UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

DESENVOLVIMENTO DE UM SISTEMA PARA CLASSIFICAÇÃO DE ANORMALIDADES NO CONSUMO DE ENERGIA ELÉTRICA

EDUARDO WERLEY SILVA DOS ÂNGELOS

São Luís – MA
2009

DESENVOLVIMENTO DE UM SISTEMA PARA CLASSIFICAÇÃO DE ANORMALIDADES NO CONSUMO DE ENERGIA ELÉTRICA

Dissertação de mestrado submetida à Coordenação do Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos para obtenção do título de mestre em Engenharia Elétrica na área de Sistemas de Energia.

Por

Eduardo Werley Silva dos Ângelos.

São Luís – MA
2009

Ângelos, Eduardo Werley Silva dos

Desenvolvimento de um sistema para classificação de anormalidades no consumo de energia elétrica/ Eduardo Werley Silva dos Ângelos. – São Luís, 2009.
100 f.

Orientador: Prof. Dr. Osvaldo Ronald Saavedra Mendez
Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, 2009.

1. Energia elétrica – consumo - anormalidades. 2. Sistema especialista FUZZY. 3. Mineração de dados. I. Título.

CDU 621.31:004.891

DESENVOLVIMENTO DE UM SISTEMA PARA CLASSIFICAÇÃO DE ANORMALIDADES NO CONSUMO DE ENERGIA ELÉTRICA

Eduardo Werley Silva dos Ângelos

Dissertação aprovada em 07 de agosto de 2009.

Prof. Osvaldo Ronald Saavedra Mendez, Dr.
(Orientador)

Prof. Omar Andres Carmona Cortes, Dr.
(Co-orientador)

Prof. Eduardo Nobuhiro Asada, Dr.
(Membro da Banca Examinadora)

Prof. Sofiane Labidi, Dr.
(Membro da Banca Examinadora)

AGRADECIMENTOS

À minha mãe, Ana Lúcia, ao meu pai, Josué e aos demais familiares.

À minha querida Sonia e demais amigos próximos.

Ao meu orientador, professor Osvaldo Saavedra, pelo compartilhar das suas muitas experiências e por seu sábio direcionamento ao longo destes anos.

Ao meu co-orientador, professor Omar, por haver levado esta pesquisa a novos horizontes e por sua frutífera parceria.

Aos professores do Grupo de Sistemas de Potência do Programa de Pós-Graduação em Engenharia de Eletricidade da UFMA, pelos valiosos conhecimentos repassados – professores Leonardo Paucar, Maria da Guia e José Pessanha.

Aos meus caros colegas do Laboratório de Sistemas de Potência – Carlos Cesar, Sérgio, Alan, Yuri, Alex, Carlos Portugal, Gabriel, Igor, Cláudio e Tayssara – pelos incontáveis e agradáveis momentos de convivência.

Ao CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo incentivo financeiro.

Sobretudo Ao grande *Eu Sou*, pela constante inspiração e proximidade.

"... Os impossíveis dos homens são possíveis para Deus". Lc 18.26

RESUMO

Este trabalho apresenta uma metodologia computacional para classificação de perfis anormais de consumo de energia elétrica. A abordagem parte da premissa que um dado cliente deve permanecer o mais próximo possível de seu padrão de consumo histórico, sendo que os desvios do padrão registrado representam possíveis fraudes de energia ou irregularidades de medição. A parte inicial da metodologia — busca de consumidores com perfis de consumo semelhantes — é efetuada por meio da técnica computacional de clusterização fuzzy. Já a tarefa de mensurar o desvio do padrão histórico é realizada por meio de uma metodologia de classificação nebulosa, baseada na matriz de partição fuzzy e distância dos elementos aos centros dos agrupamentos. Por fim, as distâncias para os grupos são normalizadas, gerando um índice no intervalo unitário, sendo que os elementos de maior chance de estarem irregular são aqueles com índices mais próximos de um. A metodologia foi validada com uma base de dados de uma concessionária local. Os resultados alcançados foram satisfatórios, sendo obtida adequada performance tanto no processo de detecção de fraudes quanto irregularidades na medição.

Palavras-Chave: Perdas Comerciais. Fraude de Energia. Mineração de Dados. Clusterização Fuzzy.

ABSTRACT

This work proposes a computational technique for classification of electricity consumption profiles. The approach is based on the assumption that it's possible to find out groups of consumers with similar patterns of energy use. So, given the found groups, which can be also viewed as a normal consumption profile, ones can associate a high chance of fraud or abnormality to that consumers lying more apart from the groups. The methodology comprises two steps. A fuzzy clustering c-means-based is done in order to search for consumers with similar consumption profiles, in the first one. Afterwards, a fuzzy classification is performed using a fuzzy membership matrix and the Euclidian distance to the cluster centers. Then, the distance measures are normalized and ordered, yielding an unitary index score, where the possible fraudulent or abnormal consumers are those with the higher scores. The approach was tested and validated with real data base, showing good performance in both fraud and metering defect detection tasks.

Keywords: Non-technical losses. Electricity Theft. Data Mining. Fuzzy Clustering.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 FRAUDES E ANORMALIDADES NA SOCIEDADE E NO SETOR ELÉTRICO

Fraude é uma prática cuja origem remonta ao início da própria civilização humana, responsável por consideráveis prejuízos na sociedade como um todo, sobretudo econômicos, com ocorrência cada vez mais crescente nos dias atuais. Em certo ponto, fraude pode ser vista como um tipo de padrão anômalo, ou anormalidade, isto é, um comportamento que se destoa consideravelmente do universo onde se encontra de forma a levantar suspeitas sobre sua autenticidade. Nesse sentido, podem também ser considerados padrões anormais em um dado processo: erros de avaliação, perfis incomuns, desvios internos, distúrbios intrínsecos e ações afins – que ocasionam impactos tão destrutivos quanto fraudes, não obstante possuírem, muitas vezes, motivações distintas.

