consistência*
Regras	**Tato**	**Material**	**Atitude**	**NRI**	φ	δ
R_1	*Moderado*	"-"	*Neutra*	2	2	0.667

Após modificação dos valores de suporte para $\varphi \leq 1$, e consistência para $\delta = 0.5$, a partir das regras generalizadas maximizadas (Tabela 4.10), obtém-se a Tabela 4.12.

Tabela 4.12 – $RR_{validas}$ para $\varphi \leq 1$ e $\delta = 0.5$.

	Atributos de Classe		*Atributo de Decisão*	*Numero de Registros Iguais*	*Suporte*	*Consistência*
Regras	**Tato**	**Material**	**Atitude**	**NRI**	φ	δ
R_1	*Moderado*	"-"	*Neutra*	2	2	0.667
R_2	"-"	*Pelúcia*	*Positiva*	1	1	1
R_3	*Macio*	*Plástico*	*Neutra*	1	1	1
R_4	*Duro*	*Madeira*	*Neutra*	1	1	0.5
R_5	"-"	*Metal*	*Positiva*	1	1	1

A alteração dos valores de suporte e consistência permite encontrar regras de classificação mais generalizadas, porém, menos precisas. Os valores ótimos para estes parâmetros dependem do grau de confiabilidade necessário para o problema a ser analisado.

As principais características dos procedimentos aqui apresentados são:

- uso específico de dois parâmetros no processo de aprendizagem, suporte e consistência;
- o parâmetro suporte é usado para generalizar um determinado sistema de informação;
- a redução da informação é efetuada somente no subconjunto do sistema de informação, que tem suficiente suporte do banco de dados. Isto é útil no

manuseio de grandes bancos de dados;

- as regras podem ter uma certa inconsistência com os dados, desde de que elas sejam suficientemente consistentes com o esperado.

Este método pode transformar-se no método apresentado no capítulo anterior, desde que se faça $\varphi = 1$ e $\delta = 1$.

4.3. Considerações Finais

Neste capítulo, apresentou-se a teoria de Rough Sets como um método para aprendizagem e obtenção de conhecimento em banco de dados. Este método é capaz de gerar regras de classificação não-determinística. O nível de consistência a ser tolerado pelas regras é especificado pelo usuário, e depende do grau de confiabilidade necessário a ser considerado no problema que será analisado. A utilização de regras não-determinísticas para a classificação de dados, é particularmente, útil para lidar com dados ruidosos, gerando regras mais gerais e confiáveis na análise de grandes bancos de dados.

Capítulo 5
Metodologia para Detecção de Fraude ou Erro de Medição em Grandes Clientes

5.1. Introdução

O objetivo deste trabalho é apresentar um sistema para detecção de fraude ou erro de medição em unidades consumidora de energia elétrica. Para tanto, a metodologia desenvolvida, define perfis de comportamentos diários de grandes clientes consumidores de energia elétrica, que são utilizados para gerar regras que classificam o consumidor em normal (ausência de suspeita de fraude ou erro de medição) ou anormal (unidade consumidora suspeita de anormalidade).

Para que se tenha êxito na utilização do sistema de detecção de fraude desenvolvido, é necessário que as empresas concessionárias de energia disponham de uma plataforma tecnológica de telemedição, que fornece informações, em tempo real, das unidades consumidoras, afim de que estas possam ser monitoradas diariamente.

Resumidamente, a telemedição divide-se em três etapas: a primeira etapa denominada de aquisição de dados que é feita por uma remota, conhecida como *gateway*, que se encontra acoplada ao medidor de energia em cada consumidor individualmente; a segunda etapa designada transmissão dos dados que é realizada através de um modem GPRS, com conexão periódica definida pela concessionária de energia; a terceira etapa consiste no armazenamento dos dados recebidos periodicamente em um servidor de dados.

Com a aquisição, transmissão e armazenamento de dados em tempo real, é possível monitorar todos as unidades consumidoras on-line, disparando alarmes caso haja anormalidade na medição. Tais alarmes estão diretamente ligados com o perfil

gerado para cada consumidor, baseados em dados históricos e dados em tempo real, utilizando técnicas de inteligência artificial.

Para apresentação da metodologia usada no desenvolvimento do sistema de detecção de fraude, faz-se necessário a apresentação de alguns conceitos e terminologias aqui empregados e apresentação das informações dos consumidores de energia elétrica ligados à alta tensão.

5.2. Terminologia

Para melhor compreender as terminologias mais amplamente empregadas neste trabalho, faz-se necessário à familiaridade com alguns dos termos do setor elétrico [7], como segue:

- Carga instalada: soma das potências nominais dos equipamentos elétricos instalados na unidade consumidora, em condições de entrar em funcionamento, expressa em quilowatts (kW).
- Consumidor: pessoa física ou jurídica, ou comunhão de fato ou de direito, legalmente representada, que solicitar a concessionária o fornecimento de energia elétrica e assumir a responsabilidade pelo pagamento das faturas e pelas demais obrigações fixadas em normas e regulamentos da ANEEL, assim vinculando-se aos contratos de fornecimento, de uso e de conexão ou de adesão, conforme cada caso.
- Contrato de adesão: instrumento contratual com cláusulas vinculadas às normas e regulamentos aprovados pela ANEEL, não podendo o conteúdo das mesmas ser modificado pela concessionária ou consumidor, a ser aceito ou rejeitado de forma integral.
- Demanda: média das potências elétricas ativas, solicitadas ao sistema elétrico pela parcela da carga instalada em operação na unidade consumidora, durante um intervalo de tempo especificado, expressa em quilowatts (kW).
- Demanda contratada: demanda de potência ativa a ser obrigatória e continuamente disponibilizada pela concessionária, no ponto de entrega, conforme valor e período de vigência fixados no contrato de fornecimento e que deverá ser integralmente paga, seja ou não utilizada durante o período de

faturamento, expressa em quilowatts (kW).

- Demanda medida: maior demanda de potência ativa, verificada por medição, integralizada no intervalo de 15 (quinze) minutos durante o período de faturamento, expressa em quilowatts (kW).
- Energia elétrica ativa: energia elétrica que pode ser convertida em outra forma de energia, expressa em quilowatts-hora (kWh).
- Energia elétrica reativa: energia elétrica que circula continuamente entre os diversos campos elétricos e magnéticos de um sistema de corrente alternada, sem produzir trabalho, expressa em quilovolt-ampère-reativo-hora (kVarh).
- Estrutura tarifária: conjunto de tarifas aplicáveis às componentes de consumo de energia elétrica e/ou demanda de potência ativas de acordo com a modalidade de fornecimento.
- Estrutura tarifária convencional: estrutura caracterizada pela aplicação de tarifas de consumo de energia elétrica e/ou demanda de potência independentemente das horas de utilização, do dia e dos períodos do ano.
- Estrutura tarifária horo-sazonal: estrutura caracterizada pela aplicação de tarifas diferenciadas de consumo de energia elétrica e de demanda de potência de acordo com as horas de utilização, do dia e dos períodos do ano, conforme especificação a seguir:

    a) Tarifa Azul: modalidade estruturada para aplicação de tarifas diferenciadas de consumo de energia elétrica de acordo com as horas de utilização do dia e os períodos do ano, bem como de tarifas diferenciadas de demanda de potência de acordo com as horas de utilização do dia.
    b) Tarifa Verde: modalidade estruturada para aplicação de tarifas diferenciadas de consumo de energia elétrica de acordo com as horas de utilização do dia e os períodos do ano, bem como de uma única tarifa de demanda de potência.
    c) Horário de Ponta: período definido pela concessionária e composto por 3 (três) horas diárias consecutivas, exceção feita aos sábados, domingos e feriados nacionais, considerando as características do seu sistema elétrico.
    d) Horário Fora de Ponta: período composto pelo conjunto das horas diárias consecutivas e complementares àquelas definidas no horário de ponta.
    e) Período Úmido: período de 5 (cinco) meses consecutivos, compreendendo os fornecimentos abrangidos pelas leituras de dezembro de um ano a abril

do ano seguinte.

f) Período Seco: período de 7 (sete) meses consecutivos, compreendendo os fornecimentos abrangidos pelas leituras de maio a novembro.

- Fator de carga: razão entre a demanda média e a demanda máxima da unidade consumidora, ocorridas no mesmo intervalo de tempo especificado.
- Fator de demanda: razão entre a demanda máxima num intervalo de tempo especificado e a carga instalada na unidade consumidora.
- Fator de potência: razão entre a energia elétrica ativa e a raiz quadrada da soma dos quadrados das energias elétricas ativa e reativa, consumidas num mesmo período especificado.
- Grupo "A": grupamento composto de unidades consumidoras com fornecimento em tensão igual ou superior a 2,3 kV, ou, ainda, atendidas em tensão inferior a 2,3 kV a partir de sistema subterrâneo de distribuição e faturadas neste Grupo nos termos definidos no art. 82, caracterizado pela estruturação tarifária binômia e subdividido nos seguintes subgrupos:

g) Subgrupo A1 - tensão de fornecimento igual ou superior a 230 kV;
h) Subgrupo A2 - tensão de fornecimento de 88 kV a 138 kV;
i) Subgrupo A3 - tensão de fornecimento de 69 kV;
j) Subgrupo A3a - tensão de fornecimento de 30 kV a 44 kV;
k) Subgrupo A4 - tensão de fornecimento de 2,3 kV a 25 kV;
l) Subgrupo AS - tensão de fornecimento inferior a 2,3 kV, atendidas a partir de sistema subterrâneo de distribuição e faturadas neste Grupo em caráter opcional.

- Ponto de entrega: ponto de conexão do sistema elétrico da concessionária com as instalações elétricas da unidade consumidora, caracterizando-se como o limite de responsabilidade do fornecimento.
- Potência: quantidade de energia elétrica solicitada na unidade de tempo, expressa em quilowatts (kW).
- Potência Ativa: que realiza o trabalho propriamente dito, gerando calor, iluminação, movimento, etc., expressa em quilowatts (kW).
- Potência Reativa: que mantém o campo eletromagnético, expresso em quilovolt-ampère-reativo (kVar).

- Unidade consumidora: conjunto de instalações e equipamentos elétricos caracterizado pelo recebimento de energia elétrica em um só ponto de entrega, com medição individualizada e correspondente a um único consumidor.

5.3. Informações dos Consumidores do Grupo “A”

Para obter informações sobre os grandes consumidores de energia, Grupo A, utiliza-se os dados coletados pelos equipamentos de medição da concessionária em cada consumidor, bem como as informações advindas dos contratos firmados entre esses consumidores e a concessionária de energia elétrica.

Os equipamentos de medição fornecem informações da unidade consumidora a cada intervalo de 15 (quinze) minutos, ocorrendo em um mês quase 3000 (três mil) registros. Em alguns casos as informações podem ser registradas a cada intervalo de 5 (cinco) minutos totalizando, aproximadamente, 9000 (nove mil) informações mensais. Para armazenar tais registros utiliza-se “memórias de massa” nos medidores.

A Figura 5.1 apresenta uma tela do *software* utilizado para obtenção das informações das unidades consumidoras que se dá através da leitura da memória de massa dos equipamentos de medição. Dentre as diversas opções de relatórios, gráficos e configurações, apresenta-se na Figura 5.1 um exemplo de listagem de uma memória de massa informando o número do registro, data e hora do registro, número de pulsos do canal 1, 2 e 3, as potências ativas e reativas com o respectivo fator de potência. Obtém-se ainda, através desta listagem, a informação referente aos horários de ponta e fora ponta e horários indutivos e capacitivos.

Portanto, nos medidores podem conter as seguintes informações:

- Relatório de parâmetros:
    - Número de série do equipamento
    - Leitora
    - Modelo
    - Data/Hora da Leitura

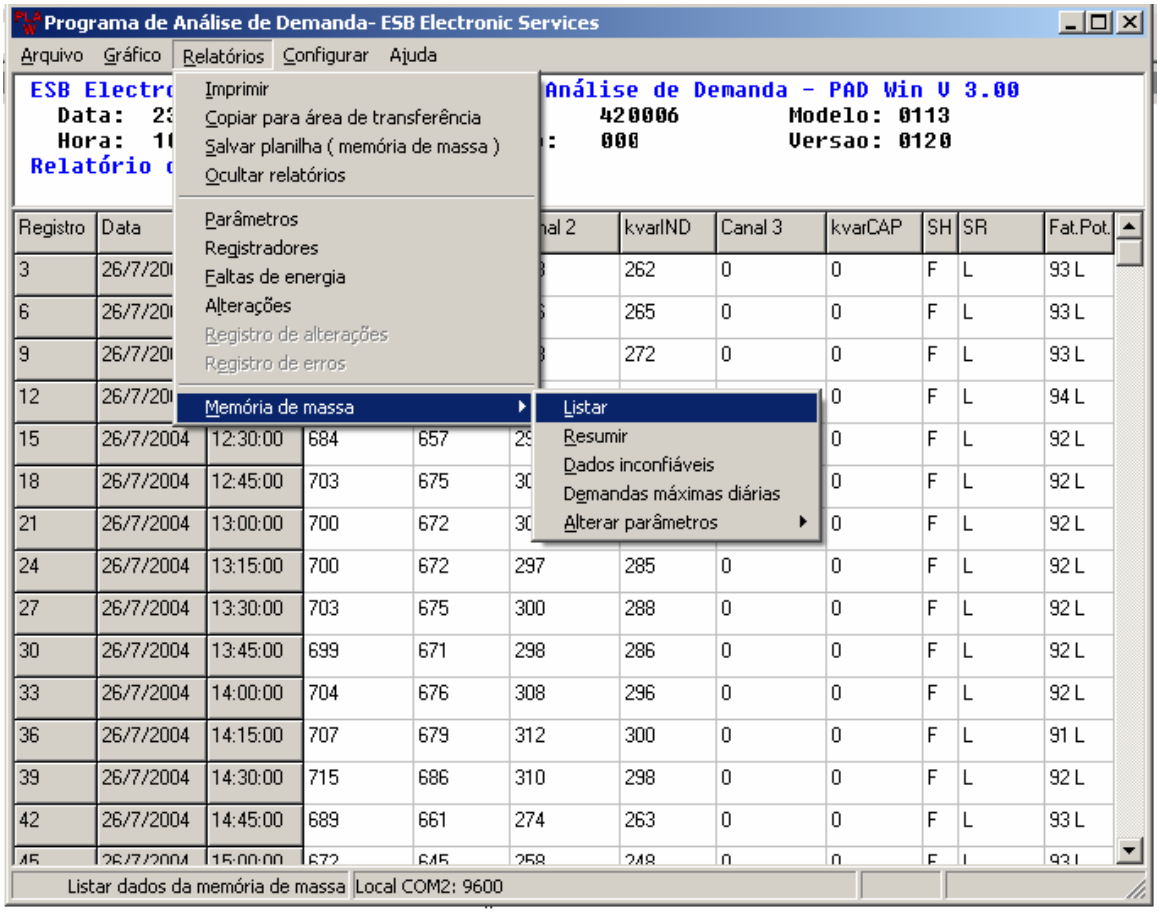


Figura 5.1 – Listagem de uma memória de massa de um consumidor do grupo “A”.

- Último Período de Integração
- Última Fatura
- Penúltima Fatura
- Intervalo de Integração
- Número de Reposições de Demanda
- Tarifa
- Fabricante do RD
- Constantes de Multiplicação dos Canais 1, 2 e 3
- Grandezas dos Canais 1, 2 e 3
- Segmentos Horários de Ponta e Fora Ponta
- Segmentos Horários Indutivo e Capacitivo
- Fator de Potência de Referência
- Feriados Nacionais

- Relatório de Totalizadores:
  - Canais 1, 2 e 3

        - Total Geral
        - Total Ponta Direta
        - Total Fora Ponta Direta
        - Demanda Último Intervalo
        - Demanda Máxima Ponta Direta
        - Demanda Máxima Fora Ponta Direta
        - Demanda Acumulada Ponta Direta
        - Demanda Acumulada Fora Ponta Direta
        - Total Energia Reativa Excedente Ponta
        - Total Energia Reativa Excedente Fora Ponta
        - Demanda Ativa Corrigida para Fator de Potência máxima Ponta
        - Demanda Ativa Corrigida para Fator de Potência máxima Fora Ponta
        - Demanda Ativa Corrigida para Fator de Potência Acumulada Ponta
        - Demanda Ativa Corrigida para Fator de Potência Acumulada Fora Ponta
- Relatório de Quedas de Energia:
    - Data/Hora do início da queda
    - Data/Hora do retorno da queda
    - Duração da queda (dias, horas, minutos, segundos)
- Relatório de Alterações
    - Descrição do Código da Alteração
    - Leitor
    - Data
    - Hora
- Relatório de Memórias de Massa:
    - Número do Registro
    - Data
    - Hora
    - Pulsos Canal 1
    - Pulsos Canal 2
    - Pulsos Canal 3
    - Segmentos Horários de Ponta e Fora Ponta

    - Segmentos Horários Indutivo e Capacitivo
    - Fator de Potência
- Resumo de Memórias de Massa:
    - Totais de Pulsos dos Canais 1, 2 e 3:
        - Geral direto
        - Ponta direta
        - Fora Ponta direta
    - Registro das 3 maiores demandas diárias Geral, Ponta e Fora Ponta
- Relatório das Demandas Máximas Diárias:
    - Canais 1, 2 e 3
        - Data
        - Hora Ponta
        - Demanda registrada na Hora Ponta
        - Hora Fora Ponta
        - Demanda registrada na Hora Fora Ponta

Os consumidores do Grupo “A”, assinam contrato de fornecimento de energia elétrica conforme estrutura tarifária que melhor se enquadrar. A ANEEL estabeleceu três estruturas tarifárias distintas para esse grupo: Horo-Sazonal Azul, Horo-Sazonal Verde e Convencional [7]. Os dados constantes nesses contratos são:

- Nome / Razão Social do Consumidor
- CNPJ / CPF
- Endereço da Sede
- Nome do Representante
- Código de Identificação do Consumidor
- Endereço das Instalações Elétricas
- Modalidade Tarifária
    - Horo-Sazonal Azul
        - Período da Vigência do Contrato
        - Demanda Contratada no Período Seco Ponta
        - Demanda Contratada no Período Seco Fora Ponta
        - Demanda Contratada no Período Úmido Ponta
        - Demanda Contratada no Período Úmido Fora Ponta
        - Tensão Nominal

        - Tensão de Medição
        - Horário de Ponta
    - Horo-Sazonal Verde
        - Período da Vigência do Contrato
        - Demanda Contratada no Período Seco
        - Demanda Contratada no Período Úmido
        - Tensão Nominal
        - Tensão de Medição
        - Horário de Ponta
        - Período de Teste
        - Posto de Transformação
    - Convencional:
        - Período da Vigência do Contrato
        - Demanda Contratada
        - Tensão Nominal
        - Tensão de Medição
        - Período de Teste
        - Posto de Transformação
- Código de Atividade
- Classe / Atividade

5.4. Metodologia Utilizada

O desenvolvimento do sistema para identificação de fraudes ou erros de medição em unidades consumidoras ligadas à alta tensão, tem por objetivo definir um padrão de comportamento para cada consumidor analisado. A metodologia para o sistema ocorre em 6 (seis) etapas: primeiramente a consolidação dos dados em que define-se o banco de dados a ser analisado, seleciona-se os atributos relevantes, indicados por especialistas e efetua-se a subdivisão e limpeza dos dados; na segunda etapa, realiza-se a seleção e pré-processamento dos dados através da geração de novos atributos e discretização das informações; na terceira etapa realiza-se as organizações dos dados; posteriormente realiza-se o processo de mineração de dados aplicando a técnica de Rough Sets para a redução de atributos,

geração de regras, definição de consistência e seleção de regras válidas para o sistema; na quinta etapa ocorre a interpretação e avaliação dos resultados obtidos através da mineração e, finalmente, na sexta etapa, é extraído o conhecimento do sistema de informação. Essas seqüências podem ser observadas na Figura 5.2

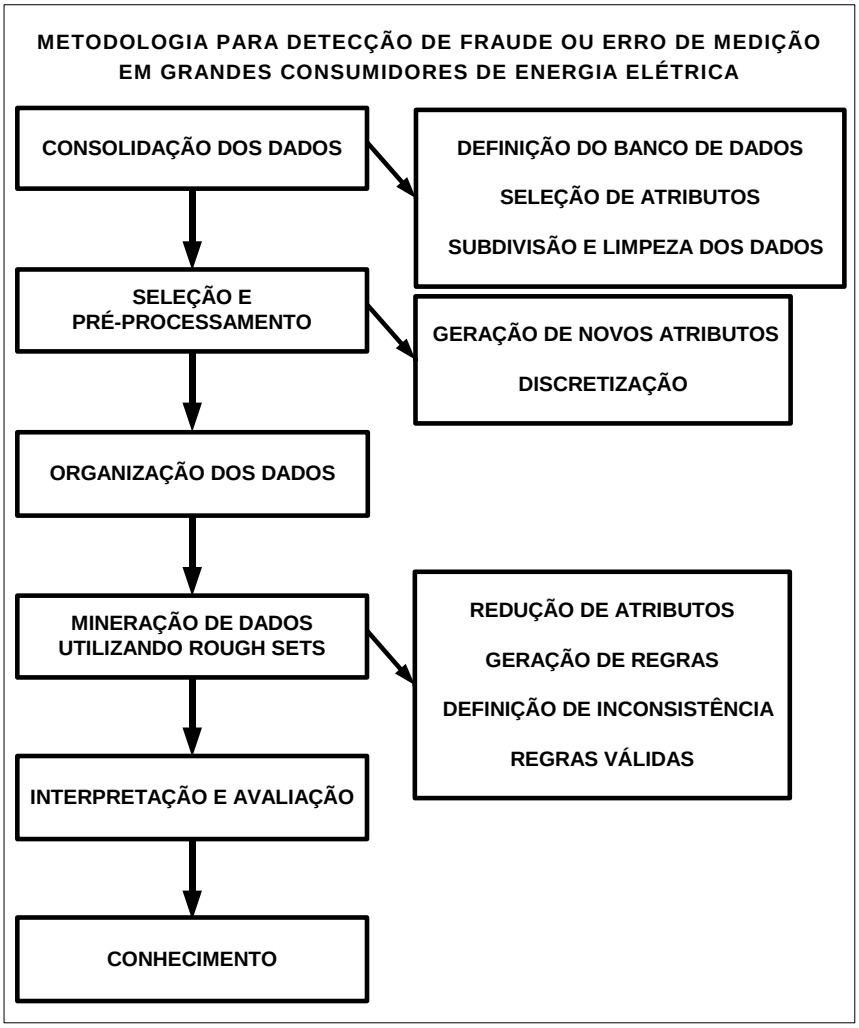


Figura 5.2 – Metodologia de desenvolvimento do trabalho.

Todas as etapas desta metodologia apresentada são descritas, em detalhes, nas seções seguintes.

5.4.1. Consolidação dos Dados

5.4.1.1. Composição do Banco de Dados

O trabalho foi desenvolvido utilizando informações de 27 (vinte e sete) unidades consumidoras, classificadas por atividade da seguinte forma: três no ramo de Alimentos, dez em Comércio, quatro no ramo Industrial e dez Frigoríficos. As informações mensais de cada consumidor diferenciam em quantidade, ou seja, para alguns existem informações referentes a dezessete meses, para outros, apenas três, respeitando, ou não, uma seqüência mensal.

Como em todas as técnicas de desenvolvimento de sistemas baseados em dados, os dados disponíveis foram divididos em dados para treinamento e dados para teste e validação dos resultados. No processo de treinamento está concentrado a maior parte das informações, em torno de 70% dos dados normais e 55% dos anormais, enquanto que para o teste e validação do sistema, admitiu-se o restante das informações.

Os dados selecionados para as duas fases foram escolhidos aleatoriamente, mas com proporção considerável de consumidores normais e anormais. Esta fase é de grande importância, talvez até a mais importante, porque se os dados não estiverem bem representados, as regras não representarão todo o sistema.

5.4.1.2. Seleção de Atributos por Especialistas

Para analisar o comportamento de um consumidor de energia quanto ao seu consumo/demanda, faz-se necessário uma verificação no seu passado, ou seja, buscar informações em seu histórico e, a partir de então, acompanhar seu comportamento em tempo real classificando-o como consumidor com comportamento normal ou anormal.

Na realização deste trabalho, foram selecionados alguns atributos considerados relevantes do ponto de visto de especialistas, classificados como estáticos ou dinâmicos, para serem utilizados na detecção de fraude ou erro de medição. Os atributos estáticos são as informações que não variam no tempo, enquanto que os dinâmicos sofrem modificações constantemente.

Atributos Estáticos

Os atributos estáticos selecionados estão diretamente relacionados com os dados de contratos de fornecimento de energia e algumas constantes referentes ao medidor da unidade consumidora:

- Número do Medidor
- Constante de Multiplicação do Medidor
- Modalidade Tarifária
- Demanda Contratada:
    - Período Seco Ponta
    - Período Seco Fora Ponta
    - Período Úmido Ponta
    - Período Úmido Fora Ponta
- Classe / Atividade

Atributos Dinâmicos

Os atributos dinâmicos estão relacionados com os dados que variam instantaneamente. As informações armazenadas em memórias de massa e selecionadas como atributos dinâmicos a serem considerados no desenvolvimento do sistema são:

- Numero do Registro
- Data
- Hora
- Pulsos Canal 1
- Pulsos Canal 2
- Pulsos Canal 3
- Segmento do Horário de Ponta e Fora Ponta
- Segmento do Horário Indutivo e Capacitivo
- Feriados Nacionais

Atributo de Decisão

O objetivo do atributo de decisão é classificar as unidades consumidoras em normais e anormais. As unidades a serem classificadas como normais serão as que não apresentarem erro de medição nem fraude. Por outro lado, serão anormais as unidades em que se constatar possível fraude ou erro de medição.

5.4.1.3. Subdivisão e Limpeza dos Dados

A concessionária de energia elétrica efetua leituras nas unidades consumidoras em intervalos de, aproximadamente, 30 (trinta) dias, podendo variar de no mínimo 27 (vinte e sete), e o máximo 33 (trinta e três) dias. A memória de massa corresponde a esse período de leitura. Através das informações coletadas na leitura, gera-se o faturamento do consumo e demanda de energia elétrica correspondente para cada consumidor.

Buscando definir um padrão de comportamento de consumo e demanda de energia para o consumidor, após alguns estudos, optou-se por analisar o desempenho da unidade consumidora por dia da semana, traçando um perfil para o domingo, a segunda, a terça e,assim por diante. Na Figura 5.3 pode-se observar a subdivisão das informações contidas nas memórias de massa. Este procedimento é realizado para cada consumidor individualmente.

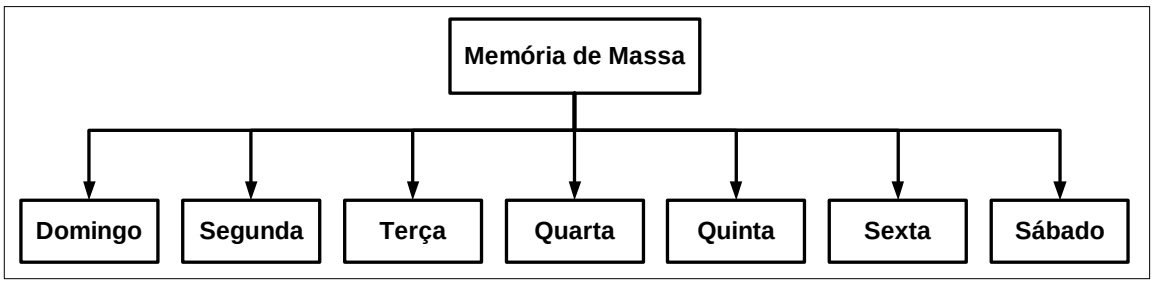


Figura 5.3 – Subdivisão das informações por dia da semana.

O algoritmo a seguir, Quadro 5.1, demonstra o processo de subdivisão do sistema. Considere ATR como sendo o conjunto dos registros de memória de massa, de um consumidor, contendo os seguintes atributos: NR – número do registro da informação; data – data da informação registrada; hora – hora em que a informação foi registrada, Canal_1 – números de pulsos de potência ativa; Canal_2 – número de

pulsos de potência reativa indutiva; Canal_3 – número de pulsos de potência reativa capacitiva, SH – segmentos horários de ponta e fora ponta, SR – segmentos horários indutivo e capacitivo.

```
Entrada: ATR = [NR data hora Canal_1 Canal_2 Canal_3 SH SR];
Saída: Registros armazenados por dia da semana
Início
        domingo = [ ]; segunda = [ ]; terca = [ ]; quarta = [ ]; quinta = [ ];
        sexta = [ ]; sabado = [ ];
        [lin,col] = size(ATR);
        Para i = 1 : lin
                [dia] = weekday (ATR(i,2));
                Se dia = 1
                        domingo = [domingo;ATR(i,:)];
                Fim
                Se dia = 2
                        segunda = [segunda;ATR(i,:)];
                Fim
                Se dia = 3
                        terca = [terca;ATR(i,:)];
                Fim
                Se dia = 4
                        quarta = [quarta;ATR(i,:)];
                Fim
                Se dia = 5
                        quinta = [quinta;ATR(i,:)];
                Fim
                Se dia = 6
                        sexta = [sexta;ATR(i,:)];
                Fim
                Se dia = 7
                        sabado = [sabado;ATR(i,:)];
                Fim
        Fim
Fim
```

Quadro 5.1 – Algoritmo: subdivisão das informações por dia da semana.

Na etapa de limpeza dos dados, levou-se em consideração dois parâmetros considerados importantes: feriados e número de registros diários.

- *Feriados* – em dias de feriados, dependendo da atividade do consumidor, normalmente o consumo de energia diminui. Considerando que a queda de consumo é normal para estes dias específicos, deduz-se que esta diminuição é aceitável. Como o comportamento de consumo em um feriado é muito próximo de um comportamento de fraude, então os dias de feriados são retirados do processo de treinamento e geração de regras para classificação

dos consumidores em normal ou anormal. Por esse motivo, decide-se eliminar os registros referentes a esses feriados. Considerou-se 4 (quatro) feriados móveis (Carnaval, Paixão, Páscoa, Corpus Christi) e feriados fixos (Ano Novo, Tiradentes, Dia do Trabalhador, Independência do Brasil, Finados, Proclamação da República e Natal). Pode-se ainda inserir feriados Estaduais e Municipais para uma melhor filtragem dos dados.

- *Número de registros diários* – a demanda de energia elétrica é dada pelo consumo de energia dividido pelo tempo no qual se verificou tal consumo. Para efeito de faturamento de energia pela concessionária, o intervalo de integração utilizado é de 15 (quinze) minutos. Assim, a demanda de energia é igual ao consumo a cada intervalo de integração dividido 15 minutos. Desta forma, as memórias de massa registram as informações das unidades consumidoras a cada 15 (quinze) minutos. Trabalhar com tantas informações, mesmo que separadas por dias da semana, torna-se um procedimento computacionalmente "caro" e, em um período de tempo tão pequeno do intervalo de integração, não é possível retratar perfis de comportamentos anormais ou normais dos consumidores. Definiu-se então, como estratégia de estudo, trabalhar com as informações diárias integralizadas, ou seja, apenas 1 (um) registro passa a representar os 96 (noventa e seis) registros diários, quantidade necessária para compor informações de 24 horas. Portanto, para ter informações de um período diário completo, são necessários 96 (noventa e seis) registros.

  Com base nas análises feitas nos registros de memória de massa, pode-se observar informações incompletas de período diário. Isso se deve ao fato de não existir um horário fixo para a leitura da memória, podendo ficar incompletos os dados do primeiro e último dia (início e término) do período lido. Ou seja, esses dias não constam das informações das 24 horas ou 96 (noventa e seis) registros. Mesmo reunindo diversas memórias de massa de um período seqüencial de meses de um mesmo consumidor, sempre existirão dias incompletos, sendo necessário excluí-los pois não agregam informações para o sistema de detecção de fraude.

A etapa de limpeza dos dados, ou seja, eliminação de registros em dias de feriados e eliminação dos dados diários incompletos, é realizada após a subdivisão por dias da semana para garantir que todos as informações diárias utilizadas no sistema sejam confiáveis.

5.4.2. Seleção e Pré-Processamento

5.4.2.1. Geração de Novos Atributos

Após a consolidação das informações descritas na subseção 5.4.1, torna-se possível calcular alguns parâmetros, como por exemplo: Fator de Potência, Fator de Demanda, Fator de Carga, Consumo de Energia, Demanda Máxima e Mínima. Estas informações geram alguns resumos diários, cuidadosamente calculados para o dia completo e para os horários de Ponta e Fora de Ponta separadamente e, são definidos como RDC - Resumo Diário Completo, RDP - Resumo Diário Ponta e RDFP - Resumo Diário Fora Ponta respectivamente. Estas informações são consideradas como atributos gerados pois são calculadas pelo próprio sistema computacional desenvolvido com a metodologia proposta, a partir dos dados de memória de massa.

O algoritmo seguinte, Quadro 5.2, descreve a rotina utilizada para o cálculo do resumo diário, referente a um dia da semana, de um período completo, ou seja, informações das 24 (vinte e quatro) horas do dia. As definições dos termos utilizados por esta rotina são denominados como sendo: NM – número do medidor; CT – classe tarifária; MT – modalidade tarifária; DCP DCFP – demanda contratada no período da Ponta e Fora da Ponta, respectivamente; CM – constante do medidor, DS – número referente ao dia da semana analisado (para o domingo o DS = 1); domingo – dados referentes aos domingos na subdivisão do sistema de informação; Cons e Dem – consumo e demanda, respectivamente, para cada período de 15 (quinze) minutos registrados na memória de massa durante o dia analisado; Max_Dem e Med_Dem - máxima e média demanda registrada no dia analisado; Fat_Car – fator de carga referente ao dia em questão; Cons_Int – consumo integralizado do dia, com o objetivo de informar o consumo diário total; RDC_dom – resumo diário completo

referente ao domingo; Dados_dom – dados integralizados por domingo; Label_dom – informações contratuais ou de constantes do consumidor para o domingo.

Quadro 5.2 – Algoritmo: cálculo dos resumos diários.

```
Entrada: NM; CT, MT, DCP, DCFP, CM, DS, domingo.
Saída:   RDC_dom, Dados_dom, Label_dom.
Início
        ATRD = [NM  CT  MT  DCP  DCFP  CM];
[Ldom,Cdom]=size(domingo);
        Dados_dom = [ ];
        Label_dom = [ ];
        RDC_dom = [ ];
        Pos = find(domingo(:,3) = 0);
        Se length(Pos)>1
                Para i =1:(length(Pos)-1)
                        A = [ ];
                        A = domingo(Pos(i):(Pos(i+1)-1),:);
                        re = find (domingo(Pos(i) : (Pos(i+1)-1), 2) = domingo
(Pos(i), 2));
                        A = A (re , :);
                        Cons = A(:,4)* CM;
                        Dem = Cons*4;
                        Max_Dem = max (Dem);
                        Med_Dem = mean (Dem);
                        Fat_Car = Med_Dem / Max_Dem;
                        Cons_Int = sum(Cons);
                        RDC_dom = [RDC_dom; Cons_Int Max_Dem
Fat_Carga];
                        Dados_dom = [ Dados_dom; A(: ,4 )' ];
                        Label_dom = [ Label_dom; ATRD(1) domingo(Pos(i),2)
DS ATRD(2:6) ];
                Fim
        Fim
Fim
```

Esta rotina deve ser realizada para todos os dias da semana de domingo à sábado e, para cada consumidor individualmente.

O mesmo procedimento utilizado para calcular o RDC, é utilizado para o cálculo do RDP e RDFP, levando-se em consideração os horários definidos em contratos, para cada consumidor, para estes respectivos períodos. Para RDP e RDFP, existe uma quantidade maior de atributos quando comparados com o RDC e, isto se deve ao fato de alguns cálculos serem considerados apenas para os horários de ponta e fora ponta, separadamente.

A partir dos resumos calculados, RDC, RDP e RDFP, executa-se o processo de discretização, como será mostrado no próximo tópico.

5.4.2.2. Discretização

A abordagem de Rough Sets não trabalha com valores contínuos e sim com valores discretos. Ou seja, para atributos dinâmicos é necessário aplicar um processo de discretização a fim de possibilitar a utilização da técnica. Portanto, esta etapa da metodologia prepara os dados para serem aplicados em Rough Sets.