O combate a anormalidades é uma atividade por demais dispendiosa, demandando tempo, dedicação e experiência. Em setores historicamente propícios à existência de fraudes, como empresas de telecomunicações, de cartões de crédito, redes de computadores, de energia elétrica, dentre outros, duas estratégias principais têm sido usadas para sua minimização: a prevenção e a detecção (SUNDARAM, 1996). A primeira, diz respeito a iniciativas de proteção dos mecanismos de prestação dos serviços, bem como alternativas para conscientização dos usuários sobre os malefícios decorrentes da prática de fraudes. A segunda, que vem sendo fortemente potencializada por técnicas computacionais inteligentes, fundamenta-se no monitoramento dos perfis de uso dos serviços concedidos, para se estimar, detectar ou prevenir comportamentos indesejáveis (KOU *et al.*, 2004).

Em sistemas de energia elétrica, entende-se por consumo anormal ou irregular aquele relacionado a fraudes, erros de medição, consumos não faturados, problemas na instalação do consumidor, ligações clandestinas, inadimplência e fenômenos afins – fatores que resultam nas chamadas *perdas comerciais*, isto é, déficits financeiros relacionados a certo montante de energia consumida, mas não faturada. Relata-se que em 2007, somente as perdas comerciais

no sistema elétrico brasileiro foram da ordem de 6% da energia ofertada, ou seja, 15 TW.h, equivalente a cerca de 10 bilhões de reais, utilizando o preço médio de aquisição de energia pelas distribuidoras, já considerando impostos (BRASIL, 2007).

As conseqüências do consumo irregular de energia manifestam-se fortemente através de sobrecargas no sistema, no decremento da qualidade do serviço prestado e, invariavelmente, no aumento das tarifas, visto que no atual modelo regulatório as perdas são oneradas na composição do preço final da energia repassado ao consumidor.

A principal estratégia usada pelas concessionárias para redução das perdas comerciais é a inspeção de campo. Cometti (2005) menciona três fatores que usualmente originam incursões em campo: varreduras por regiões, denúncias e análise do histórico de consumo. Ocorre que a realização de inspeções motivadas por tais fatores não têm atingido os níveis de eficiência desejados, de forma a cobrir os elevados custos operacionais do processo. Como forma de contornar o problema, recentemente vem sendo aplicadas técnicas computacionais de estatística, inteligência artificial, mineração de dados, ou a combinação entre estas, fundamentadas na análise de dados históricos em busca de padrões ou perfis que possam ser associados à fraude ou irregularidades. Através da pré-identificação dos consumidores cujos perfis de consumo estejam mais próximos de uma situação de fraude, pode-se otimizar a quantidade de equipes de campo, maximizando por fim o número de irregularidades descobertas.

1.2 JUSTIFICATIVA

A despeito das inúmeras formas de se fraudar o consumo de energia elétrica, grande parte delas pode ser rotulada em termos de padrões bem-definidos. A existência de perfis característicos traz forte embasamento para a aplicação da técnica de clusterização – processo computacional que visa a formação de agrupamentos, cujos elementos de um mesmo grupo sejam os mais semelhantes possíveis e os elementos entre os grupos distintos sejam os mais diferentes possíveis.

A teoria de agrupamentos, geralmente utilizada como uma primeira etapa para execução de algoritmos de classificação, vem obtendo bons resultados quando aplicada a problemas de detecção de fraudes (ROCHA, 2002; CALILI, 2005; LAZO, 2005). Neste trabalho será desenvolvido um modelo não-supervisionado fundamentado no processo de

agrupamentos, porém apto à identificação de irregularidades no domínio dos sistemas de distribuição e viável à aplicação em sistemas reais.

Há que se ressaltar também a afirmação feita no trabalho de Bolton and Hand (2002), onde se menciona que "*a modelagem para detecção de fraudes é, no mínimo, um difícil problema de se estimar a probabilidade de pertinência de elementos em classes, ao invés de um simples problema de classificação*". Em citações como essa, pode ser vislumbrada a aplicação da Lógica *Fuzzy* a problemas de identificação de fraudes; aplicabilidade que é comprovada nos trabalhos de Phuong *et al.* (2002), Pathak *et al.* (2005), Thang *et al.* (2006), Lazo *et al.* (2005) e Dos-Angelos *et al.* (2007), sendo estes dois últimos voltados à detecção de fraudes no consumo de energia. Portanto, uma das premissas do presente trabalho é o desenvolvimento de uma metodologia pautada na Lógica *Fuzzy*, tanto no processo inicial de clusterização quanto no processo de classificação do perfil de consumo de novos clientes.

1.3 OBJETIVOS

1.3.1 GERAL

Desenvolver um sistema para classificação do perfil de consumo dos usuários do sistema de distribuição de energia elétrica de fácil aplicabilidade e com um grau adequado de assertividade, de forma a incrementar a eficiência das inspeções de campo em busca de irregularidades.

1.3.2 ESPECÍFICOS

- Reavaliar a importância do processo de *Descoberta de Conhecimento em Bases de Dados* no desenvolvimento de sistemas para detecção de irregularidades;
- Aprofundar o conhecimento sobre a aplicabilidade dos agrupamentos *Fuzzy* à classificação de padrões de consumo irregulares em Sistemas de Distribuição de Energia;
- Validar o sistema proposto com uma base de dados reais;

- Introduzir novo conhecimento decorrente dos estudos de caso para melhorar o grau de assertividade do sistema proposto.