Os dados dinâmicos diários desse trabalho são discretizados utilizando estatística descritiva. Para todos os atributos dos resumos obtidos através do algoritmo apresentado no Quadro 5.2, define-se um intervalo de variação, denotado por Δ, e indicado como sendo:

$$\Delta(i,j) = \overline{X}(i,j) - \sigma(i,j) \tag{5.1}$$

sendo que i refere-se aos resumos RDG, RDP, RDFP; j refere-se aos atributos dos resumos considerados; $\overline{X}(i,j)$ e $\sigma(i,j)$ representam a média e desvio padrão, respectivamente, do resumo i e atributo j. Este procedimento ocorre para todos os 7 (sete) dias da semana.

Em uma análise diária, avalia-se o comportamento do consumidor através dos intervalos das variações previamente estabelecidos. Para o atributo j analisado, referente ao resumo i, que apresentar uma variação acima do $\Delta(i,j)$ estabelecido, atribui-se o valor de 1 (um) ao atributo “analise”, caso contrário atribui-se o valor 2 (dois). Este procedimento pode ser representado por uma regra “se / então” conforme algoritmo do Quadro 5.3.

Quadro 5.3 – Algoritmo: criação do atributo “Análise”.

```
Se Δ(i,j)calculado  >  Δ(i,j)estabalecido
        Analise(k) = 1
Se não
        Analise(k) = 2
Fim
```

sendo k a quantidade de análises devido aos atributos de cada resumo (RDG, RDP, RDFP).

5.4.3. Organização dos Dados Transformados

O sistema de decisão $\Re$, previamente definido na equação (3.2) como sendo $\Re = (U, C \cup D)$ na abordagem de Rough Sets, consolidou-se da seguinte forma, na metodologia proposta:

- $U =$ {registros das unidades consumidoras a cada 15 (quinze) minutos}

  O conjunto de objetos formado por todas as informações registradas na memória de massa.

- $C =$ {atributos estáticos e dinâmicos}

  Atributos Estáticos. Considerou-se seis atributos estáticos, sendo que dois são os atributos Classe Tarifária e Modalidade Tarifária, cada qual com seus respectivos valores nominais. Já os atributos remanescentes, referem-se à alteração contratual quanto aos atributos Classe Tarifária, Modalidade, Demanda Contratada na Ponta ou Fora Ponta, Constante do Medidor. Os valores nominais admitidos para estes quatro atributos será 1 (um) se não houver mudança nos valores contratuais para os respectivos atributos, ou 2 (dois) caso haja alteração.

  Atributos Dinâmicos. Os atributos dinâmicos abordados são: Dia da Semana, Consumo Geral, Demanda Máxima Geral, Fator de Carga Geral, Consumo na Ponta e Fora da Ponta, Demanda Máxima na Ponta e Fora da Ponta, Demanda Mínima na Ponta e Fora Ponta, Fator de Carga na Ponta e Fora Ponta, Fator de Demanda na Ponta e Fora Ponta, Máximo Fator de Potência na Ponta e Fora Ponta, Mínimo Fator de Potencia na Ponta e Fora da Ponta.

  O número total de Atributos Condicionais considerados são 24 (vinte e quatro).

- $D =$ {estado da unidade consumidora}

Os valores nominais admitidos para o conjunto D são: normal ou anormal.

5.4.4. Mineração de Dados Usando Rough Sets

A teoria de Rough Sets é utilizada na metodologia de detecção de fraude ou erro de medição, com o objetivo de eliminar atributos desnecessários para a classificação do sistema e gerar regras. Algumas regras geradas apresentam incertezas quanto a classificações dos dados. Assim, o usuário define o valor de consistência mínima a ser permitido para as regras, selecionando apenas as regras que satisfaça a consistência pré-definida para serem utilizadas na classificação dos dados. Portanto, o processo de mineração de dados aplicando a teoria de Rough Sets está compreendido em 3 (três) fases: redução do sistema de informação, geração de regras e, seleção de regras válidas para a classificação dos dados a partir da consistência permitida pelo usuário.

5.4.4.1. Redução do sistema de Decisão

A fase de redução do sistema de decisão tem como objetivo selecionar os atributos relevantes para a geração de regras.

O primeiro procedimento para a redução do sistema de decisão $\Re$ é agrupar todos os registros que apresentam os mesmos valores nominais para C e D. A seguir, apresenta-se a rotina utilizada para realizar este agrupamento, Quadro 5.4.

Quadro 5.4 – Algoritmo: redução do sistema de decisão $\Re$.

```
Entrada: ℜ = [C D]
Saída:   ℜ
Início
        V = [ ];
        [r1,r2] = size( ℜ );
        ℜ = [ℜ  ones(r1,1)];
        Para j = 1 : (r2 -1)
                Para i = (j + 1) : r2
                        Se ℜ (j,(1:r2)) ≠ 0  e  ℜ (j,(1:r2)) = ℜ (i,(1:r2))
                                ℜ (j,(r2+1)) = ℜ (j,r2) + ℜ (i,r2);
                                V = [ V,i ];
```

```
                                        ℜ (i,:) = 0;
                                Fim
                        Fim
                Fim
                ℜ (V,:) = [ ];
        Fim
```

Nesta rotina apresentada, incrementou-se uma coluna na matriz formada pelo sistema de decisão com o objetivo de informar a quantidade de registros totalmente iguais existentes em $\Re$.

O segundo procedimento para a redução do sistema de decisão é determinar a relação de indiscernibilidade $IND_{\Re}(C)$, definida pela equação (3.3). Neste procedimento, todos os registros de $\Re$, que apresentarem os mesmos valores nominais apenas para os atributos condicionais, são agrupados formando uma classe. Na rotina apresentada anterior, agrupam-se todos os registros com os mesmos valores nominais para os conjuntos de atributos C e D. Portanto, havendo registros iguais neste segundo procedimento, isso implica automaticamente que os valores de decisão são diferentes, ou seja, são indiscerníveis entre si. No Quadro 5.5, tem-se o algoritmo usado para determinar $IND_{\Re}(C)$.

Quadro 5.5 – Algoritmo : relação de indiscernibilidade $IND_{\Re}(C)$.

```
        Entrada: ℜ
        Saída:   IND_ℜC , Classes
        Início
                V = [ ];
                Classes = [ ];
                [r1,r2] = size( ℜ );
                IND_ℜC = [ℜ  ones(r1,1)];
                [r1,r2] = size( IND_ℜC );
                Para j = 1 : (r2 -1)
                        Para i = (j + 1) : r2
                                Se  IND_ℜC (j,(1:r2-3)) ≠ 0   e   IND_ℜC (j,(1:r2-3))= IND_ℜC (i,
(1:r2-3))
                                        IND_ℜC (j,r2) = IND_ℜC (j,(r2-1)) + IND_ℜC (i,(r2-1));
                                        Classes (j,j)=j;
                                        Classes (j,i)=i;
                                        V = [ V,i ];
                                        IND_ℜC (i,:) = 0;
                                Se não
                                        Classes (j,j)=j;
                                Fim
```

```
                Fim
            Fim
            IND_ℜC (V,:) = [ ];
        Fim
```

A matriz denominada de *Classes* no algoritmo acima, indica quais os registros pertencentes a uma mesma classe.

Definir a relação de indiscernibilidade do sistema de informação é a principal etapa antes de iniciar o processo de redução desse mesmo sistema, pois a $IND_{\Re}(C)$ deve ser mantida para o conjunto de atributos mínimos encontrado, o reduto.

Na teoria apresentada nos capítulos 3 e 4, obteve-se a redução da informação ao encontrar os redutos a partir da matriz de discernibilidade e função de discernibilidade, que utiliza simplificações de expressões booleanas através de seus teoremas, propriedades e postulados.

A utilização da função de discernibilidade pode não ser a melhor forma para encontrar redutos devido ao tamanho do banco de dados. Uma alternativa é utilizar outros métodos, por exemplo: algoritmos genéticos, heurísticas, ou outras formas válidas para a simplificação de função, que possam ser utilizadas na substituição da função discernibilidade apresentada pela teoria de Rough Sets.

Há uma ferramenta desenvolvida chamada de Rosetta (*A Rough Sets Toolkit for Analysis of Data*), desenvolvida e implementada em [26], podendo auxiliar no processo de DCBD, pois possibilita realizar operações de visualização, pré-processamento (eliminação de atributos, preenchimento de valores nulos, etc.), redução de informação e geração de regras na forma $< \text{Se} \ldots \text{então} \ldots >$. Também faz validação e análise das regras obtidas. Este *software* é gratuito e disponível para *download*[2], tendo sido utilizado em alguns trabalhos científicos e citado como um sistema que apresenta bons resultados na redução do sistema de informação.

Neste trabalho, utilizou-se a função de discernimento para encontrar o reduto, uma vez que os registros iguais encontrados em $\Re$ foram agrupados, reduzindo significativamente o tamanho da matriz em relação às linhas.

[2] www.idi.ntnu.no/~aleks/rosetta/

5.4.4.2. Geração das Regras

Nesta fase da aplicação da teoria de Rough Sets, geram-se as regras de classificação para definir a normalidade ou anormalidade das unidades consumidoras à partir da redução do sistema de informação.

Devido à existência de indiscernibilidade no sistema, é conveniente gerar regras com a prioridade de classificação de anormalidade, pois o objetivo é identificar fraudes e erro de medição em grandes clientes consumidores de energia elétrica. Por exemplo, na Tabela 5.1, tem-se uma situação de indiscernibilidade em que uma única regra conduz a decisões diferentes.

Tabela 5.1 – Regra indiscernível.

Regra (*R*)	*Atributos Condicionais Reduzidos*										*Número de registros*	
	b1	*b7*	*b8*	*b9*	*b10*	*b11*	*b12*	*b13*	*b14*	*b15*	*D = anormal*	*D = normal*
r	1	1	2	2	2	2	1	2	1	2	18	15

Para a regra r tem-se 18 (dezoito) registros com decisão anormal e 15 (quinze) registros com decisão normal. Neste caso, definindo a prioridade de decisão para as regras como sendo $D = anormal$, na classificação dos 33 (trinta e três) registros referentes à r, classificam-se corretamente 18 (dezoito) e incorretamente o restante dos registros. Para obter-se a informação do quanto à regra r classifica corretamente os registros considerando $D = anormal$, calcula-se a o valor da consistência δ de r, conforme o algoritmo apresentado no Quadro 5.6. Para que uma regra seja consistente não pode haver situação de indiscernibilidade, ou seja, a regra deve conduzir a uma única decisão e a consistência torna-se unitária $(\delta = 1)$.

Seja R o conjunto de regras extraído do banco de dados reduzido. No Quadro 5.6, o algoritmo calcula R_δ que é o conjunto de R acrescido do valor de δ.

Priorizar a decisão da regra em anormal não é uma obrigatoriedade, pode-se avaliar primeiramente os resultados e definir qual o conjunto de regras que melhor classifica o sistema e apresenta menor taxa de erro quanto à classificação incorreta de registros, regras com decisão “normal” ou regras com decisão “anormal”. Com isto, a flexibilidade quanto ao valor nominal do atributo de decisão que irá definir as aproximações deve existir.

Quadro 5.6 – Algoritmo: cálculo do valor da consistência δ para cada regra r .

```
Entrada: R
Saída:   R_δ
Início
        δ = [ ];
        [r1,r2] = size(R );
        Para i = 1 : r1
                δ (i)= R (i,( r2-1)) / R (i,( r2-1))+ R (i, r2)
        Fim
        R_δ = [R  δ ];
Fim
```

5.4.4.3. Definição de Consistência e Seleção de Regras Válidas

Nesta fase, o objetivo é definir o valor do parâmetro consistência para selecionar as regras inconsistentes que serão utilizadas na classificação do sistema. O algoritmo para a seleção de regras válidas denominado de $R_\delta valida$, é descrito a seguir, sendo x o valor da consistência admitida.

Quadro 5.7 – Algoritmo: seleção de regras válidas - $R_\delta valida$.

```
Entrada: R_δ , x
Saída:   R_δvalida
Início
        R_δvalida = [ ];
        [r1,r2] = size(R_δ );
        Para i = 1 : r1
                Se R_δ (i, r2) >= x
                        R_δvalida = [R_δvalida ; R_δ (i,:)];
                Fim
        Fim
Fim
```

Para definir qual o valor ideal da consistência, devem ser realizados vários testes com diferentes valores, a fim de encontrar uma melhor classificação de fraude ou erro de medição em unidades consumidores de energia elétrica.

5.4.5. Interpretação e Avaliação

Depois da definição do conjunto de regras para classificação dos dados na fase de treinamento do sistema, analisa-se os resultados quanto à correta classificação dos consumidores de energia selecionados para a etapa de Teste e Validação.

Os dados utilizados para o teste compreendem informações de unidades consumidoras com comportamento normal e anormal. Para validar as regras definidas na seção 5.4.4.3 é necessário que estas apresentem bons resultados quanto à classificação correta das unidades consumidoras.

Os resultados são apresentados em forma de relatório, contendo as seguintes informações:

- Atributo de Decisão Considerado na Geração das Regras
- Consistência Admitida
- Total de Registros Analisados na Base de Dados
- Total de Registros "Normais" na Base de Dados
- Total de Registros "Normais" "identificados"
- Total de Registros "Normais" "não identificados"
- Total de Registros “Anormais" na Base de Dados
- Total de Registros “Anormais” "identificados"
- Total de Registros “Anormais” "não identificados"

As informações descritas acima apresentam apenas o total de registros classificados ou não corretamente. Pode ocorrer um grande número de registros classificados incorretamente, mas que estejam relacionados com poucos consumidores. Por exemplo, suponhamos que existam 40 (quarenta) registros classificados incorretamente. Estes registros estão relacionados com um mesmo consumidor e para quatro dias da semana, ou seja, apresentou todos estes erros de classificação, mas isto não quer dizer que não tenha encontrado a anormalidade do consumidor. Portanto, gerou-se a opção de visualizar a classificação dos registros por unidade consumidora. Através desta visualização podemos ter informações de:

- Classificação de "Normal" Corretamente
- Classificação de "Anormal" Corretamente
- Classificação de "Normal" Indevidamente

- Classificação de "Anormal" Indevidamente

Os resultados das classificações têm as seguintes informações: número da unidade consumidora e número de registros para cada dia da semana por unidade. A Tabela 5.2 apresenta o resultado de classificação para o exemplo citado acima, sendo que o somatório de todos os registros dos dias da semana é igual a quantidade de registros classificados incorretamente.

Tabela 5.2 – Classificações dos registros por unidade consumidora.

	Dias da Semana						
Unidade Consumidora	*Domingo*	*Segunda*	*Terça*	*Quarta*	*Quinta*	*Sexta*	*Sábado*
UC		**10**	**10**	**10**		**10**	

A partir dos resultados encontrados nas classificações dos registros, é que se analisa a influência dos valores da consistência, em que se busca sempre um resultado satisfatório quanto a classificação das unidades consumidoras com fraude ou erro de medição como anormais e com uma taxa de erro aceitável.

5.5. Considerações Finais

Neste capítulo apresentou-se algumas das terminologias do setor elétrico que são utilizadas no trabalho para uma melhor compreensão. Levantou-se também as principais informações geradas por contratos e medições de energia referentes às unidades consumidoras do Grupo “A”, consumidores atendidos à alta tensão. E por fim, apresentou-se a metodologia utilizada no desenvolvimento do trabalho para a detecção de fraude ou erro de medição em grandes clientes consumidores de energia elétrica, baseado em dados históricos e dados de tempo real recebidos periodicamente pela telemedição, utilizando a teoria de Rough Sets.

No próximo capítulo serão apresentados alguns resultados da aplicação da metodologia apresentada através de estudo de caso. Será apresentada a comparação quanto aos resultados obtidos através das variações dos valores do Atributo de Decisão para a geração de regras e, o valor da consistência admitida para essas regras. Portanto, a eficiência da metodologia será testada no próximo capítulo.

Capítulo 6
Estudo de Caso

6.1. Introdução

O objetivo deste capítulo é apresentar alguns resultados quanto ao treinamento e teste do sistema computacional desenvolvido com a metodologia proposta no Capítulo 5. Para identificar o tipo de comportamento dos grandes clientes consumidores de energia elétrica, estes são classificados como normal e anormal. Logo, os clientes com comportamento anormal indicam possibilidades de fraude ou erro de medição e são selecionados para inspeções técnicas, finalizando assim o processo de identificação de anormalidades.

6.2. Aquisição dos Dados

Neste trabalho, para definir perfis de comportamentos diários de unidades consumidoras de energia elétrica, foram necessárias informações históricas dos consumidores. O conjunto de dados utilizados no desenvolvimento da metodologia proposta conforme mencionado na seção 5.2.1 do capítulo 5, é oriundo de uma concessionária de energia, e estão compreendidos em informações de comportamentos normais, ou seja, sem fraude ou erro de medição e, informações de fraude de algumas unidades consumidoras de energia. As informações de fraudadores obtidas através do banco de dados cedido não foram suficientes para traçar tais perfis, por não terem sido fornecidos dados históricos do período que antecedeu ou sucedeu a fraude. As informações de memória de massa, do período da fraude, fornecidas pela concessionária foram utilizadas para analisar os efeitos provocados por tais fraudes com relação, principalmente, as potências ativa e reativa. A partir das observações encontradas, simulou-se alguns casos de fraudes em consumidores com históricos normais para a execução desse trabalho. Portanto, os

dados de comportamento normal são reais e os dados de comportamento anormal foram criados. Os dados referentes à identificação dos consumidores não foram fornecidos para garantir o sigilo sobre a identidade dos mesmos.

Utilizou-se o software PLA_Win (Programa de Leitura e Análise para Ambiente Windows) para leitura e análise dos arquivos de memória de massa dos medidores de energia elétrica, analisados nesse trabalho. O PLA_Win é um aplicativo que executa as tarefas de carga de programa, leituras e parametrização nos medidores SAGA1000[3] e em outros medidores compatíveis. Este aplicativo roda em microcomputadores com arquitetura compatível com IBM-PC sob o sistema operacional Windows95 ou superior.

Utilizou-se, também, do *Software Matlab - The Language of Technical Computing*, versão 7.0 *Release* 14, para a implementação das rotinas desenvolvidas para o sistema de detecção de fraude ou erro de medição em grandes clientes consumidores de energia elétrica.