1.4 ESTRUTURA DO TRABALHO

Para um melhor entendimento do assunto proposto, este trabalho está organizado da seguinte forma: neste **primeiro** capítulo efetuou-se a introdução ao tema, no que diz respeito a suas generalidades, formulação do problema, objetivos e justificativas.

No Capítulo **2** aborda-se o problema de fraude e irregularidades tanto em âmbitos gerais como na questão específica do consumo de energia elétrica. Ainda no Capítulo 2 é discutido o processo de identificação de irregularidades e suas peculiaridades.

No Capítulo **3** contemplam-se os princípios fundamentais do processo de Descoberta de Conhecimento em Bases de Dados, seus conceitos, técnicas e aplicações. Também aborda-se o conceito de Clusterização, enfocando a análise dos clusters do tipo *Fuzzy*.

No Capítulo **4** realiza-se a explanação da metodologia desenvolvida – sua justificativa e formulação. A caracterização da metodologia se dá no âmbito da Descoberta de Conhecimento em Bases de Dados, onde se descreve suas etapas e os requisitos para sua aplicação.

Já no Capítulo **5**, analisa-se sua aplicação a um sistema real, seus resultados e discussão sobre seu desempenho. Por fim, no Capítulo **6** são tecidas as considerações finais do trabalho, contribuições e possíveis trabalhos futuros.

2 ANORMALIDADES

2.1 INTRODUÇÃO

Neste capítulo discorre-se sobre a problemática das anormalidades no consumo de energia elétrica. O termo “anormalidades” aqui se refere a qualquer tipo de padrão anormal de uso da energia elétrica, tanto aqueles decorrentes de ações fraudulentas de consumidores quanto os provocados por defeitos ou propriedades dos dispositivos da rede elétrica.

No decorrer do capítulo analisa-se o processo de identificação de anormalidades, com suas duas abordagens principais: supervisionadas e não-supervisionadas. É descrito, também, o fenômeno das anormalidades no consumo de energia elétrica, focando-se a parcela referente às chamadas perdas comerciais, associadas à prática de fraude, furto de energia e fenômenos afins. Analisa-se a relevância das perdas comerciais a nível mundial e nacional, discorrendo-se sobre sua segmentação, estratificação e levantando-se as principais causas do fenômeno, que serão de particular importância na definição da estrutura do sistema inteligente desenvolvido neste trabalho.

Por fim, descrevem-se as principais formas de prevenção e combate a anormalidades no consumo de energia, enfatizando o papel das estratégias computacionais inteligentes na solução da problemática.

2.2 DEFINIÇÕES

Fraude pode ser entendida como qualquer atitude proposital e deliberada para se usufruir de vantagens pessoais mediante a utilização inapropriada de um bem ou serviço. Neste contexto, podem ser correlacionados: invasões em redes de computadores, desvios em operações de crédito, fraudes em serviços de telecomunicações, de distribuição de energia elétrica e de fornecimento de água, fraudes em operações de seguros e demais ações

semelhantes. Nesses setores, fraudes respondem por prejuízos bilionários, afetando não somente os lucros empresariais, mas toda a cadeia de produção e distribuição dos serviços.

O escopo geral do que realmente se constitui fraude nem sempre está perfeitamente definido. Um problema em particular surge na diferenciação de fraudes e outras perdas originadas por deficiências administrativas, negligência, erros em procedimentos internos, falhas de gestão ou riscos de negócios. Tais ações acarretam prejuízos tão relevantes quanto os advindos por fraudes em si, porém se encontram em universos diferentes, devendo ser consideradas apenas como perfis irregulares. Contudo, devido a seus impactos similares, a problemática geral de fraudes e irregularidades tem sido tratada como um fenômeno científico comum (SUNDARAM, 1996), sendo, portanto, referenciada neste texto simplesmente como *anormalidades.*

2.3 PROCESSO DE IDENTIFICAÇÃO DE ANORMALIDADES

Com a difusão das tecnologias de manipulação e armazenamento de dados em massa, órgãos prestadores de serviços têm acumulado uma quantia infindável de informações sobre o comportamento e trajetória do uso dos seus recursos. A premissa principal dos sistemas computacionais de detecção de anormalidades é que, a partir da análise de dados históricos, é possível encontrar e formular padrões associados a irregularidades, de forma a prevenir e identificar futuras fraudes, maximizando as corretas predições e mantendo as incorretas em um nível aceitável.

A análise e detecção de fraudes é uma tarefa fortemente restrita a cada domínio de aplicação, requerendo, muitas vezes, conhecimentos sócio-econômicos, de legislação e de mercado específicos. O fenômeno, no entanto, se constitui de instâncias similares, com métodos que se perpetuam diversas vezes, o que dá margem às empresas para que possam se precaver quanto ao processo e suas conseqüências.

Os métodos inteligentes para identificação de anormalidades podem ser estatísticos (clássicos) ou baseados em inteligência artificial (BOLTON; HAND, 2002). Qualquer tipo de metodologia, entretanto, pode ser desenvolvida seguindo duas estratégias principais – supervisionadas ou não-supervisionadas – tratadas a seguir.

2.3.1 ABORDAGENS SUPERVISIONADAS

São fundamentadas na hipótese que “há uma forma de se representar comportamentos anormais partindo de um padrão ou assinatura” (SUNDARAM, 1996). Nestas estratégias, inicialmente há um processo de treinamento onde amostras de casos anormais e legítimos são usadas para a definição de classes normais ou anormais, por intermédio de um modelo supervisor. Posteriormente, novos casos são então classificados (preditos), tendo como base os grupos pré-definidos, conforme ilustrado na Figura 2.3.1.

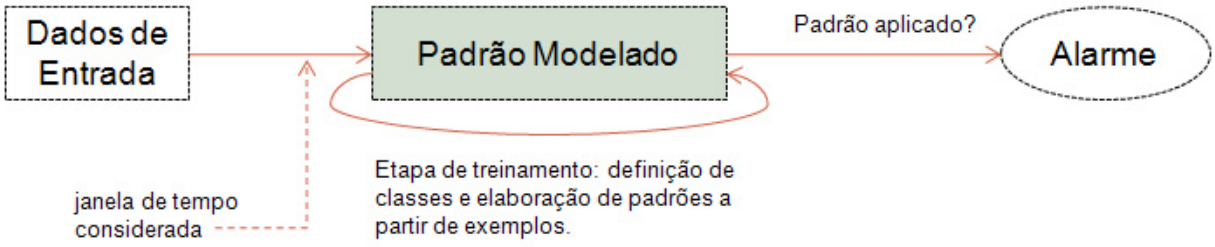


Figura 2.3.1 – Diagrama de blocos de uma estratégia supervisionada

Nas abordagens supervisionadas, padrões existentes nas classes de treinamento associados a anormalidades poderão ser identificados posteriormente em novos casos classificados, entretanto, não há garantia que perfis não existentes sejam identificados. Este fato representa a maior desvantagem dos modelos supervisionados, ou seja, a dificuldade em se estabelecer um padrão geral que contemple todos os casos previstos de anormalidades.

O conceito de método supervisionado é compartilhado no domínio de Aprendizado de Máquina (do inglês, *Machine Learning),* quando referente à aprendizagem do tipo **supervisionada**. Nesse tipo de abordagem, também conhecida como aprendizagem por exemplos, é estabelecida uma referência para o comportamento a ser atingido pelo sistema, de forma que para cada conjunto de atributos de entrada é pré-associada uma saída correspondente. Em outras palavras, o problema caracteriza o aprendizado de uma função através de amostras de sua entrada e saída, sendo denominado *classificação*, para rótulos discretos e *regressão*, para valores contínuos. Nesse sentido, o desenvolvimento de um

sistema supervisionado para classificação de anormalidades pode ser associado às seguintes etapas:

I. Aquisição de um volume considerável de amostras das classes pré-definidas;

II. Definição dos conjuntos de treinamento e de testes;

III. Treinamento e sintonização da função de mapeamento entrada-saída, utilizando o conjunto de exemplos;

IV. Avaliação da assertividade do modelo utilizando o conjunto de testes;

V. Repetição dos passos anteriores para outros conjuntos de treinamento e testes.

As principais técnicas supervisionadas aplicadas a problemas de identificação de fraudes são: técnicas estatísticas tradicionais, como Análise de Discriminantes Lineares e Discriminação Logística (HAND, 1981; MCLACHLAN, 1992); Redes Neurais; Redes Bayesianas; Árvores de Decisão; Sistemas Especialistas e outros modelos baseados em regras.

2.3.2 ABORDAGENS NÃO-SUPERVISIONADAS

Nestes métodos, procura-se estabelecer dinamicamente uma distribuição de referência de comportamento, de forma que seja possível a partir dela identificar elementos que mais se desviam. Assim, não se faz necessário o uso de classes pré-definidas (amostras de casos anormais e legítimos), ocorrendo, ao invés disso, apenas a detecção de mudanças em relação à norma estabelecida, conforme ilustrado na Figura 2.3.2.

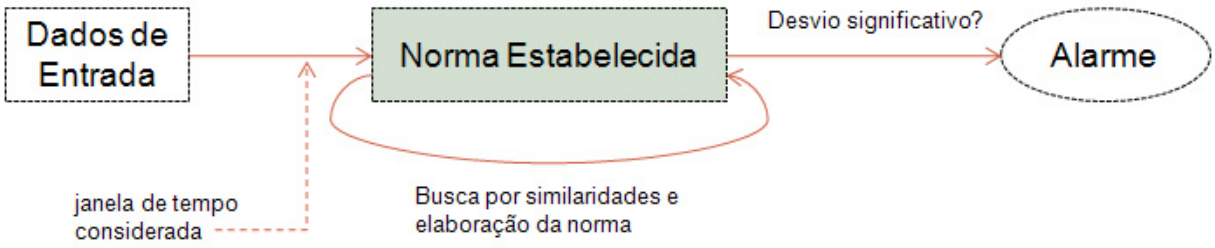


Figura 2.3.2 – Diagrama de blocos de um sistema não-supervisionado.

Uma das desvantagens associada a tais abordagens é o grande esforço computacional demandado, em termos de geração de padrões, buscas ao banco de dados e outras tarefas (SUNDARAM, 1996). Este fator se torna de menor relevância diante da capacidade de processamento dos computadores atuais, mas pode exercer grande influência em sistemas de detecção de anormalidades em tempo real. Outro ponto preponderante é a definição do nível de corte além do qual um dado perfil analisado seja considerado como anormal; um limiar abaixo do ideal pode acarretar o aumento de falsas predições, mas acima do adequado pode ocasionar uma redução de desempenho. Por outro lado, dada a inexistência de um elemento supervisor, pode-se deduzir que estratégias não-supervisionadas são mais factíveis de serem generalizadas, meta muito aclamada para o desenvolvimento de técnicas de detecção de fraudes (BOLTON; HAND, 2002).

Sistemas não-supervisionados de identificação de anormalidades estão relacionados à aprendizagem do tipo não-supervisionada, que envolve a aprendizagem de padrões a partir de uma determinada entrada quando não são apresentados valores de saída específicos. Geralmente, este tipo de aprendizado é usado para encontrar conjuntos de dados com características semelhantes, onde as técnicas mais utilizadas são: *C-Means* (também denominado *K-Means*), *Fuzzy C-Means* (FCM), K-Vizinhos mais Próximos (do inglês, *K-Nearest Neighbor – KNN*) e Mapa Auto-Organizável de Kohonen (do inglês, *Self-Organizing Map – SOM*).