Conjunto de Treinamento e Teste

Contabilizando todas as informações dos consumidores utilizados no desenvolvimento da metodologia, obtém-se um total de 1112832 registros sendo 997.536 registros de consumidores de comportamento normal e 115296 registros de anormalidades. Os registros de anormalidades foram “criados” para o desenvolvimento do trabalho devido à falta de dados de fraudadores no banco de dados analisado. Analisou-se, inicialmente, para a criação destes dados alguns casos verdadeiros de fraude ocorridas em unidades consumidoras de energia, e através da análise de como estas fraudes interferiram nos dados de memórias de massa, desenvolveu-se, posteriormente, os dados fraudulentos. As fraudes criadas foram do tipo: “alteração da constante do medidor”, “desligamento de uma fase no medidor”, “desligamento total de energia no medidor nos finais de semana”, “desligamento total de energia no medidor no horário de ponta”. Essas informações foram divididas em dados para treinamento e dados para teste e validação de resultados, conforme Tabela 6.1.

[3] Medidores eletrônicos desenvolvidos pela ESB Eletronic Services.

Tabela 6.1 – Divisão do conjunto de dados

	Registro	
Conjunto de Dados	**NORMAL**	**ANORMAL**
TREINAMENTO	716256	62784
TESTE	281280	52512

Para os conjuntos de dados obtidos através da divisão mencionada na Tabela 6.1, realizou-se o processo de subdivisão por dia da semana, eliminação de feriados nacionais do conjunto de dados de treinamento e limpeza de informações diárias incompletas. Como optou-se trabalhar com informações diárias, para os 96 (noventa e seis) registros diários, considerando um intervalo de 15 (quinze) minutos, passam a ser representados por apenas 1 (um) registro com informação integralizada do dia, obtendo-se, assim, os dados consolidados conforme Tabela 6.2.

Tabela 6.2 – Conjunto de dados consolidados.

	Registro	
Conjunto de Dados	**NORMAL**	**ANORMAL**
TREINAMENTO	7.201	631
TESTE	2.930	547

6.3. Processo de Treinamento

Alguns atributos estáticos, quando utilizados como atributos condicionais do sistema de informação, apresentam um aumento de regras de classificação e não melhora o resultado quanto à classificação incorreta de consumidores em normal ou anormal. Por esta razão, os atributos “Modalidade Tarifária” e “Classe/Atividade”, foram removidos dos atributos condicionais do sistema de informação. Os atributos dinâmicos referentes ao Horário de Ponta foram desprezados para este sistema, pois no período, intervalo entre as 17 e 20 horas, os consumidores podem buscar soluções alternativas para redução de consumo, devido o preço diferenciado da tarifa para os contratos horo-sazonais.

Logo, no processo de treinamento, dentre os atributos previamente analisados para a detecção de fraude, selecionou-se um total de 15 (quinze) atributos condicionais e 1 (um) atributo de decisão para a geração de regras, com o objetivo de classificar o comportamento das unidades consumidoras em normal e anormal.

Desta forma, definiu-se:

- Conjunto de Registros: $U =$ {registro de parâmetros diários das unidades consumidoras}
- Conjunto de Atributos Condicionais: $B = \{b1, b2, b3, b4, b5, b6, \cdots, b15\}$
- Conjunto de Atributos de Decisão: $D = \{d1, d2\}$

Portanto, o sistema de decisão a ser utilizado pode ser representado por $\Re = (U, B \cup D)$.

Agrupamento de Registros Iguais

Os registros ou dados com os mesmos valores nominais para os atributos condicionais B e de decisão D, encontrados no conjunto de dados de treinamento, foram agrupados com a finalidade de reduzir o esforço computacional para encontrar o reduto do sistema de informação. Este procedimento totalizou 402 registros diferentes para um total de 7.832 dados.

Relação de Indiscernibilidade

Após o processo de agrupamento dos dados, definiu-se a relação de indiscernibilidade $IND_{\Re}(B)$ existente entre os registros, resultando em classes discerníveis e indiscerníveis. Resumidamente, o total de classes obtidas para os dados de treinamento pode ser visto na Tabela 6.3.

Tabela 6.3 – Relação de indiscernibilidade do conjunto de dados de treinamento.

Classes	Total
Discerníveis	204
Indiscerníveis	99

Qualidade das Aproximações

Considerando o valor nominal do atributo de decisão igual a “normal”, pode-se analisar a qualidade das aproximações do conjunto de classes formado por esse atributo, denotado por X, como sendo:

- $\alpha_B(X) = \frac{|1130|}{|7748|} = 0.1458$, ou seja, 14.58% das classes pertencentes ao conjunto X são precisas.
- $\alpha_B(\overline{B}(X)) = \frac{|7748|}{|7832|} = 0.9893$, ou seja, 98.93% do total de classes possivelmente pertencem a X.
- $\alpha_B(\underline{B}(X)) = \frac{|1130|}{|7832|} = 0.1443$, ou seja, 14.43% do total de classes certamente pertencem a X.

Logo, considerando o atributo de decisão igual a "anormal", a qualidade das aproximações do conjunto de classes formado por esse atributo, denotado por Y, apresenta os seguintes resultados:

- $\alpha_B(Y) = \frac{|84|}{|6702|} = 0.0125$, ou seja, 1.25% das classes pertencentes ao conjunto Y são precisas.
- $\alpha_B(\overline{B}(Y)) = \frac{|6702|}{|7832|} = 0.8557$, ou seja, 85.57% dos registros possivelmente pertencem a Y.
- $\alpha_B(\underline{B}(Y)) = \frac{|84|}{|7832|} = 0.0107$, ou seja, 1.07% do total dos registros certamente pertencem a Y.

Portanto, o conjunto de classes formado pelo valor nominal do atributo de decisão igual a "normal", obtém o maior número de regras consistência.

Reduto

Encontrou-se 2 (dois) redutos para o sistema de informação original B_1^* e B_2^*, sendo que para o conjunto de atributos mínimos B_1^*, foram considerados irrelevantes os atributos condicionais $b2$, $b3$, $b4$, $b5$ e $b6$, referente às alterações contratuais da classe tarifária, modalidade tarifária e demandas contratadas, alteração da

constante do medidor e máxima demanda encontrada no período fora de ponta, respectivamente. Para o segundo conjunto de atributos mínimos B_2^*, os atributos condicionais eliminados foram $b2$, $b3$, $b4$, $b5$ e $b9$, sendo o atributo $b9$ referente ao fator de demanda para o período fora de ponta. Portanto, os dois possíveis redutos são:

$$B_1^* = \{b1, b7, b8, b9, b10, b11, b12, b13, b14, b15\}$$

$$B_2^* = \{b1, b6, b7, b8, b10, b11, b12, b13, b14, b15\}$$

Geração de Regras

Considerando os atributos de B_1^*, 303 (trezentos e três) regras gerais foram totalizadas para o sistema de informação reduzido, sendo:

- Total de regras consistentes: 204;
- Total de regras inconsistentes: 99.

Devido à existência de apenas duas possibilidades de decisão no processo de classificação, não se fez necessário a utilização de todas as regras descritas acima. Definindo um dos dois valores normais para a decisão (normal, anormal), utilizou-se apenas as regras referentes à decisão selecionada e, para os registros que não classificarem com uma das regras desse conjunto, atribuiu-se o valor de classificação oposto ao da decisão considerada. Assim, reduz-se a quantidade de regras para o processo de classificação. Entretanto, a busca de um conjunto de regras para a classificação dos consumidores se dá através da variação de valores de dois parâmetros: valor nominal do atributo de decisão para definição das regras a serem utilizadas na classificação e, valor numérico de consistência aceitável para essas regras.

O conjunto de treinamento contém informações de 27 (vinte e sete) clientes, totalizando 7832 (sete mil oitocentos e trinta e dois) dados, sendo 7201 (sete mil duzentos e um) registros normais e 631 (seiscentos e trinta e um) registros anormais.

Logo, simulou-se, para cada valor nominal do conjunto D, 4 (quatro) variações para o valor da consistência. Os resultados quanto à classificação dos registros pelas regras consideradas podem ser observadas nos tópicos a seguir.

Utilizou-se dois métodos de avaliação para os resultados de classificação: análise diária e semanal.

Análise Diária. Na análise diária, para cada registro diário classificado como anormal, gera-se uma inspeção técnica para a unidade referente ao registro.

Análise Semanal. Na análise semanal, o objetivo é reduzir a quantidade de classificações incorretas de anormalidades. Seguindo a idéia de que um fraudador apresenta irregularidades constantemente, para que uma unidade seja selecionada para inspeção técnica é preciso apresentar mais de um registro classificado com anormalidade num intervalo de 7 (sete) dias. Havendo mais de um registro neste período de uma semana, então se realiza uma inspeção no consumidor classificado como "anormal".

- Atributo de Decisão = Normal

Consistência = 0.3

Considerando $\delta = 0.3$ para as regras de classificação, obtém-se um total de:

- Regras consistentes: 185
- Regras inconscientes: 97

A quantidade dos registros diários classificados como normal e anormal para a consistência considerada, encontra-se na Tabela 6.4.

Tabela 6.4 – Classificação de dados de treinamento para $\delta = 0.3$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	7196	5
	Anormal	533	98

Considerando o conjunto de registros normais do processo de treinamento, ou seja 7201 (sete mil duzentos e um) registros, obteve-se a classificação correta de 99.93% dos dados. Para o conjunto formado pelos 631 (seiscentos e trinta e um) registros anormais utilizados no mesmo processo, apenas 15.5% destes dados foram classificados corretamente.

As análises dos resultados da classificação dos consumidores foram realizadas levando-se em consideração os registros classificados como anormais, uma vez que o propósito do trabalho é detectar fraude ou erro de medição em unidades consumidoras de energia elétrica. Assim, realizaram-se duas análises: diária e semanal.

Análise Diária. Dos 98 (noventa e oito) registros anormais classificados corretamente, pode-se observar na Tabela 6.5, que foram encontrados 10 (dez) dos 12 (doze) consumidores fraudadores pertencentes ao conjunto de dados de treinamento. Os 5 (cinco) registros classificados incorretamente como anormal, referem-se a 2 (dois) consumidores.

Tabela 6.5 – Identificação dos consumidores anormais - análise diária.
Dados de treinamento - ($\delta = 0.3$ e $D = \{normal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	3A	0	0	0	0	1	0	0	1
	2A	4	0	0	0	2	0	0	6
	2B	0	0	0	0	2	0	0	2
	4A	0	0	0	1	0	0	0	1
	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2D	0	0	0	0	1	0	0	1
	2E	0	1	0	0	2	0	0	3
	3I	0	2	0	0	0	0	0	2
	2I	7	3	0	0	2	0	0	12
Incorreta	4B	0	0	0	0	1	0	0	1
	2D	4	0	0	0	0	0	0	4

A Tabela 6.5 apresenta informações da quantidade de consumidores classificados correta e incorretamente como anormal em uma análise diária, separando por dias da semana a quantidade de registros classificados para cada consumidor. Considerando-se que as inspeções técnicas em consumidores classificados como anormal são realizadas a partir das

informações diárias, tem-se, portanto, um total de 12 (doze) inspeções, das quais 83.3% apresentam resultado satisfatório quanto à detecção de fraude ou erro de medição na análise diária.

Análise Semanal. Para $\delta = 0.3$, selecionou-se 5 (cinco) consumidores para inspeção, conforme Tabela 6.6. Através dessa análise, encontrou-se 4 (quatro) fraudadores de um conjunto de 12 (doze) e, obteve-se 80% de sucesso quanto a taxa de acerto na identificação de anormalidade.

Tabela 6.6 – Identificação dos consumidores anormais - análise semanal. Dados de treinamento - ($\delta = 0.3$ e $D = \{normal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	4	0	0	0	2	0	0	6
	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2I	7	3	0	0	2	0	0	12
Incorreta	2D	4	0	0	0	0	0	0	4

Com base nos resultados das análises realizadas para $\delta = 0.3$, pode-se concluir que a quantidade de fraudadores identificados foi bem maior na análise diária, sendo encontrados 10 (dez) clientes contra 4 (quatro) detectados através da análise semanal. Quanto ao sucesso das inspeções realizadas, a análise diária apresentou um melhor desempenho, uma vez que obteve uma taxa de acerto de 83.3%, sendo inspecionados, indevidamente, apenas 2 (dois) clientes.

Consistência = 0.5

Para $\delta = 0.5$, a quantidade de regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 185
- Regras inconscientes: 85

Os registros diários classificados como normal e anormal para a consistência considerada, podem ser observados na Tabela 6.7. A classificação correta de registros pertencentes ao conjunto de anormais foi de 35.3% dos dados e para o conjunto de registros normais, obteve-se a classificação correta de 98.8% dos registros.

Tabela 6.7 – Classificação de dados de treinamento para $\delta = 0.5$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	7116	85
	Anormal	408	223

Os comentários dos resultados obtidos nas análises diária e semanal, considerando o atributo de decisão normal para a seleção de regras de classificação e consistência 0.5, encontram-se descritos a seguir.

Análise Diária. Na classificação correta e incorreta dos registros na análise diária, considerando a consistência admitida, obteve-se os seguintes resultados: dos 223 (duzentos e vinte e três) registros anormais classificados corretamente, selecionou-se os 12 (doze) consumidores fraudadores pertencentes ao conjunto de dados de treinamento e, dos 85 (oitenta e cinco) registros classificados incorretamente como anormal, selecionou-se 20 (vinte) consumidores. Assim, um total de 32 (trinta e dois) clientes foram selecionados para inspeção técnica, sendo que o sucesso quanto a detecção de fraude para esta seleção foi de 37.5%.

Análise Semanal. Considerando $\delta = 0.5$, foram selecionados para inspeção técnica 22 (vinte e duas) unidades consumidoras de energia elétrica, conforme Tabela 6.8. Através desta análise, detectou-se 11 (onze) fraudadores de um conjunto de 12 (doze) e, obteve-se 50% de sucesso quanto às inspeções realizadas.

Comparando os resultados obtidos da análise diária e semanal, considerando a consistência igual 0.5, pode-se concluir que a análise semanal obteve melhor desempenho, uma vez que detectou 91.7% dos fraudadores utilizados no processo de treinamento e apresentou diminuição de 9 (nove) clientes classificados incorretamente como anormal para a inspeção técnica.

Tabela 6.8 – Identificação dos consumidores anormais - análise semanal. Dados de treinamento - ($\delta = 0.5$ e $D = \{normal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	3A	0	0	1	1	1	3	0	6
	2A	4	7	6	6	3	0	0	26
	2B	6	7	5	7	3	0	2	32
	1A	0	0	2	0	0	0	0	2
	4A	1	1	0	1	0	0	1	4
	1B	8	0	0	0	0	0	8	16
	4C	3	4	4	3	3	6	5	28
	2D	5	0	0	0	1	0	0	6
	2E	0	8	3	4	2	1	0	18
	3I	0	3	0	0	0	0	3	6
	2I	6	7	2	4	3	2	4	28
Incorreta	3A	0	0	1	1	0	2	0	4
	2A	0	0	0	1	1	0	0	2
	2B	3	3	1	3	0	0	0	10
	3C	0	1	1	0	0	0	0	2
	3D	0	0	2	0	0	0	0	2
	2C	0	1	0	0	0	0	1	2
	4C	1	0	0	0	0	0	1	2
	2D	2	0	0	0	0	0	0	2
	3J	0	2	2	1	0	0	1	6
	4D	1	1	1	0	0	0	1	4
	2I	0	0	3	3	0	0	0	6

Consistência = 0.7

Considerando agora $\delta = 0.7$, as regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 185
- Regras inconscientes: 54

A Tabela 6.9 apresenta o resultado da classificação dos registros diários como normal e anormal considerando $\delta = 0.7$. Um total de 67.5% dos registros de fraude foram classificados corretamente contra 5.7% de dados normais classificados incorretamente como fraudadores (anormais). Assim, um total de 7198 (sete mil cento e noventa e oito) registros foram classificados corretamente e 634 (seiscentos e trinta e quatro) classificados incorretamente. Analisando apenas os número de acertos na classificação dos registros, a consistência considerada apresenta bons resultados e

as análises abaixo, mostram os resultados quanto ao número de consumidores envolvidos na classificação anormal.

Tabela 6.9 – Classificação de dados de treinamento para $\delta = 0.7$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	6791	410
	Anormal	224	407

Análise Diária. Com os 407 (quatrocentos e sete) registros anormais classificados corretamente, encontraram-se todos os consumidores fraudadores do conjunto de dados de treinamento e, para os 410 (quatrocentos e dez) registros classificados incorretamente como anormal, os 27 (vinte e sete) consumidores pertencentes ao conjunto de dados de treinamento foram selecionados. Considerando estas informações, selecionou-se 39 (trinta e nove) clientes para inspeções, sendo que 30.7% desta seleção apresentaram sucesso quanto à detecção de fraude ou erro de medição.

Análise Semanal. Considerando a análise semanal, com $\delta = 0.7$, selecionou-se 31 (trinta e um) clientes para serem tecnicamente inspecionados, dos quais, 10 (dez) fraudadores de um conjunto de 12 (doze) foram localizados. Os clientes classificados indevidamente para a inspeção foram 21 (vinte e um), ou seja, 77.8% do total das unidades consumidoras de energia normais utilizadas no processo de treinamento. O resultado das inspeções apresentou 32.2% de sucesso quanto à detecção.

Observando os resultados das análises realizadas para $\delta = 0.7$, pode-se concluir que a análise semanal, mesmo identificando uma quantidade menor de fraudadores, apresentou melhor resultado na quantidade de clientes selecionados para a inspeção técnica que reduziu de 39 (trinta e nove) para 31 (trinta e um) o número de inspeções com relação à análise diária.

Consistência = 1

Para $\delta = 1$, apenas as regras consistentes, 185 (cento e oitenta e cinco), são utilizadas na classificação dos dados.

A Tabela 6.10 apresenta o resultado da classificação dos registros diários como normal e anormal para $\delta = 1$. Neste caso, 100% dos registros de fraude do conjunto de treinamento foram classificados corretamente e, classificaram-se indevidamente como anormais 84.3% de dados normais.

Tabela 6.10 – Classificação de dados de treinamento para $\delta = 1$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	1130	6071
	Anormal	0	631

Assim, apenas um total de 1661 registros foram classificados corretamente. Analisando os dados classificados como anormal, a consistência admitida não apresenta bons resultados. Estas informações podem ser reafirmadas nas análises a seguir.