2.4 ANORMALIDADES NO CONSUMO DE ENERGIA ELÉTRICA

Durante o processo de geração, transmissão e distribuição de energia elétrica, invariavelmente ocorrem perdas de energia. O fenômeno se manifesta desde elementos condutores, como linhas e cabos, elementos transformadores de tensão até diversos dispositivos de manobra e proteção, sendo, de certa forma, um processo físico natural. O conhecimento do nível de perdas por parte das concessionárias de energia é um fator de capital importância para o correto planejamento e dimensionamento dos sistemas de distribuição, bem como para a manutenção da qualidade da energia ofertada.

Em sistemas de energia elétrica, as perdas possíveis de serem estimadas, associadas a fenômenos físicos no transporte e armazenamento de energia são denominadas de **perdas técnicas**, definidas pela Agência Nacional de Energia Elétrica (ANEEL), em nível de distribuição, nos seguintes termos:

> (...) o montante de energia elétrica, expresso em megawatt-hora por ano (MW.h/ano), dissipado entre o suprimento e o ponto de entrega, decorrente das leis físicas relativas aos processos de transporte e transformação de tensão, mais as perdas na medição de energia elétrica da unidade consumidora de responsabilidade da distribuidora; corresponde à soma de duas parcelas, incluindo as perdas por efeito joule e por efeito corona nos cabos, condutores, ramais, medidores, conexões, sistemas supervisórios, relés fotoelétricos, capacitores, transformadores de corrente e de potencial e as devidas a fugas de correntes em isoladores e pára-raios, além da referente às perdas nos transformadores (no ferro ou em vazio e nos condutores ou no cobre), ambas decorrentes exclusivamente da energia fornecida às unidades consumidoras regulares, outras concessionárias e ao consumo próprio. (ARAÚJO; SIQUEIRA, 2006, p. 69)

A priori, os montantes de perdas em uma determinada área de concessão são conhecidos, sendo melhor estimados quanto mais bem estruturada for a distribuidora, em termos de elementos de medição e do conhecimento da estrutura de distribuição do sistema – localização de alimentadores, carga dos transformadores, extensão dos ramais, etc. Ainda assim, a real magnitude das perdas de energia projeta-se além dos montantes previstos pela maioria das concessionárias, onde se deduz que há um consumo de energia paralelo à rede estabelecida, desconhecido pela empresa, mas de conseqüências notáveis e catastróficas para o equilíbrio do sistema como um todo. A energia adicional para suprir as cargas não previstas é, de fato, consumida, mas não é faturada, ocasionando perdas expressivas nas receitas das distribuidoras, referenciadas pelo termo **perdas não-técnicas.** O cálculo da parcela de energia referente a tais perdas é dado pela seguinte expressão:

$$E_{PNT} = E_{OUT} - E_{FAT} - E_{PT} \tag{2.1.1}$$

Onde:

E_{PNT}: energia referente à perdas não-técnicas [kWh];
E_{OUT}: energia estimada na saída do alimentador [kWh];
E_{FAT}: energia faturada [kWh];
E_{PT}: energia referente à perdas técnicas [kWh];

O somatório das perdas técnicas e não-técnicas, $E_{PNT} + E_{PT} = E_{OUT} - E_{FAT}$, é denominado **perdas globais**.

2.4.1 PERDAS GLOBAIS NO MUNDO E NO BRASIL

Perdas de energia é um assunto de notável importância nos órgãos gestores do setor energético. Toda a cadeia produtiva das empresas de energia depende de uma forma correta de lidar com o problema, já que a própria sustentabilidade das concessionárias é influenciada. Estatísticas comprovam que diferentes visões do problema, em diferentes países, levam a distintos níveis de perdas globais, como mostrado na Tabela 2.4.1(DOS-ANGELOS *et al.*, 2007) .

Tabela 2.4.1 – Índices de perdas globais na distribuição em países selecionados.

PAIS	1980	1990	2000	2004
Finlândia	6,2	4,8	3,7	3,4
Holanda	4,7	4,2	4,2	4,3
Coréia do Sul	6,6	5,6	4,7	4,4
Dinamarca	9,3	8,8	7,1	4,4
Japão	5,8	5,7	5,4	4,5
Bélgica	6,5	6,0	4,8	4,7
Áustria	7,9	6,9	7,8	4,7
Alemanha	5,3	5,2	5,1	5,5
França	6,9	9,0	7,8	5,6
Suíça	9,1	7,0	7,4	6,1
Austrália	11,6	8,4	9,1	6,1
Estados Unidos	10,5	10,5	7,1	6,4
Canadá	10,6	8,2	9,9	6,6
Itália	10,4	7,5	7,0	6,9
Suécia	9,8	7,6	9,1	7,2
Irlanda	12,8	10,9	9,9	8,0
Noruega	9,5	7,1	9,8	8,0
Reino Unido	9,2	8,9	9,4	8,1
Espanha	11,1	11,1	10,6	8,6
Portugal	13,3	9,8	9,4	9,0
Nova Zelândia	14,4	13,3	11,5	13,2

Mesmo levando em conta as diferentes configurações e parâmetros das redes elétricas de cada país, tem-se observado, via de regra, que as perdas de energia são maiores nas nações em desenvolvimento. Este fato está exemplificado na Tabela 2.4.2, onde se ressalta que as perdas na distribuição concentram-se em patamares inferiores a sete por cento nos países membros da OCDE – organização para a cooperação e desenvolvimento econômico (do inglês, OECD) – que reúne majoritariamente países desenvolvidos.

Tabela 2.4.2 – Perdas na distribuição por blocos geopolíticos

BLOCO	Energia Gerada (GW.h)	Perdas (GW.h)	Perdas (%)
Pacífico (OECD)	1.729.598	81.785	4,73
China e Hong Kong (OECD)	2.236.733	143.614	6,42
América do Norte (OECD)	4.997.072	339.937	6,80
Europa (OECD)	3.468.931	242.898	7,00
África	539.427	59.758	11,08
Oriente Médio	588.196	65.631	11,16
Antiga União Soviética	1.379.568	176.490	12,79
Europa (demais países)	197.040	25.775	13,08
América Latina	875.473	146.478	16,73
Ásia (exceto China)	1.518.952	258.361	17,01
MUNDO - TOTAL	**17.530.990**	**1.540.727**	**8,79**

Fonte: INTERNATIONAL ENERGY AGENCY, 2007.

No caso específico brasileiro, o órgão regulador do setor elétrico informa que as perdas totais no país, atualmente, concentram-se no patamar de 16% da energia requerida, conforme ilustrado na Figura 2.4.1 (SACIC, 2007). Observa-se no gráfico um forte crescimento das perdas na década de noventa, motivado possivelmente pelo incipiente investimento no setor elétrico no período em questão (VIEIRALVES, 2005). No ano de 2007, cerca de cinco bilhões de reais referentes às perdas de energia estiveram embutidos nas tarifas de energia – um valor 24% maior que no ano 2003 (BRASIL, 2007).

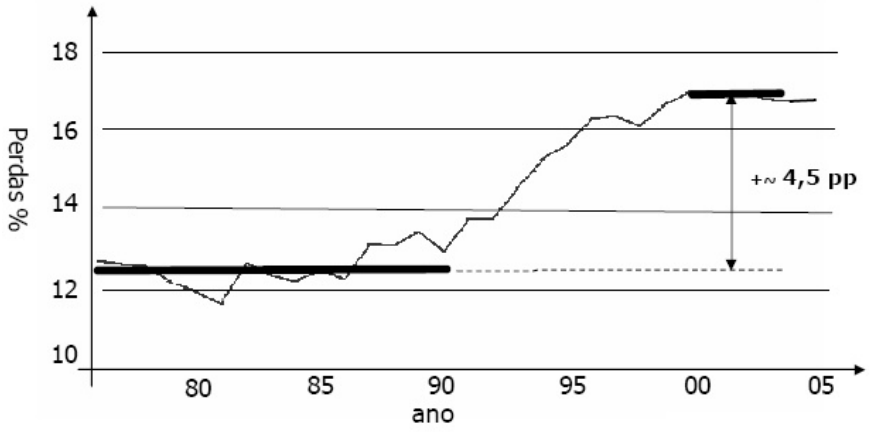


Figura 2.4.1 – Evolução das perdas no Brasil

As perdas de energia no Brasil podem ser consideradas elevadas, em comparação a outros países em desenvolvimento. Em 2004, por exemplo, DOS-ANGELOS *et al.* (2007,

p.26) relata que as perdas no país foram da ordem de 16,85%, índice superior ao verificado no México (15,84%), Rússia (12,08%), Chile (8,01%) e China (6,42%) no mesmo ano. Não obstante, a distribuição das perdas no Brasil não é a mesma ao longo das 64 concessionárias de distribuição, sendo as regiões Norte e Nordeste as que possuem os maiores índices (VIEIRALAVES, 2005). Algumas distribuidoras chegam a ter o dobro de perdas comerciais em relação às perdas técnicas (REIS, 2005).

Outro ponto a ser considerado é sobre a segmentação das perdas de energia no país. Guimarães (2004) afirma que, do montante de 62.788 GW.h de energia perdido em 2003, cerca de 53% concentrou-se somente nos segmentos de média e baixa tensão, relativo ao nível de distribuição (vide Figura 2.4.2). Segundo o gráfico, metade das perdas na distribuição, em média, é de caráter não-técnico, ressaltando a crescente carência de estratégias eficazes para a minimização do problema.

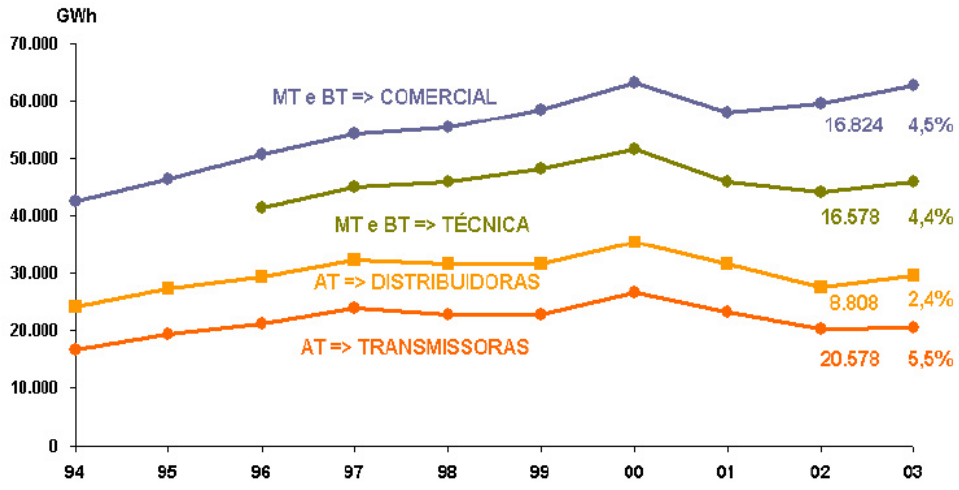


Figura 2.4.2 – Brasil: Índice de perdas por subsistemas em 2003.
Nota: AT >= 69 kV; MT e BT <= 69 kV.