Análise Diária. Para as classificações anormais, os dados classificados corretamente na análise diária foram 631 (seiscentos e trinta e um) registros, ou seja, 100% dos dados anormais utilizados no processo de treinamento. Os registros classificados indevidamente como anormais, referem-se a todos os consumidores pertencentes ao conjunto de dados normais, ou seja, 27 (vinte e sete) consumidores. Considerando que foram inspecionados 39 (trinta e nove) clientes, o resultado do sucesso na execução desta tarefa é de 30.7%.

Análise Semanal. Na análise semanal, com $\delta = 1$, os resultados de consumidores selecionados como anormais foram os mesmos da análise diária.

Considerando a consistência igual a 1, pode-se concluir que os resultados quanto à seleção incorreta de consumidores normais para inspeção, ao considerar apenas as regras consistentes, são ruins, uma vez que todos os clientes do conjunto de dados de treinamento são inspecionados.

Pode-se observar que as consistências mais viáveis para o conjunto de regras formados através da definição do atributo de decisão quando normal, conforme os resultados aqui apresentados, são 0.3 e 0.5. Para $\delta = 0.3$ a análise diária apresentou

bom desempenho em relação a quantidade de acertos no resultado das inspeções realizadas, 80%. Nesta análise localizou-se 11 (onze) dos 12 (doze) consumidores anormais existentes e selecionou-se apenas 2 (dois) consumidores indevidamente. Para $\delta = 0.5$ a análise semanal apresentou o melhor resultado e o índice de acerto nas inspeções realizadas foi de 50%. Nesta análise encontrou-se 11 (onze) dos 12 (doze) fraudadores existentes, e selecionou-se indevidamente 11 (onze) clientes normais, 40.7% do total utilizado nesse processo. As demais consistências analisadas, $\delta = 0.7$ e $\delta = 1$, apesar de detectarem todos os registros anormais do banco de dados, não são recomendadas devido ao alto índice de acusações indevidas.

- Atributo de Decisão = Anormal

Consistência = 0.3

Considerando $\delta = 0.3$, as regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 19
- Regras inconscientes: 46

A Tabela 6.11 apresenta o resultado da classificação dos registros diários como normal e anormal considerando $\delta = 0.3$.

Tabela 6.11 – Classificação de dados de treinamento para $\delta = 0.3$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	6784	417
	Anormal	221	410

Pode-se observar, na Tabela 6.11, que um total de 5.8% de dados normais são classificados incorretamente como anormais e 65% dos registros anormais são classificados corretamente. Assim, classificou-se corretamente um total de 7194 (sete mil cento e noventa e quatro) registros e, quanto à classificação incorreta, apresentou um total de 638 (seiscentos e trinta e oito) registros. Analisando apenas os número de acertos na classificação dos registros, a consistência considerada apresenta bons

resultados. As análises a seguir mostram os resultados quanto ao número de consumidores envolvidos na classificação anormal.

Análise Diária. Para os 417 (quatrocentos e dezessete) registros normais classificados incorretamente como anormais, os 27 (vinte e sete) clientes pertencentes ao conjunto de dados de treinamento foram selecionados para inspeção. Para os 410 (quatrocentos e dez) registros classificados corretamente como anormal encontraram-se todos os consumidores fraudadores do mesmo conjunto de dados considerado. Com base nestas informações, selecionou-se 39 (trinta e nove) consumidores para inspeções, sendo que apenas 30.7% destas inspeções apresentaram sucesso quanto à detecção de fraude ou erro de medição.

Análise Semanal. Na análise semanal, considerando $\delta = 0.3$, a quantidade de clientes selecionados para a inspeção reduz de 39 (trinta e nove) para 32 (trinta e dois), com relação a análise diária. Assim, 10 (dez) fraudadores de um total de 12 (doze) foram localizados e um total de 22 (vinte e dois) consumidores classificados indevidamente foram inspecionados, ou seja, 75.9% do total dos clientes com comportamento normal utilizados no treinamento. O resultado das inspeções apresenta 31.2% de sucesso quanto à detecção.

Comparando os resultados das duas análises para a consistência = 0.3, conclui-se que a análise diária torna-se inviável devido o alto índice de acusação indevida de fraudadores.

Consistência = 0.5

Para $\delta = 0.5$, a quantidade de regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 19
- Regras inconscientes: 27

Os registros diários classificados como normal e anormal para a consistência considerada, podem ser observados na Tabela 6.12. A classificação correta de

registros pertencentes ao conjunto de anormais foram de 38.4% dos dados e para o conjunto de registros normais, obteve-se a classificação correta para 98.5% dos registros.

Tabela 6.12 – Classificação de dados de treinamento para $\delta = 0.5$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	7097	104
	Anormal	389	242

Os comentários das análises diária e semanal encontram-se descritos a seguir.

Análise Diária. Considerando a consistência admitida, obteve-se os seguintes resultados para a classificação correta e incorreta dos registros na análise: 242 (duzentos quarenta e dois) registros anormais classificados corretamente, selecionando-se os 12 (doze) consumidores fraudadores pertencentes ao conjunto de dados de treinamento para a inspeção e, dos 104 (cento e quatro) registros classificados incorretamente como anormal, selecionou-se 23 (vinte e três) consumidores. Assim, um total de 35 (trinta e cinco) clientes foram indicados para inspeção técnica, sendo que o sucesso quanto a detecção de fraude para todos os clientes inspecionados foi de 37%.

Análise Semanal. Para $\delta = 0.5$, foram selecionados 21 (vinte e uma) unidades consumidores para inspeção, conforme Tabela 6.13. Através desta análise, detectou-se 91.6% dos fraudadores do banco de dados de treinamento e acusou indevidamente 37% dos registros normais. Logo, um total de 21 (vinte e um) clientes foram selecionados para inspeção técnica, resultando em 52.4% de acertos quanto aos fraudadores detectados nas inspeções.

Comparando os resultados obtidos nas análises realizadas, conclui-se que a análise semanal apresenta um melhor desempenho, mesmo não detectando todos os fraudadores utilizados no treinamento. O fato de apresentar uma quantidade menor de clientes classificados indevidamente como fraudadores, contribuiu para a definição de que esta análise é melhor na classificação de dados.

Tabela 6.13 – Identificação dos consumidores anormais - análise semanal. Dados de treinamento - ($\delta = 0.5$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	3A	0	1	2	2	1	4	0	10
	2A	4	7	6	6	3	0	0	26
	2B	8	7	4	7	3	1	2	32
	1A	0	0	3	1	0	0	0	4
	4A	1	1	0	1	0	2	1	6
	1B	8	0	0	0	0	0	8	16
	4C	3	4	4	3	3	6	5	28
	2D	5	0	0	0	1	0	0	6
	2E	0	8	4	4	3	1	0	20
	3I	0	3	0	0	0	0	3	6
	2I	6	6	3	6	3	4	4	32
Incorreta	3A	0	1	1	1	0	3	0	6
	2A	0	0	0	1	1	0	0	2
	2B	3	4	1	3	0	1	0	12
	3C	0	1	1	0	0	0	0	2
	3D	0	0	2	0	0	0	0	2
	4C	1	0	0	0	0	0	1	2
	2D	2	0	0	0	0	0	0	2
	3J	0	2	3	2	0	0	1	8
	4D	1	1		1	0	1	1	6
	2I	0	0	3	3	0	0	0	6

Consistência = 0.7

Considerando $\delta = 0.7$ para as regras de classificação, obtém-se um total de:

- Regras consistentes: 19
- Regras inconscientes: 2

A Tabela 6.14 apresenta a quantidade dos registros diários classificados como normal e anormal para a consistência considerada.

Tabela 6.14 – Classificação de dados de treinamento para $\delta = 0.7$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	7196	5
	Anormal	533	98

Para o conjunto de registros normais utilizados no processo de treinamento, obteve-se a classificação correta de 99.9% dos dados. Para o conjunto formado pelos registros anormais utilizados no mesmo processo, apenas 15.5% dos dados foram classificados corretamente.

Análise Diária. Com os 98 (noventa e oito) registros anormais classificados corretamente, pode-se observar através da Tabela 6.15 que foram encontrados 10 (dez) dos 12 (doze) consumidores fraudadores pertencentes ao conjunto de dados de treinamento. Os 5 (cinco) registros classificados incorretamente como anormal, referem-se a 2 (dois) consumidores deste mesmo conjunto. Considerando que as inspeções técnicas em consumidores classificados como anormal são realizadas a partir das informações diárias, um total de 12 (doze) inspeções técnicas são geradas, obtendo-se 83% de sucesso quanto à taxa de acerto na identificação de anormalidades.

Tabela 6.15 – Identificação dos consumidores anormais - análise diária.
Dados de treinamento - ($\delta = 0.7$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	3A	0	0	0	0	1	0	0	1
	2A	4	0	0	0	2	0	0	6
	2B	0	0	0	0	2	0	0	2
	4A	0	0	0	1	0	0	0	1
	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2D	0	0	0	0	1	0	0	1
	2E	0	1	0	0	2	0	0	3
	3I	0	2	0	0	0	0	0	2
	2I	7	3	0	0	2	0	0	12
Incorreta	4B	0	0	0	0	1	0	0	1
	2D	4	0	0	0	0	0	0	4

Análise Semanal. Para $\delta = 0.7$, selecionou-se, através da análise semanal, 5 (cinco) consumidores para inspeção, conforme Tabela 6.16. Assim, encontrou-se nesta análise, 4 (quatro) fraudadores de um conjunto de 12 (doze) e, obteve-se 80% de sucesso quanto taxa de acertos na detecção de anormalidade.

Assim, comparando os resultados de consumidores selecionados para a verificação técnica na análise diária, a análise semanal encontra uma quantidade

menor de fraudadores com relação ao conjunto de treinamento, mas aumenta a taxa de acertos em relação às inspeções realizadas.

Tabela 6.16 – Identificação dos consumidores anormais - análise semanal. Dados de treinamento - ($\delta = 0.7$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	4	0	0	0	2	0	0	6
	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2I	5	3	0	0	2	0	0	10
Incorreta	2D	2	0	0	0	0	0	0	2

Consistência = 1

Admitindo $\delta = 1$, um total de 19 (dezenove) regras são selecionadas para a classificação dos dados, sendo todas consistentes. Assim, o resultado da classificação dos registros diários em normal e anormal podem ser observados na Tabela 6.17.

Tabela 6.17 – Classificação de dados de treinamento para $\delta = 1$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	7201	0
	Anormal	547	84

Neste caso, classificou-se como anormal um total de 84 (oitenta e quatro) registros sendo que todos foram classificados corretamente e identificado 13.3% dos registros anormais utilizados no processo de treinamento, foram identificados. Os resultados obtidos na classificação dos dados para $\delta = 1$, podem ser examinados nas análises abaixo.

Análise Diária. Os dados classificados corretamente na análise diária referem-se a 9 (nove) dos 12 (doze) clientes anormais utilizados no processo de treinamento, conforme Tabela 6.18. Assim, o resultado das inspeções técnicas quanto aos clientes visitados é de 100% de acertos na busca de fraudadores.

Tabela 6.18 – Identificação dos consumidores anormais - análise diária. Dados de treinamento - ($\delta = 1$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	0	0	0	0	2	0	0	2
	2B	0	0	0	1	0	0	0	1
	4A	0	0	0	1	0	0	0	1
	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2D	0	0	0	0	1	0	0	1
	2E	0	1	0	0	1	0	0	2
	3I	0	2	0	0	0	0	0	2
	2I	1	3	0	0	1	0	0	5

Análise Semanal. Na análise semanal, para a mesma consistência, $\delta = 1$, são selecionados 74 (setenta e quatro) registros, referentes a 3 (três) consumidores anormais para serem inspecionados conforme Tabela 6.19. Este resultado apresenta 100% de acertos quanto à detecção da fraude na inspeção, mas detecta apenas 25% do total das unidades consumidoras de energia com anormalidade utilizadas no processo de treinamento.

Tabela 6.19 – Identificação dos consumidores anormais - análise semanal. Dados de treinamento - ($\delta = 1$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	1B	8	0	0	0	0	0	8	16
	4C	8	8	8	8	6	8	8	54
	2I	0	3	0	0	1	0	0	4

Para os resultados das análises realizadas, considerando $\delta = 1$, conclui-se que a análise diária apresenta melhor resultado pois detecta 75% do total de fraudadores utilizados no conjunto de dados de treinamento.

Pode-se observar que através da definição do atributo de decisão igual a anormal para a geração de regras de classificação, obteve-se resultados diferentes para as consistências analisadas. Todavia, os resultados mais significativos para a detecção de fraude são para $\delta = 0.5$, $\delta = 0.7$ e $\delta = 1$. A consistência $\delta = 0.5$ apresentou um bom resultado na análise semanal, encontrando 11 (onze) dos 12 (doze) fraudadores existentes e acusando indevidamente 10 (dez) clientes. Assim a taxa de acertos na detecção de fraudes referente às inspeções realizadas foi de 52.4%. Para $\delta = 0.7$, a análise diária apresentou bom desempenho em relação a

quantidade de acerto no resultado das inspeções realizadas, 80%, localizando 10 (dez) dos 12 (doze) clientes anormais existentes e selecionou indevidamente 2 (dois) consumidores. Para $\delta = 1$, a análise diária também apresentou um bom desempenho, pois não selecionou indevidamente nenhum cliente para ser inspecionado. Detectou 9 (nove) dos 12 (doze) clientes anormais existentes nos dados de treinamento e, a taxa de acerto nas inspeções realizadas foi de 100%. As demais análises realizadas, apesar de detectarem todos os registros anormais do banco de dados, não são recomendadas devido ao alto índice de acusações indevidas.

Comparando a quantidade de regras utilizadas para a classificação dos consumidores, pode-se observar que para $D = \{normal\}$ são requeridas, no mínimo, 185 (cento e oitenta e cinco) regras, enquanto que para $D = \{anormal\}$ utiliza-se em média 38 (trinta e oito) regras. Esta diferença é obvia, uma vez que se têm muito mais consumidores normais do que anormais no banco de dados utilizado neste trabalho.

Considerando que o objetivo é detectar fraudadores ou erro de medição, a melhor opção é considerar $D = \{anormal\}$ para a geração de regras de classificação de dados pois apresenta um menor custo e tempo computacional para o sistema. Quanto à definição do valor da consistência a ser admitida, esta é definida pelo usuário do sistema.

6.4. Processo de Teste

A partir dos resultados encontrados nas classificações dos registros através do processo de treinamento, será analisada a influência das regras selecionadas pela definição dos valores da consistência e atributo de decisão, no banco de dados de teste. Este procedimento é definido como processo de teste e tem como objetivo encontrar resultados satisfatórios quanto à classificação correta das unidades consumidoras anormais buscando encontrar o maior número desses consumidores e, apresentar uma menor taxa possível de erro quanto à classificação incorreta de consumidores como fraudadores.

O conjunto de regras utilizados neste processo é o mesmo utilizado no treinamento. Portanto, considerou-se o conjunto mínimo de atributos B^*_1, sendo que foram geradas por este reduto um total de 204 (duzentos e quatro) regras consistentes e 99 (noventa e nove) regras inconsistentes, totalizando 303 (trezentos e três) regras gerais.

O banco de dados utilizados para o teste é composto por 3477 (três mil quatrocentos e setenta e sete) registros, sendo 2930 (dois mil novecentos e trinta) registros com $D = \{normal\}$, referentes aos mesmos 27 (vinte e sete) consumidores utilizados no treinamento e, 547 (quinhentos e quarenta e sete) registros com $D = \{anormal\}$, referentes a 16 (dezesseis) clientes.

Simulou-se, para cada valor nominal do conjunto D, 4 (quatro) variações para o valor da consistência, da mesma forma como foi realizado na processo de treinamento. Os resultados quanto da classificação dos registros pelas regras consideradas podem ser observados nos tópicos a seguir.

- Atributo de Decisão = Normal

  Consistência = 0.3

  Considerando $\delta = 0.3$ para as regras de classificação, obtém-se um total de:

  - Regras consistentes: 185
  - Regras inconscientes: 97

A quantidade dos registros diários classificados como normal e anormal para a consistência considerada, encontram-se na Tabela 6.20.

Tabela 6.20 – Classificação de dados de teste para $\delta = 0.3$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2878	52
	Anormal	413	116

Considerando o conjunto total de registros do processo de teste, obteve-se a classificação correta de 86% dos dados.

Considerando apenas os registros classificados como anormais, foco de nosso trabalho, os resultados das análises diária e semanal, podem ser observados a seguir.

Análise Diária. Com um total de 116 (cento e dezesseis) registros anormais classificados corretamente, selecionou-se 11 (onze) dos 16 (dezesseis) clientes fraudadores pertencentes ao conjunto de dados de teste e, dos 52 (cinqüenta e dois) registros classificados incorretamente como anormal, 16 (dezesseis) consumidores foram selecionados.

Considerando a seleção de 27 (vinte e sete) clientes suspeitos de anormalidades para as inspeções técnicas, um total de 11 (onze) unidades consumidoras apresentaram resultados positivos quanto à confirmação de anormalidade. Assim, 40.7% do total das inspeções realizadas foram satisfatórias quanto à detecção de fraude ou erro de medição.

Análise Semanal. Através da análise semanal para $\delta = 0.3$, selecionou-se 11 (onze) consumidores para inspeção, conforme Tabela 6.21.

Tabela 6.21 – Identificação dos consumidores anormais - análise semanal. Dados de teste - ($\delta = 0.3$ e $D = \{normal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	2	0	0	0	0	0	0	2
	4A	3	0	0	0	0	0	3	6
	4B	8	8	8	7	7	8	8	54
	2C	3	3	3	3	4	4	4	24
	2F	6	0	0	0	0	0	6	12
	2I	1	0	0	0	0	0	1	2
Incorreta	3A	2	2	2	1	1	1	1	10
	3C	0	0	1	0	1	0	0	2
	3D	0	0	0	0	2	0	0	2
	4C	0	1	2	0	1	0	0	4
	2D	2	0	0	0	0	0	0	2

A quantidade de registros para esta análise é menor que a quantidade da análise diária, devido ao fato de serem selecionados os registros que apresentam mais de uma acusação de anormalidade num período de uma

semana. Portanto, selecionou-se 100 (cem) registros anormais corretamente, identificando 6 (seis) fraudadores de um conjunto de 16 (dezesseis) e, 20 (vinte) registros anormais indevidamente referentes a 5 (cinco) clientes, obtendo-se 54.5% de sucesso nas 11 (onze) inspeções realizadas.