Os altos índices de perdas técnicas no país podem ser justificados, em certo aspecto, pela característica do sistema elétrico nacional, em particular sua dimensão continental e a predominância hidrelétrica, que resultam em longos sistemas de transmissão e elevados fluxos energéticos entre regiões. Focando-se apenas no segmento de distribuição de energia, por analogia, Araújo (2007) afirma que o valor das perdas de energia é função direta do tamanho da área de concessão de cada distribuidora de energia.

A iluminação pública também pode favorecer as perdas quando o cadastro dos pontos de iluminação é ineficiente. Semelhantemente, fatores climáticos podem acarretar um maior consumo e, conseqüentemente, maiores perdas, como ocorre nos sistemas isolados do Norte do país (VIERALVES, 2005).

2.4.2 ANORMALIDADES SOB A ÓTICA DAS PERDAS COMERCIAIS

Considerando o caráter previsível e diretamente remediável das perdas técnicas, a principal componente das anormalidades no consumo de energia elétrica vincula-se a perdas de caráter não-técnico. Estas perdas, também denominadas *perdas comerciais,* devido a suas conseqüências catastróficas no faturamento das concessionárias, são definidas pela ANEEL nos seguintes termos:

> "*Perdas não técnicas:* apuradas pela diferença entre as perdas totais e as perdas técnicas, considerando, portanto, todas as demais perdas associadas à distribuição de energia elétrica, tais como furtos de energia, erros de medição, erros no processo de faturamento, unidades consumidoras sem equipamento de medição, dentre outros." (ANEEL, 2006, p.97)

Alguns autores propõem outra classificação para as perdas não-técnicas, como por, exemplo, Reis (2005) que as subdivide em **perdas administrativas** – ocasionadas pela própria empresa no processo de leitura dos medidores e emissão de faturamento – e perdas comerciais em si, causadas por furto de energia direto da rede elétrica. Penin (2008), em sua tese de doutorado, propõe uma abordagem multidisciplinar das perdas não-técnicas (PNT), categorizando-as em **perdas não-técnicas do tipo um**, que envolve furto de energia e fraude no fornecimento ou medidor, e **perdas não-técnicas do tipo dois**, que engloba falhas dos medidores, erros de leitura ou imprecisão no faturamento e demais perdas.

Existem ainda outras fontes de perdas comerciais pouco referenciadas na literatura técnica. Pode-se citar, por exemplo, o efeito das perdas técnicas nas perdas comerciais, ocorrendo quando dispositivos do sistema elétrico se comportam de forma adversa à modelada no cálculo das perdas técnicas, devido a defeitos ou obsolescência (Penin, 2008). Outra fonte de perdas não-técnicas é o próprio ciclo de leitura dos consumidores, que dificilmente coincide com mês civil, havendo a necessidade de se aproximar o consumo referente aos dias desconsiderados.

2.4.2.1 A problemática no Brasil

Os prejuízos no sistema elétrico brasileiro, decorrentes de perdas não-técnicas, têm sido tão representativos nos últimos anos que as distribuidoras de energia, o órgão regulador e demais entidades do setor vêm despendendo consideráveis esforços humanos e tecnológicos para a resolução do problema. Não obstante, estatísticas mostram que a questão tem se agravado, como citado na Tabela 2.4.3 (BRASIL, 2007). Entre 2003 e 2007, as perdas comercias cresceram 27%, enquanto a energia injetada no sistema aumentou somente 12 %.

Tabela 2.4.3 – Evolução das perdas 2003 e 2007

Atributo	2003	2007	Δ 2003 / 2007
Energia Injetada [TWh]	339	379	12%
Perdas Técnicas [TWh]	24	26	8%
Perdas Comerciais [TWh]	15	19	27%
Perdas Totais na Distribuição [TWh]	39	45	15%
Perdas Técnicas na Rede Básica [TWh]	8	7	-13%

Para se ter uma noção mais exata do impacto das perdas comerciais nas finanças do setor elétrico, relata-se que o valor da energia faturada referente às perdas comerciais alcançou R$ 2,9 bilhões em 2003 e R$ 5,3 bilhões em 2007 (BRASIL, 2007). Essas cifras, somadas aos valores das perdas embutidos nas tarifas, representaram um déficit de arrecadação, pelo setor elétrico e o Estado, de R$ 6,7 bilhões em 2003 e R$ 10 bilhões em 2007, em valores nominais. Este último montante representou cerca de 11% do faturamento de todas as concessionárias de energia elétrica no referido ano, perfazendo um total de 19 TWh de energia perdida, que seria suficiente para suprir todo o mercado cativo do Estado de Minas Gerais – com seus 6,2 milhões de consumidores – durante um ano inteiro.

Além dos prejuízos financeiros, perdas comerciais ainda acarretam sobrecargas nas linhas e transformadores da rede, devido ao consumo não-programado, minando a eficiência e a imagem das concessionárias. Há também a deterioração da segurança do sistema com o uso de cabos e dispositivos improvisados na prática de fraudes. Por fim, em um âmbito externo, o consumo inesperado traz a necessidade de reforços no parque gerador, com a antecipação de investimentos e expansões.

2.4.2.2 A segmentação das perdas não-técnicas no país

Além da distribuição de perdas não-técnicas no sistema elétrico brasileiro não se manter uniforme, observa-se também que ela encontra-se estratificada por classes de consumo e natureza da perda. Na Tabela 2.4.4 (VIEIRALVES, 2005), mostra-se uma estimativa da segmentação das perdas comerciais no país por classe de consumo, onde se conclui que metade da energia perdida ocorre na classe residencial, parcela que responde por 6% da energia faturada na referida classe.