Com base nos resultados das análises realizadas para $\delta = 0.3$, pode-se concluir que a quantidade de fraudadores identificados foi bem maior na análise diária, sendo encontrados 11 (onze) clientes contra 6 (seis) detectados através da análise semanal. Quanto ao sucesso em inspeções realizadas, a análise semanal apresentou um melhor desempenho uma vez que obteve uma taxa de acerto de 54.5% sendo inspecionados indevidamente 5 (cinco) clientes contra 40.7% de acerto na análise diária em que foram inspecionados 16 (dezesseis) consumidores

Consistência = 0.5

Para $\delta = 0.5$, a quantidade de regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 185
- Regras inconscientes: 85

Os registros diários classificados como normal e anormal para a consistência considerada, podem ser observados na Tabela 6.22. A classificação correta de registros pertencentes ao conjunto de anormais, no processo de teste, foi de 31.3% dos dados e para o conjunto de registros normais, obteve-se a classificação correta para 97.2% dos dados.

Tabela 6.22 – Classificação de dados de teste para $\delta = 0.5$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2849	81
	Anormal	376	171

Análise Diária. Para a classificação correta de dados anormais, ou seja, 171 (cento e setenta e um) registros, selecionou-se para inspeção 13 (treze) clientes, apontando 81.2% do total de fraudadores utilizados no teste . A

classificação incorreta de dados anormais se deu através de 81 (oitenta e um) registros, selecionando indevidamente 18 (dezoito) consumidores como fraudadores. Assim, um total de 31 (trinta e um) clientes foram inspecionados, sendo que o sucesso quanto a detecção de fraude para estas inspeções foi de 35.5%.

Análise Semanal. Para $\delta = 0.5$, ao todo, são indicadas 19 (dezenove) unidades consumidoras de energia elétrica para serem inspecionadas, conforme Tabela 6.23. Obtive-se os seguintes resultados para a análise semanal: seleção correta de 140 (cento e quarenta) registros, indicando 11 (onze) clientes para inspeção e, seleção incorreta de 36 (trinta e seis) registros referentes a 8 (oito) consumidores normais acusados de anormalidade. Considerando que estas unidades selecionadas através da análise semanal passaram por uma verificação técnica, a taxa de acertos obtida quanto ao sucesso de detecção de fraude ou erro de medição foi de 57.9%. Através desta análise, 68.7% do total de fraudadores do banco de dados utilizado no teste foram identificados.

Tabela 6.23 – Identificação dos consumidores anormais - análise semanal. Dados de teste - ($\delta = 0.5$ e $D = \{normal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	2	3	2	4	1	0	0	12
	3B	0	1	4	3	1	0	3	12
	3C	0	2	1	3	0	0	0	6
	4A	2	0	0	0	0	0	2	4
	4B	8	8	8	7	7	8	8	54
	2C	3	1	1	0	4	4	3	16
	2F	7	0	0	0	0	0	7	14
	3I	3	3	0	0	0	0	2	8
	3J	0	0	0	3	1	0	0	4
	4D	2	2	3	0	0	0	1	8
	2I	1	0	0	0	0	0	1	2
Incorreta	3A	2	2	2	1	1	1	1	10
	2B	1	1	0	1	1	0	0	2
	3C	2	4	1	0	1	0	0	8
	3D	0	0	0	1	3	0	0	4
	4C	0	1	2	0	1	0	0	4
	2D	2	0	0	0	0	0	0	2
	3J	0	0	2	0	0	0	0	2
	2G	1	0	0	0	1	0	0	2

Comparando os resultados das duas análises para a consistência = 0.5, conclui-se que na análise diária encontra-se mais fraudadores do que na análise semanal. Entretanto, na análise semanal inspeciona-se uma quantidade bem menor de clientes normais classificados incorretamente, 44.4% a menos.

Consistência = 0.7

Considerando $\delta = 0.7$, as regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 185
- Regras inconscientes: 54

A Tabela 6.24 apresenta os resultados da classificação dos registros diários como normal e anormal, considerando $\delta = 0.7$, em que um total de 47.9% dos registros de fraude foram classificados corretamente e apenas 8.1% de dados normais foram classificados incorretamente como anormais. Assim, 2954 (dois mil novecentos e cinqüenta e quatros) registros foram classificados corretamente e 523 (quinhentos e vinte e três) foram classificados incorretamente.

Tabela 6.24 – Classificação de dados de teste para $\delta = 0.7$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2692	238
	Anormal	285	262

Análise Diária. Para os 262 (duzentos e sessenta e dois) registros anormais classificados corretamente, encontraram-se 14 (quatorze) clientes fraudadores do conjunto de dados de teste. Para os 238 (duzentos e trinta e oito) registros classificados incorretamente como anormal, 25 (vinte e cinco) clientes foram identificados indevidamente como fraudadores. Portanto, selecionou-se 39 (trinta e nove) clientes para inspeção técnica, sendo que 35.9% destas inspeções apresentaram sucesso quanto à detecção de fraude ou erro de medição.

Análise Semanal. Considerando a análise semanal, com $\delta = 0.7$, a quantidade

de clientes selecionados para a inspeção reduz de 39 (trinta e nove) para 27 (vinte e sete), com relação a análise diária. Na análise semanal, dos 194 (cento e noventa e quatro) registros identificados, selecionou-se 11 (onze) fraudadores de um conjunto de 16 (dezesseis). Quanto à classificação incorreta, 138 (cento e trinta e oito) registros identificam indevidamente 16 (dezesseis) consumidores. Portanto, as inspeções realizadas nas unidades selecionadas, resultaram em 40.7% de acertos na detecção de fraude ou erro de medição.

Comparando os resultados das duas análises para $\delta = 0.7$ conclui-se que, na análise diária encontra-se mais fraudadores do que na análise semanal. Porém, na análise semanal inspeciona-se 9 (nove) clientes acusados indevidamente a menos que na análise diária.

Consistência = 1

Para $\delta = 1$, somente as regras consistentes são selecionadas:

- Regras consistentes: 185
- Regras inconscientes: 0

A Tabela 6.25 apresenta o resultado da classificação dos registros diários como normal e anormal para $\delta = 1$.

Tabela 6.25 – Classificação de dados de teste para $\delta = 1$ e $D = \{normal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	438	2492
	Anormal	60	487

Análise Diária. Para as classificações anormais, os dados classificados corretamente na análise diária foram 487 (quatrocentos e oitenta e sete) registros, ou seja, 89% dos dados anormais utilizados no processo de teste, selecionando os 16 (dezesseis) clientes para inspeção. Os registros classificados indevidamente como anormais referem-se a 85% dos dados normais pertencentes ao conjunto de teste, identificando para inspeção um

total de 27 (vinte e sete) clientes, ou seja, todos utilizados para a realização dos testes. Considerando que são inspecionados 43 (quarenta e três) consumidores, o resultado de sucesso na execução desta tarefa é de 37.2%.

Análise Semanal. Na análise semanal, com $\delta = 1$, dos 276 (duzentos e setenta e seis) registros identificados, selecionou-se 12 (doze) fraudadores de um conjunto de 16 (dezesseis). Quanto à classificação incorreta, dos 566 (quinhentos e sessenta e seis) registros identificou-se indevidamente 20 (vinte) consumidores. Portanto, a quantidade de clientes selecionados para a inspeção totalizou 32 (trinta e dois), dos quais 37.5% dos resultados das inspeções realizadas nas unidades foram de sucesso na detecção de fraude ou erro de medição.

Comparando os resultados das duas análises para $\delta = 1$ conclui-se que, na análise diária encontram-se todos os fraudadores pertencentes ao conjunto de teste enquanto que a análise semanal detecta-se apenas 12 (doze). Entretanto, a quantidade de clientes acusados indevidamente na análise diária é bem maior do que na análise semanal.

Comparando todos os resultados obtidos na variação do valor da consistência, conclui-se que os resultados dos testes estão condizentes com os resultados de treinamento, apresentando um melhor desempenho para os conjuntos de regras formados por $\delta = 0.3$ e $\delta = 0.5$. Entretanto, para análise diária e análise semanal, com consistência igual a 0.3 e 0.5, respectivamente, encontrou-se 11 (onze) dos 16 (dezesseis) fraudadores nas duas situações, mas, a acusação indevida para a primeira análise, é o dobro da análise semanal. Quanto ao sucesso em inspeções realizadas, a análise semanal, para $\delta = 0.5$, apresentou um melhor desempenho obtendo uma taxa de acerto de 57.9%, enquanto que a analise diária, para $\delta = 0.3$, obteve uma taxa de acerto de 40.7% na detecção de anormalidades.Os demais resultados obtidos na classificação de dados de teste através da variação do valor da consistência, mesmo encontrando um número maior de fraudadores, apresentam um ponto negativo, uma taxa muito alta de acusação indevida de clientes normais como fraudadores. Para $\delta = 0.7$ e $\delta = 1$, na análise semanal acusa-se indevidamente como fraudador, em média, 18 (dezoito) de um total de 27 (vinte e sete) clientes e, na

análise diária acusa-se indevidamente como anormal, em média, 26 (vinte e seis) consumidores.

- Atributo de Decisão = Anormal

Consistência = 0.3

Considerando $\delta = 0.3$, as regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 19
- Regras inconscientes: 46

A classificação dos registros diários como normal e anormal considerando $\delta = 0.3$, pode ser observada na Tabela 6.26.

Tabela 6.26 – Classificação de dados de teste para $\delta = 0.3$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2733	197
	Anormal	316	231

Assim, foram classificados corretamente 85.2% dos registros do banco de dados de teste.

Análise Diária. Dos 231 (duzentos e trinta e um) registros classificados corretamente como anormal, selecionou-se 14 (quatorze) consumidores fraudadores pertencentes ao conjunto de dados de teste. Para os 197 (cento e noventa e sete) registros normais classificados incorretamente como anormais, 24 (vinte e quatro) dos 27 (vinte e sete) clientes do conjunto de teste foram selecionados. Com base nestas informações, selecionou-se 38 (trinta e nove) consumidores para inspeções, sendo que apenas 36.8% destas inspeções apresentaram sucesso quanto à detecção de fraude ou erro de medição.

Análise Semanal. Na análise semanal para $\delta = 0.3$, selecionou-se

corretamente 170 (cento e setenta) registros como anormais e 11 (onze) fraudadores, de um total de 16 (dezesseis), foram localizados. A classificação incorreta como anormais totalizou 104 (cento e quatro) registros e um total de 14 (quatorze) clientes foram selecionados para inspeção, ou seja, 51.8% do total dos clientes com comportamento normal utilizados no conjunto de teste. As inspeções foram realizadas nos 25 (vinte e cinco) consumidores selecionados e o resultado quanto ao acerto na detecção de fraude ou erro de medição foi de 44%.

Consistência = 0.5

Para $\delta = 0.5$, a quantidade de regras utilizadas na classificação dos dados de entrada são:

- Regras consistentes: 19
- Regras inconscientes: 27

Os registros diários classificados como normal e anormal para a consistência considerada, podem ser observados na Tabela 6.27.

Tabela 6.27 – Classificação de dados de teste para $\delta = 0.5$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2883	47
	Anormal	404	143

O índice de classificação correta dos registros pertencentes ao conjunto de dados anormais foi de 26% e para o conjunto de registros normais, obteve-se um índice de classificação correta de 98.4%. Os comentários das análises diária e semanal encontram-se descritos a seguir.

Análise Diária. Considerando a consistência admitida, obteve-se os seguintes resultados para a classificação dos registros na análise diária: 143 (cento e quarenta e três) registros anormais classificados corretamente, seleção de 13 (treze) consumidores fraudadores pertencentes ao conjunto de dados de teste para a inspeção e, seleção de 14 (quatorze) consumidores dos 47 (quarenta e

sete) registros classificados incorretamente como anormal. Assim, um total de 27 (vinte e sete) clientes foram indicados para inspeção técnica, sendo que o sucesso na detecção de fraude para todos os clientes inspecionados foi de 48%.

Análise Semanal. Através da análise semanal, para $\delta = 0.5$, foram selecionados 17 (dezessete) unidades consumidores para inspeção com 64.7% de acertos na detecção de fraude, conforme Tabela 6.28. Nesta análise semanal, detectou-se 68.7% dos fraudadores do banco de dados de teste e acusou-se indevidamente 22.2% dos clientes normais utilizados neste processo.

Tabela 6.28 – Identificação dos consumidores anormais - análise semanal. Dados de teste - ($\delta = 0.5$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	3	3	2	4	2	2	2	18
	3B	5	2	4	4	6	6	3	30
	3C	0	3	1	3	5	4	0	16
	4A	1	0	0	0	0	0	1	2
	4B	0	0	8	7	7	8	0	30
	2C	2	1	1	0	0	1	3	8
	2F	7	0	0	0	0	0	7	14
	3I	3	4	2	0	4	4	3	20
	3J	0	0	0	4	5	5	0	14
	4D	3	3	3	0	3	3	1	16
	2I	0	0	0	0	0	0	2	2
Incorreta	3A	0	1	1	1	0	3	0	6
	2A	0	0	0	1	1	0	0	2
	2B	3	4	1	3	0	1	0	12
	3C	0	1	1	0	0	0	0	2
	3D	0	0	2	0	0	0	0	2
	4C	1	0	0	0	0	0	1	2

Comparando o resultado obtido da análise diária, em que foram selecionados 27 (vinte e sete) consumidores para a inspeção, com o resultado obtido da análise semanal, em que 17 (dezessete) clientes foram inspecionados, pode-se concluir que a análise semanal apresenta o melhor desempenho, pois, além de detectar um bom número de fraudadores, acusou indevidamente de anormalidade uma quantidade pequena de clientes em relação ao banco de dados utilizado no teste.

Consistência = 0.7

Considerando $\delta = 0.7$ para as regras de classificação, obtém-se um total de:

- Regras consistentes: 19
- Regras inconscientes: 2

A quantidade dos registros diários classificados como normal e anormal para a consistência considerada encontra-se na Tabela 6.29. Para o conjunto de registros normais utilizados no processo de teste, obteve-se a classificação correta de 99.6% dos dados. Para o conjunto formado pelos registros anormais utilizados no mesmo processo, apenas 15.5% dos dados foram classificados corretamente.

Tabela 6.29 – Classificação de dados de teste para $\delta = 0.7$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2920	10
	Anormal	462	85

Análise Diária. Com os 85 (oitenta e cinco) registros anormais classificados corretamente, pode-se observar, através da Tabela 6.30, que foram encontrados 9 (nove) dos 16 (dezesseis) consumidores fraudadores pertencentes ao conjunto de dados de teste. Os 10 (dez) registros classificados incorretamente como anormal, referem-se a 5 (cinco) consumidores deste mesmo conjunto. Considerando que as inspeções técnicas nos consumidores classificados como anormais são realizadas a partir das informações diárias, conforme a Tabela 6.30, um total de 66.7% dessas inspeções apresentam resultados satisfatórios na detecção de fraude ou erro de medição.

Análise Semanal. Para $\delta = 0.7$, selecionou-se do conjunto de dados de teste, através da análise semanal, 76 (setenta e seis) registros anormais classificados corretamente, sendo encontrados 6 (seis) dos 16 (dezesseis) consumidores fraudadores. Dos 6 (seis) registros classificados incorretamente como anormal, selecionou-se 2 (dois) consumidores (Tabela 6.31). Nesta análise obteve-se 75% de sucesso quanto às inspeções realizadas.

Tabela 6.30 – Identificação dos consumidores anormais - análise diária.
Dados de teste - ($\delta = 0.7$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	2	0	0	0	0	0	0	2
	3C	0	1	0	0	0	0	0	1
	4A	3	0	0	0	0	0	3	6
	4B	0	0	8	7	7	8	0	30
	2C	3	3	3	3	4	4	4	24
	2F	8	0	0	0	0	0	8	16
	3I	0	1	0	0	0	0	0	1
	3J	0	0	0	1	0	0	0	1
	2I	0	0	0	0	0	0	4	4
Incorreta	3A	0	0	2	1	1	1	0	5
	3C	0	0	0	0	1	0	0	1
	2C	0	0	0	1	0	0	0	1
	2D	2	0	0	0	0	0	0	2
	2G	1	0	0	0	0	0	0	1

Comparando os resultados das duas análises realizadas, considerando a $\delta = 0.7$, pode-se concluir que na análise diária detecta-se 56.2% do total de fraudadores existentes no conjunto de teste enquanto que na análise semanal, apenas 37.5% são encontrados. Quanto a classificação indevida de clientes normais em anormais, a análise semanal apresenta o melhor resultado pois apenas 7.4% do total de clientes normais são inspecionados indevidamente.

Tabela 6.31 – Identificação dos consumidores anormais - análise semanal.
Dados de teste - ($\delta = 0.7$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	2A	2	0	0	0	0	0	0	2
	4A	3	0	0	0	0	0	3	6
	4B	0	0	8	7	7	8	0	30
	2C	3	3	3	3	4	4	4	24
	2F	6	0	0	0	0	0	6	12
	2I	0	0	0	0	0	0	2	2
Incorreta	3A	0	0	1	1	1	1	0	4
	2D	2	0	0	0	0	0	0	2

Consistência = 1

Admitindo $\delta = 1$, um total de 19 (dezenove) regras são selecionadas para a classificação dos dados, sendo todas consistentes. O resultado da classificação dos registros diários em normal e anormal podem ser observados na Tabela 6.32.

Tabela 6.32 – Classificação de dados de teste para $\delta = 1$ e $D = \{anormal\}$.

		Classificação	
		Normal	Anormal
Dados	Normal	2923	7
	Anormal	464	83

Análise Diária. Os dados classificados corretamente na análise diária referem-se a 8 (oito) dos 16 (dezesseis) clientes anormais utilizados no processo de teste e, das classificações incorretas como anormais, são indevidamente selecionados 3 (três) clientes (Tabela 6.33). Logo, o resultado das inspeções técnicas quanto aos clientes visitados é de 72.7% de acertos na busca de fraudadores.

Tabela 6.33 – Identificação dos consumidores anormais - análise diária. Dados de teste - ($\delta = 1$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	3C	0	1	0	0	0	0	0	1
	4A	3	0	0	0	0	0	3	6
	4B	0	0	8	7	7	8	0	30
	2C	3	3	3	3	4	4	4	24
	2F	8	0	0	0	0	0	8	16
	3I	0	1	0	0	0	0	0	1
	3J	0	0	0	1	0	0	0	1
	2I	0	0	0	0	0	0	4	4
Incorretas	3A	0	0	2	1	1	1	0	5
	3C	0	0	0	0	1	0	0	1
	2C	0	0	0	1	0	0	0	1

Análise Semanal. Na análise semanal, para a mesma consistência, $\delta = 1$, são classificados corretamente 74 (setenta e quatro) registros anormais, referentes a 5 (cinco) fraudadores selecionados para a inspeção e, 4 (quatro) registros classificados indevidamente referentes a 1 (um) consumidor normal, também selecionado para inspeção (Tabela 6.34). Este resultado apresenta 83.3% de acertos na detecção da fraude na inspeção, mas detecta apenas 31.25% do total das unidades consumidoras de energia com anormalidades utilizadas no processo de teste.

Comparando os resultados das duas análises realizadas, considerando a $\delta = 1$, pode-se concluir que na análise diária detecta-se 8 (oito) dos 16 (dezesseis)

fraudadores existentes no conjunto de teste, enquanto que na análise semanal, apenas 5 (cinco) são encontrados. Quanto à classificação indevida de clientes normais em anormais, a análise semanal apresenta o melhor resultado pois apenas 3.7% do total de clientes normais são inspecionados indevidamente.

Tabela 6.34 – Identificação dos consumidores anormais - análise semanal. Dados de teste - ($\delta = 1$ e $D = \{anormal\}$).

Classificação	Identificação dos Consumidores	Registros por dias da semana							Total de Registros
		Dom	Seg	Ter	Qua	Qui	Sex	Sab	
Correta	4A	3	0	0	0	0	0	3	6
	4B	0	0	8	7	7	8	0	30
	2C	3	3	3	3	4	4	4	24
	2F	6	0	0	0	0	0	6	12
	2I	0	0	0	0	0	0	2	2
Incorreta	3A	0	0	1	1	1	1	0	4

Através da definição do atributo de decisão igual a anormal e, variação do valor da consistência para selecionar as regras a serem utilizadas na classificação dos registros, obteve-se resultados diferentes na identificação dos fraudadores do conjunto de dados de teste. Contudo, os resultados mais significativos para a detecção de fraude ou erro de medição são para $\delta = 0.5$, $\delta = 0.7$ e $\delta = 1$, confirmando os resultados obtidos no processo de treinamento. Entretanto, a consistência $\delta = 0.5$ apresentou um bom resultado na análise semanal, encontrando 11 (onze) dos 16 (dezesseis) fraudadores existentes e acusando indevidamente 6 (seis) clientes e, contribuindo com uma considerável taxa de acerto de 64.7% em relação às inspeções realizadas. Na análise diária encontrou-se 13 (treze) dos 16 (dezesseis) fraudadores e acusou indevidamente 14 (quatorze) clientes. Assim a taxa de acerto na detecção de fraudes referente às inspeções realizadas foi de 48.1%. Nenhum valor de consistência analisado, classificou 100% dos fraudadores utilizados no processo de teste.

6.5. Comparação dos Resultados

Das discussões precedentes, pode-se concluir que encontrar um resultado ideal, ou seja, detectar todos os clientes anormais, com uma alta taxa de acerto nas inspeções na detecção fraude ou erro de medição e, com isto um baixo índice de acusações indevidas de anormalidades, não é simples. A estratégia adotada neste

trabalho para buscar a melhorar performance do sistema, foi a criação da consistência admitida.

As comparações entre os resultados obtidos nos processos de treinamento e teste, conforme variação no valor de consistência admitida, podem ser observados nas figuras seguintes e encontram-se organizadas da seguinte forma: classificação correta dos registros anormais como fraudadores, tanto para o valor nominal do atributo de decisão igual a normal como para anormal; quantidade dos consumidores identificados através dos registros de classificação correta dos fraudadores, tanto para $D = \{normal\}$ quanto para $D = \{anormal\}$; classificação incorreta dos registros normais como fraudadores considerando as regras de classificação gerada pelos dois valores nominais de D; quantidade dos consumidores identificados indevidamente através dos registros de classificação incorreta dos fraudadores, também para $D = \{normal\}$ e $D = \{anormal\}$; e por fim, a comparação das taxas de acerto na identificação de anormalidades.

Como relatado anteriormente, uma grande quantidade de registros classificados correta ou incorretamente, não significa selecionar muitos consumidores, porque muitos registros selecionados podem referir-se a um mesmo consumidor. Portanto, as análises das Figuras 6.1 e 6.2 serão realizadas em conjunto, bem como as análises das Figuras 6.3 e 6.4.

Fazendo uma análise apenas qualitativa das Figuras 6.1 e 6.2, a maior quantidade de fraudadores classificados corretamente, tanto para a análise diária quanto para a análise semanal, ocorrem em duas situações: para $\delta = 1$ com $D = \{normal\}$ e para $\delta = 0.3$ com $D = \{anormal\}$. Conforme relatado anteriormente, analisar apenas a quantidade de fraudadores corretamente identificados, não é suficiente para definir os valores de δ e D. Continuando a análise dos resultados obtidos para as consistências 1 e 0.3 e seus respectivos valores de D, pode-se observar, através das Figuras 6.4 e 6.5 respectivamente, que a classificação indevida de fraudadores e a taxa de acerto nas inspeções técnicas apresentam valores insatisfatórios para o problema.

Fazendo essa mesma análise qualitativa, considerando agora $\delta = 0.3$ com $D = \{normal\}$ e $\delta = 1$ com $D = \{anormal\}$, pode-se observar que a quantidade de

fraudadores encontrados através da classificação dos dados é bem menor quando comparados com os resultados obtidos na análise anterior. Já com relação a classificação indevida de fraudadores, acusa-se poucos clientes, indevidamente, de fraudes. Quanto à taxa de acerto na inspeção técnica, os resultados apresentam as melhores taxas comparadas com os resultados obtidos pelas demais consistências. Entretanto, as consistências consideradas para esta análise, permitem a localização de poucos fraudadores, objetivo maior deste trabalho. Com isto, tornam-se insatisfatórias para a solução do problema.

Analisando os resultados obtidos considerando as seguintes situações: $\delta = 0.5$ com $D = \{normal\}$ e $\delta = 0.7$ com $D = \{anormal\}$, a quantidade de fraudadores corretamente identificados é maior para a consistência 0.5, tanto na análise diária quanto na semanal. Já para a quantidade de clientes classificados indevidamente como anormais, o resultado das análises diária e semanal para a consistência 0.7, apresenta melhor desempenho. As taxas de acerto para $\delta = 0.7$ com $D = \{anormal\}$, é mais elevada.

Analisando os resultados obtidos considerando agora as seguintes situações: $\delta = 0.7$ com $D = \{normal\}$ e $\delta = 0.5$ com $D = \{anormal\}$, tem-se: a quantidade de fraudadores identificados corretamente para as duas situações é bem parecida. Continuando a análise dos resultados obtidos para estas duas consistências, a quantidade de clientes classificados indevidamente como anormais, com $\delta = 0.5$ e $D = \{anormal\}$ apresenta um melhor desempenho, pois acusa indevidamente uma quantidade menor de clientes, tanto na análise diária, quanto na análise semanal. Com relação à taxa de acerto na inspeção técnica, com $\delta = 0.5$ e $D = \{anormal\}$ é relativamente melhor.

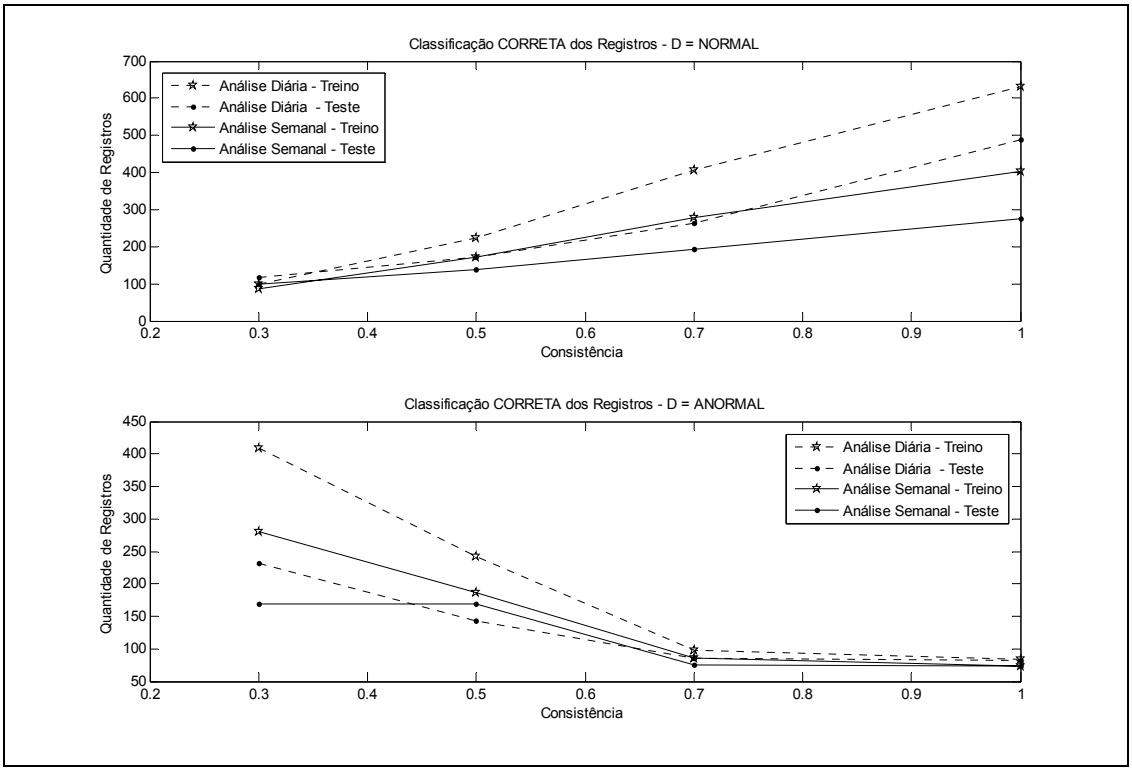


Figura 6.1 – Classificação correta dos registros anormais.

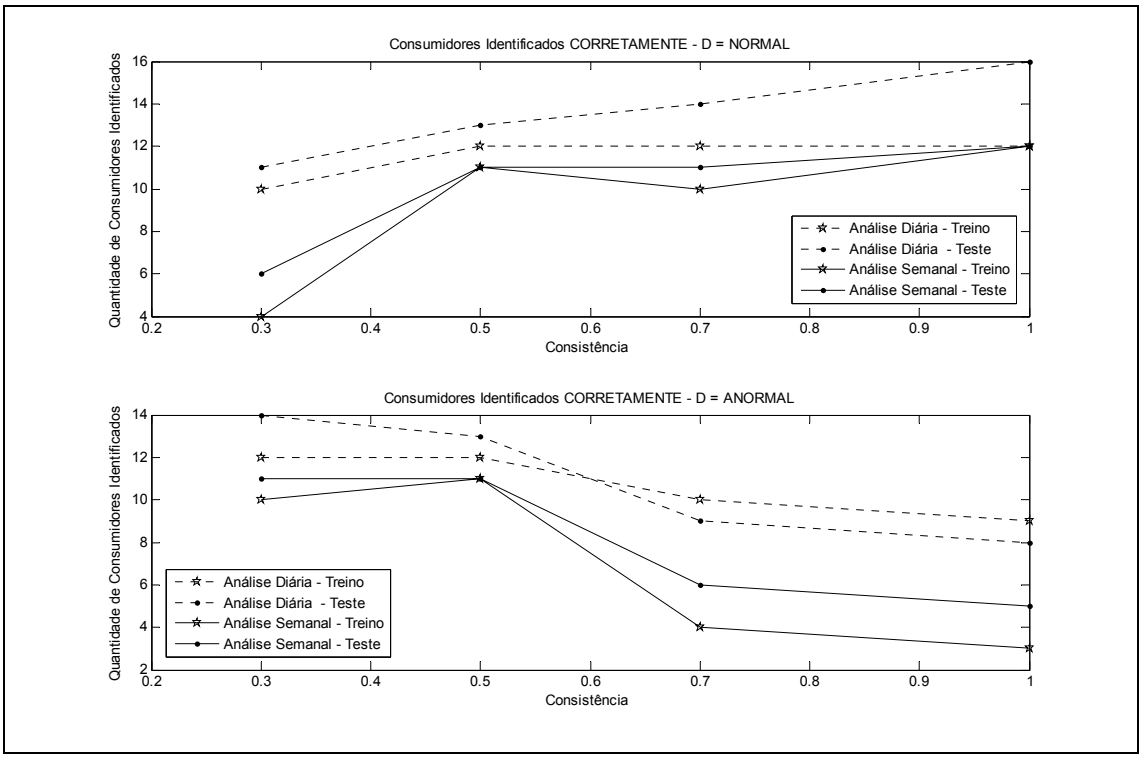


Figura 6.2 – Quantidade de consumidores classificados corretamente como anormais.

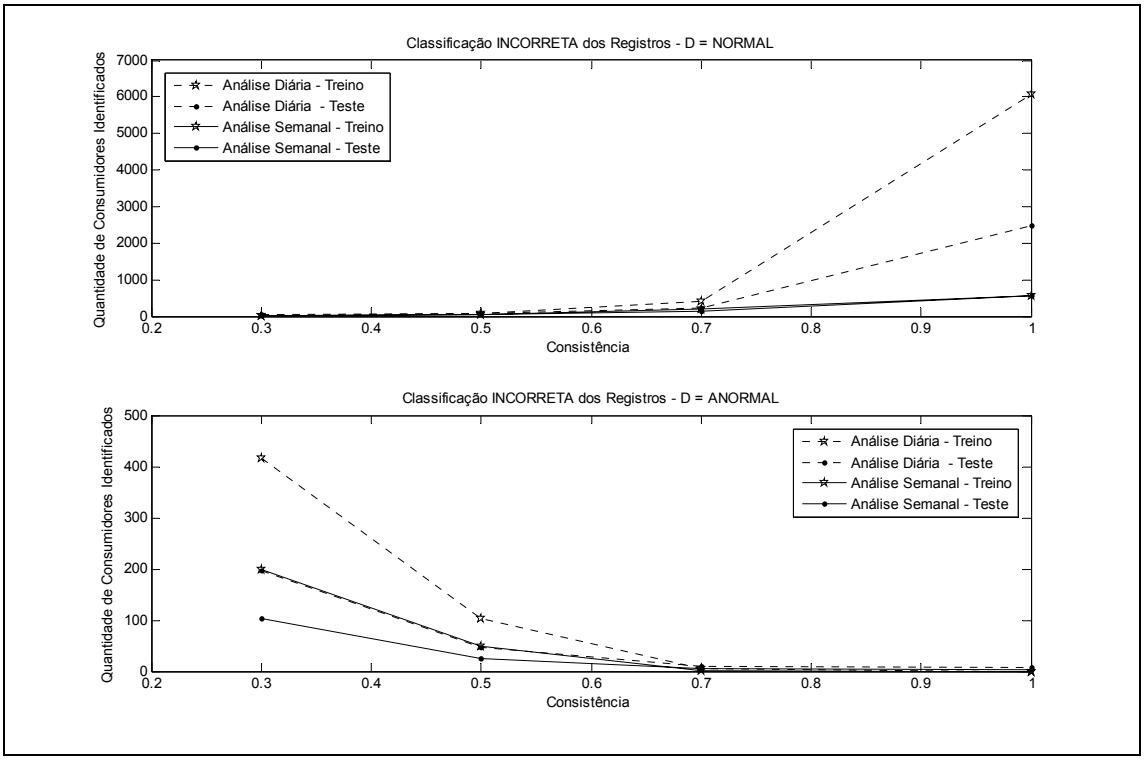


Figura 6.3 – Classificação indevida dos registros como anormais.

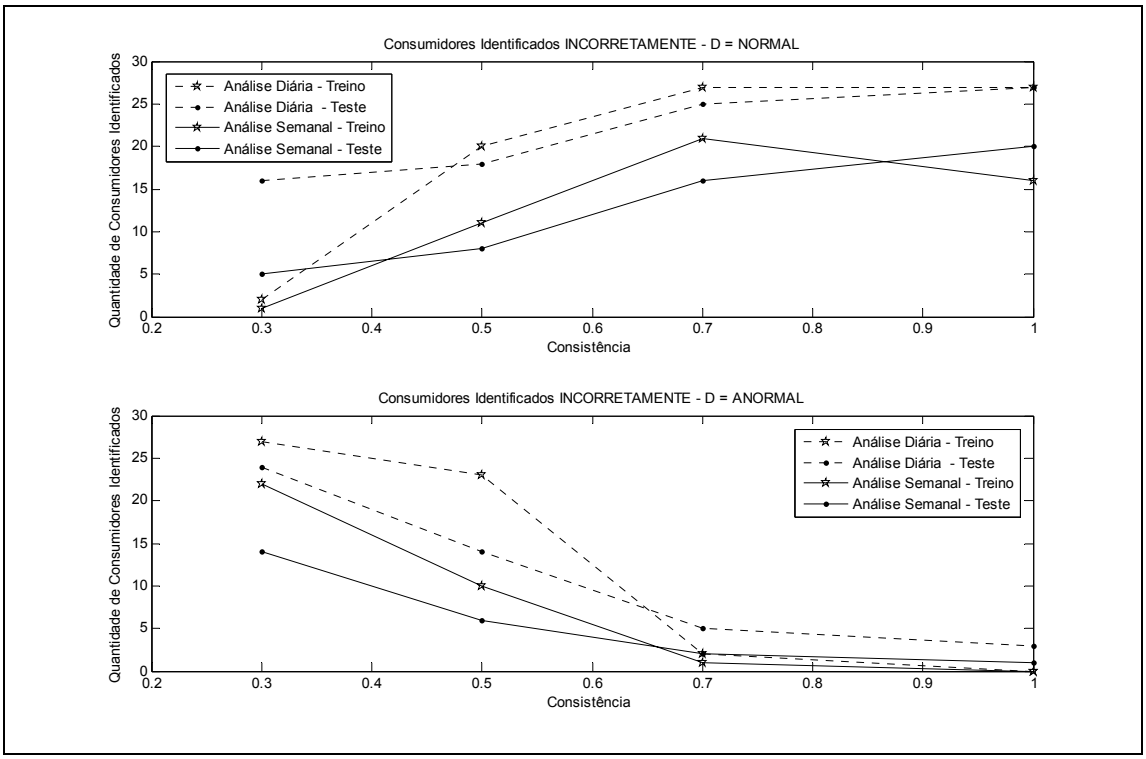


Figura 6.4 – Quantidade de consumidores classificados indevidamente como anormais.

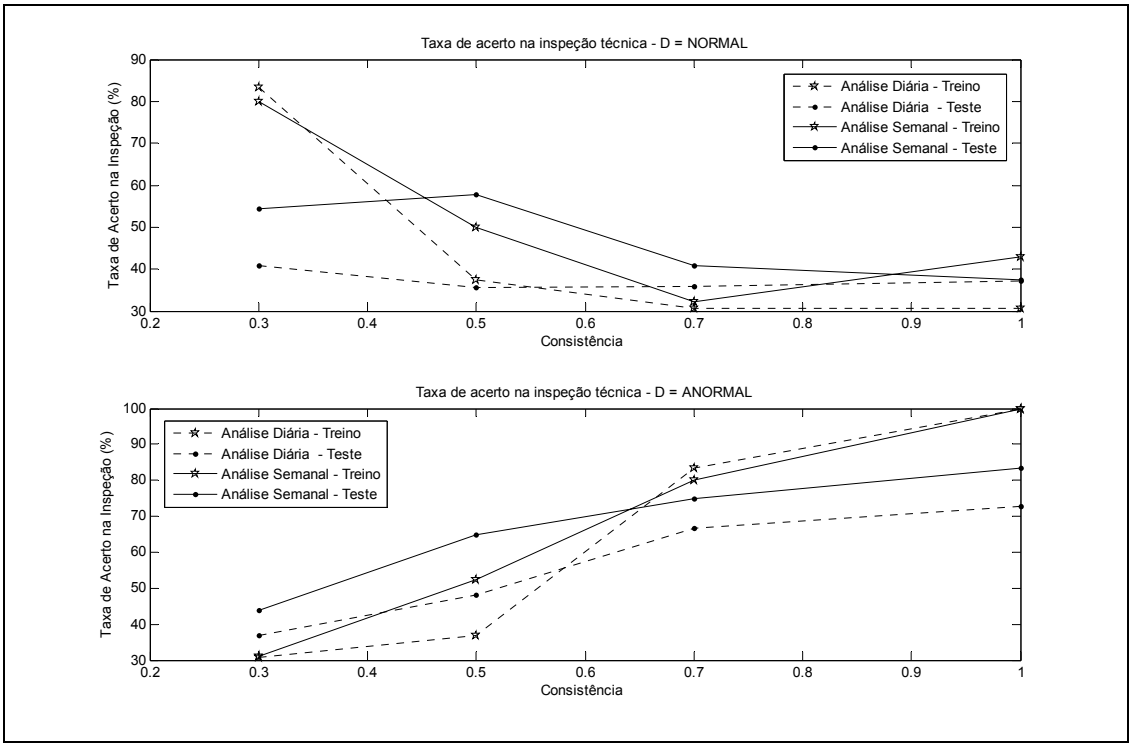


Figura 6.5 – Taxa de acerto na identificação dos consumidores classificados como anormais.

Levando em consideração o objetivo de detectar clientes anormais, com uma alta taxa de acerto nas inspeções quanto à detecção de fraude, conseqüentemente, baixo índice de acusações indevidas de anormalidades, os dois melhores desempenhos na detecção de fraude ou erro de medição, obtidos pelo sistema desenvolvido, considerando os dados utilizados no processo de teste, são:

Resultado 1:

- Tipo de análise: semanal
- Consistência admitida: 0.5
- Atributo de decisão para geração das regras: anormal
- Quantidade de regras utilizadas na classificação dos dados: 46
- Quantidade de fraudadores identificados: 11/16 (68.75%)
- Quantidade de consumidores acusados indevidamente de fraude: 6/27 (22.22%)
- Taxa de acerto na inspeção técnica realizada: 11/17 (64.70%)

Resultado 2:

- Tipo de análise: diária
- Consistência admitida: 0.7
- Atributo de decisão para geração das regras: anormal
- Quantidade de regras utilizadas na classificação dos dados: 21
- Quantidade de fraudadores identificados: 9/16 (56.25%)
- Quantidade de consumidores acusados indevidamente de fraude: 5/27 (18.52%)
- Taxa de acerto na inspeção técnica realizada: 9/14 (64.28%)

Portanto, pode-se concluir, com base nos estudos desenvolvidos neste trabalho, que considerar a consistência $\delta = 0.5$ e $D = \{anormal\}$ e análise semanal resulta em um melhor desempenho na detecção de fraude ou erro de medição em grandes clientes consumidores de energia elétrica.

6.6. Considerações Finais

Neste capítulo foram apresentados diversos resultados obtidos no processo de treinamento e teste, com o objetivo de encontrar um melhor resultado na classificação de fraudadores como anormais. Para alguns casos de fraude, o perfil do consumidor fraudador torna-se muito próximo de comportamentos normais de outros consumidores, causando indiscernibilidade no sistema, e gerando incertezas nas regras de classificação.

Para a classificação de dados considerando o valor nominal do atributo de decisão igual a normal, seleciona-se, nos processos de treinamento e teste, uma maior quantidade de regras para realizar essa tarefa, do que quando o valor nominal da decisão é igual a anormal. Pode-se observar também, através dos resultados, que alguns valores de consistência apresentam melhores resultados quanto ao número de fraudadores encontrados e outros apresentam uma melhor taxa de acerto quanto aos clientes anormais detectados nas inspeções técnicas. Contudo, o melhor resultado obtido no sistema e viável para a detecção de fraude ou erro de medição foi para a análise semanal, $\delta = 0.5$ e $D = \{anormal\}$.

Capítulo 7
Conclusões e Propostas de Trabalhos Futuros

7.1. Conclusões

O objetivo principal deste trabalho foi desenvolver um sistema computacional para detectar fraude ou erro de medição em unidades consumidoras de energia elétrica, atendidas em alta tensão, baseados em dados históricos e dados de tempo real, utilizando a técnica de Rough Sets. Devido os comportamento dos consumidores anormais, para alguns casos de fraudes serem muito parecidos com os comportamentos normais de alguns clientes, gera-se no sistema uma incerteza quanto à anormalidade, o que torna este problema adequado para a aplicação de Rough Sets. Esta teoria oferece ferramentas matemáticas para descobrir padrões ocultos nos dados, identificando dependência parcial ou total dos registros, eliminando dados redundantes.

Os clientes consumidores de energia elétrica, através de seu perfil de consumo e natureza da atividade desenvolvida em suas instalações, são divididos em várias categorias, em função dos níveis de tensão que são alimentados. Para os consumidores enquadrados nas tarifas horo-sazonais, as concessionárias utilizam medidores eletrônicos, com capacidade para o armazenamento, em memória de massa, de informações de consumo ativo e reativo, entre outros parâmetros. As diversas informações fornecidas pelos medidores e as informações contratuais, serviram de estudo para a definição dos atributos que foram utilizados no desenvolvimento desta pesquisa.

Em relação aos atributos condicionais selecionados para execução deste trabalho, levou-se em consideração a forte dependência que se tem entre o número de regras utilizadas para a classificação de dados e o número de atributos utilizados. Outro fato importante a ser relatado é que a relação de indiscernibilidade existente

entre os registros está diretamente relacionada com a quantidade de atributos utilizados no sistema de informação.

Para B com 17 (dezessete) atributos condicionais, os registros ou dados com os mesmos valores nominais encontrados no conjunto de treinamento, contendo 7832 (sete mil oitocentos e trinta e dois) dados, foram agrupados totalizando 841 (oitocentos e quarenta e um) registros diferentes entre si, considerando os atributos condicionais e de decisão. Obteve-se ainda, através da relação de indiscernibilidade, 140 (cento e quarenta) conjunto de classes indiscerníveis e 561 (quinhentos e sessenta e um) conjuntos de classes discerníveis. Vale lembrar que cada classe formada pela relação de indiscernibilidade gera uma regra de classificação. Portanto, para o conjunto de atributos formado por B, tem-se um total de 701 (setecentos e uma) regras.

Considerando agora, o conjunto de atributos utilizados na metodologia desenvolvida, ou seja, B com 15 (quinze) atributos condicionais, obteve-se um conjunto de 99 (noventa e nove) classes indiscerníveis e 204 (duzentos e quatro) classes discerníveis para o mesmo conjunto de dados de treinamento, resultando em 303 (trezentos e três) regras para a classificação dos dados.

Os resultados para os dois conjuntos de B, quando comparados com os resultados de classificações corretas de clientes fraudadores como anormais, não sofreram alteração significativa. Logo, optou-se pelo conjunto de 15 (quinze) atributos condicionais, pois reduzir a quantidade de regras na utilização da classificação dos dados, representa reduzir custo e tempo computacional na implementação do sistema desenvolvido na detecção de fraude ou erro de medição.

O sistema desenvolvido empregando a metodologia apresentada, mostrou-se capaz e eficiente para ajudar a solucionar problemas das perdas comerciais relacionadas à fraude ou erro de medição nas concessionárias de energia elétrica.

Os resultados são considerados satisfatórios, uma vez que a taxa de acerto na identificação de fraude obtida pelo sistema, utilizando-se análise semanal na unidade consumidora, a partir da pré-seleção dos consumidores com suspeita de fraude foi de 64,7%.

Os resultados obtidos através da classificação dos comportamentos dos clientes em normal e anormal, utilizando-se análise semanal, foram considerados satisfatórios, uma vez que foi possível localizar a maior parte dos consumidores anormais com baixo índice de acusação indevida de fraude (cerca de 64,70% dos fraudadores utilizados no processo de teste foram localizados e apenas 22.22% dos clientes normais foram acusados indevidamente de anormais). De acordo com os resultados apresentados no Estudo de Caso – Capitulo 6, pode-se concluir , para o banco de dados utilizado no desenvolvimento do trabalho, que para o valor de consistência igual a 0.5, atributo de decisão igual a anormal e análise semanal, obteve-se o melhor desempenho na classificação dos fraudadores.

7.2. Trabalhos Futuros

Como trabalhos futuros, pretende-se continuar a análise de outros atributos que não foram selecionados no sistema de informação inicialmente, como por exemplo o atributo referente à falta de energia na unidade consumidora, que foi desprezado devido aos fraudadores utilizados no estudo terem sido criados, não apresentando ausência de energia no seu histórico.

Alguns medidores de energia apresentam um número maior de “canais de medição” portanto, podem dispor de informações do tipo corrente e tensão nas memórias de massa por exemplo. Estas informações devem ser exploradas pois podem apresentar resultados significativos na análise de fraude.

Uma informação que não foi levada em consideração no desenvolvimento do sistema computacional, devido à ausência de dados, é quanto aos dois períodos do ano: Período Seco - cinco meses consecutivos (de dezembro de um ano a abril do ano seguinte) e Período Úmido - sete meses consecutivos (de maio a novembro). Conforme a atividade exercida pelo consumidor, estes dois períodos podem ter comportamentos diferentes quanto aos dias da semana. Pretende-se dividir o sistema e traçar dois perfis, com base nos dados de consumo e demanda, para cada unidade: perfis de período seco e perfis de período úmido.

Inserir um Calendário de Feriados (Nacionais, Estaduais e Municipais), que abranja todas as localizações de consumidores do Grupo A, será uma etapa importante em trabalhos futuros. Estas informações são importantes na etapa de consolidação dos dados pois elimina os registros nos dias de feriados, no processo de treinamento, na criação do perfil de cada consumidor. Já nos resultados de classificação dos consumidores, pode-se verificar as datas das acusações indevidas de anormalidade, pois na maioria das unidades consumidoras, nos feriados, o consumo e demanda de energia diminuem significativamente. No desenvolvimento do sistema computacional com a metodologia proposta, implementou-se apenas os feriados nacionais, pois as informações referentes às localidades de cada cliente utilizado neste trabalho, foram omitidas pela concessionária que cedeu as memórias de massa e dados contratuais.

Pretende-se ainda, desenvolver o sistema de detecção de fraude ou erro de medição, em ambiente Delphi, para poder ser utilizado em qualquer concessionária de energia que tenha por objetivo diminuir suas perdas comerciais.

7.3. Artigos Submetidos e Aceitos

Foram submetidos e aceitos dois artigos no XXVIII CNMAC – Congresso Nacional de Matemática Aplicada e Computacional, a ser realizado em São Paulo no Centro de Convenções do Campus – SENAC, de 12 a 15 de setembro de 2005. Os títulos dos artigos são:

- Rough Sets – Técnica de redução de atributos e geração de regras para classificação de dados;
- Sistema de detecção de fraude em grandes consumidores de energia elétrica, utilizando a teoria de Rough Sets, baseado em dados dinâmicos e estáticos.

Referências Bibliográficas

[1] GHOSH, S.; REILLY, D. L. Credit card fraud detection with a neural-network. System Sciences, 1994. Information Systems: Decision Support and Knowledge-Bases System, Proceedings of the Twenty-Seventh Hawaii International Conference, v.3, p. 621-630, 1994.

[2] BRAUSE, R.; LANGSDORF, T.; HEPP, M. Neural data minig for credit card fraud detection. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference, p. 103-106, 1999.

[3] TANIGUCHI, M.; HAFT, M.; HOLLMEN, J.; TRESP, V. Fraud detection in communication networks using neural and probabilistic methods. Acoustics, Speech, and Signal Processing, 1998. ICASSP'98. Proceedings of the 1998 IEEE International Conference, v.2 p. 1241-1244, 1998.

[4] REDNER, R. A.; WALKER, H. F. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 26(2): 195-234, 1984.

[5] BURGE, P.; SHAWE-TAYLOR, J.; COOKE, C.; MOREAU, Y.; PRENEEL, B.; STOERMANN, C. Fraud detection and management in mobile telecommunications networks. Security and Detection, 1997. ECOS 97, European Conference, p. 91-96,, 1997.

[6] CABRAL, J. E.; PINTO, J. O. P.; GONTIJO, E. M.; REIS, J. Rough Sets Based Fraud Detection in Electrical Energy Consumers. WSEAS International Conference on MATHEMATICS AND COMPUTERS IN PHYSICS, Cancun, Mexico, Apr. 2004, 2 ed.,v.3, p. 413-416.

[7] CABRAL, J. E.; PINTO, J. O. P.; GONTIJO, E. M.; REIS, J. Fraud detection in electrical energy consumers using rough sets. In: 2004 IEEE International Conference on Systems, Man, and Cybernetics., p. 3625–3629, 2004.

[8] REIS, J.; GONTIJO, E. M.; MAZINA, E.; CABRAL, J. E.; PINTO, J. O. P. Fraud identification in electricity company customers using decision tree. In: 2004 IEEE International Conference on Systems, Man, and Cybernetics, p. 3730–3734, 2004.

[9] BREIMAN, L.; FRIEDMAN, J. H.; OHLSEN, R. A.; STONE, C. J. Classification and regression trees. Chapman & Hall/CRC, 1993.

[10] Resolução Normativa Nº 456 da ANEEL – Agência Nacional de Energia Elétrica, de 29 de novembro de 2000. http:// www.aneel.gov.br/ccdoc/res2000456.pdf

[11] LINARES, K. S. C. Aspectos Teóricos do Datamining: Descoberta de Conhecimento em Medicina. Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis (Brasil), 2003.

[12] BARRETO, J. M. Inteligência Artificial no Limiar do século XXI, 3a. edição ed. $\rho\rho\rho$ Edições, Florianópolis-Brasil, 2001.

[13] MANNILA, H. Methods and problems in data mining (a tutorial). In *Proceedings of International Conference on Database Theory (ICDT'97)*, January 1997, Delphi-Greece, F. Afrati & P. Kolaitis, Eds., Springer-Verlag, p. 41-55, 1997.

[14] HOLSHEIMER, M.; SIBES, A. Descoberta do conhecimento em base de imagens mamográficas. IX Congresso Brasileiro de Informática em Saúde – CBIS'2004, Ribeirão Preto - São Paulo, 2004.

[15] ROMÃO, W.; PACHECO, R. C. S.; NIEDERAUER, C. A. P. Planejamento em C&T: uma Abordagem para Descoberta de Conhecimento Relevante em Banco de Dados de Grupos de Pesquisa. Revista Tecnológica, UEM - Maringá - Paraná, v. 9, p. 139-152, 2000.

[16] KLÖSGEN, W.; ZYTKOW, J. M. Knowledge Discovery in Databases Terminology. In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Editors, MIT Press, p.573-593 , 1996.

[17] HOLSHEIMER, M.; SIBES, A. Data mining: The search for knowledge in database. Relatório técnico., CWI – The Netherlands, 1998, Amsterdam - The Netherlands.

[18] FAYYAD, U; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery AAAI Press/The MIT Press, Usama M. Fayyad and Gregory Piatetsky-Shapiro and Smyth Padhraic and Ramasamy Uthurusamy Editors, 1996, ch. 1, p 1-34.

[19] SILVER, D. Knowledge discovery and data mining. Relatório técnico., CogNona Technologies, London Health Science Center, 1998.

[20] PACHECO, M.A.; VELLASCO, M.; LOPES, C. H. Decoberta do Conhecimento e Mineração de Dados. Relatório técnico., Laboratório de Inteligência Computacional, Aplicada, DEE, PUC – Rio de Janeiro, 1999.

[21] FAGIN, U,; VARDI, M. Y. *The theory of data dependencies a survey*. Mathematics of Information Processing, Americam Mathematical Society, M. Anshel and W. Gewirtz Editors, 1986, p. 19-71.

[22] QUINLAN, J. R. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[23] PAWLAZ, Z. Rough sets. International Jornal of Computer and Information Sciences, p. 341-356, 1982.

[24] PAWLAZ, Z. Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, 1991.

[25] OHRN, A. Rosetta. Technical reference manual. Relatório técnico., Knowledge Systems Group. Department of Computer and Information Science. Norwegian University of Science and Technology, Norway, 1999.

[26] SLOWINSKI, R. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, 1992.

[27] SHAN, N.; ZIARKO, W.; HAMILTON, H. J.; CERCONE, N. Using rough sets as tool for knowledge discovery. International Conference on Knowledge Discovery and Data Mining, p.263-268, 1995.