Tabela 2.4.4 – Perdas Estimadas por classe de consumo

Classe	% em relação ao total de perdas	% em relação à energia faturada
Residencial	49,23	6
Comercial	25,43	11
Rural	19,45	20
Outras	3,08	6
Industrial	2,81	7

Já na Tabela 2.4.5, mostra-se um estudo estatístico com os tipos de irregularidades mais freqüentes, onde nota-se que o furto de energia é uma das principais fontes de perdas comerciais. Tal prática não é sinônimo de “fraude de energia” – termo que tem sido mais comumente associado à práticas de adulteração da medição de energia (PENIN, 2008, p.15), referentes as demais irregularidades listadas na tabela.

Tabela 2.4.5 – Tipos de irregularidades mais freqüentes

Tipo	%
Desvio antes do medidor	23,59
Medidor com defeito	22,01
Circuito de potencial interrompido	17,92
Desvio embutido na parede	14,78
Medidor com selo violado	6,28
Medidor danificado	4,72
Medidor com disco parado	3,78
Constante errada	3,46
Fraude na chave de aferição	3,46

FONTE: VIEIRALVES, 2005.

Fraude e Furto de energia fazem parte das chamadas “perdas por ação do consumidor” (PENIN, 2008), que compõem, além das formas de irregularidades mencionadas, as listadas a seguir:

- Ligações do medidor invertidas;
- Seqüência de fases invertida (reativo);
- Elemento móvel do medidor empenado ou bloqueado por meio de perfuração da tampa de vidro ou da base e posterior introdução de corpo estranho;
- Ponteiros do medidor deslocados;
- Engrenagem do medidor substituída;
- Dentes da engrenagem desgastados;
- Ponteiro da demanda retrocedido;
- Curto-circuito nos secundários dos transformadores de corrente;
- Alimentação do motor de temporização de demanda interrompida;
- Fraude no medidor digital;

Convém ressaltar que há ainda inúmeros tipos de irregularidades não mencionados. A crescente “criatividade” dos fraudadores de energia requer também uma constante capacitação humana e tecnológica dos órgãos que combatem o fenômeno.

2.4.2.3 Motivações

Diversas pesquisas têm sido realizadas na tentativa de realçar os fatores motivadores das perdas comerciais. De início, associava-se puramente o problema ao nível de pobreza de uma determinada região (SIMAS, 2003, *apud* ARAÚJO, 2007). Posteriormente, estudos mais aprofundados (CALILI, 2005; PENIN, 2008; ARAÚJO, 2007) têm despertado a atenção para aspectos igualmente relevantes na problemática; questões de cunho social, político, econômico e cultural, como abordadas a seguir.

- **Questões Econômicas**

É indiscutível a relação direta entre perdas comerciais e o grau de desenvolvimento de um país ou estado. Isso pode ser constatado no fato já comentado, que as regiões consideradas desenvolvidas possuem menores níveis de perdas (vide Tabela 2.4.2). Além disso, a situação econômica de uma região também tem impacto direto no grau de inadimplência da população, isto é, a capacidade de se honrar os acordos financeiros pelos serviços prestados.

Particularmente no Setor de Energia Elétrica, é possível se comprovar a alta correlação entre perdas comerciais e inadimplência (CALILI, 2005). O problema decorre quando, por motivo de estar inadimplente, determinado cliente tem o fornecimento de energia interrompido, o que pode o levar a pagar seus débitos ou a lançar mão de técnicas fraudulentas para ter o bem de volta. Por outro lado, se um cliente fraudulento é identificado e autuado, a multa a ele imposta poderá torná-lo inadimplente, o que forma um ciclo vicioso.

Outro aspecto relevante para o incremento das perdas comerciais é valor da tarifa de energia elétrica repassada aos consumidores. Araújo (2007) em sua tese de doutorado identificou, através de tratamento probabilístico, uma alta correlação entre perdas comerciais e níveis de tarifas de energia pagas. Ocorre que a ANEEL impõe atualmente a completa oneração das perdas nas tarifas de energia, o que as torna mais elevadas quanto maiores forem as perdas em uma região. O contínuo aumento das tarifas de energia nos últimos anos é, portanto, um grande obstáculo para a redução das perdas comerciais.

- **Aspectos Sócio-Culturais**

De maneira geral, quanto maior o acesso a serviços básicos, menores serão os níveis de perdas comerciais registrados. Altas taxas de escolaridade, por exemplo, motivam diretamente o correto e eficiente uso da energia elétrica. Por outro lado, baixos níveis de urbanização, altas taxas de violência urbana e de favelização intensificam a problemática mencionada (CALILI, 2005).

O componente cultural ou comportamental também afeta diretamente as perdas comerciais, sendo justamente o mais difícil de ser identificado. Nesse sentido, pode-se citar a não consciência do serviço de eletricidade como bem de consumo e o sentimento de impunidade arraigado em algumas culturas. Kaufmann *et al.* (1999, *apud* ARAÚJO, 2007) afirma que o furto de energia tem uma forte relação com a sensação de Governo, sendo o problema crítico em regiões com ausência do Poder Público, com instabilidade política ou com altos índices de corrupção. Percebe-se também uma falsa idéia na cultura brasileira de

que quando se furta energia, se está penalizando uma empresa “rica”, quando na verdade se está prejudicando toda a cadeira de produção de energia, e, conseqüente, os demais consumidores.

- **Aspectos de Natureza Regulatória**

Além da já mencionada questão tarifária, vários elementos impostos pela Lei têm desfavorecido a proliferação de estratégias de redução de perdas comerciais. Como exemplo, menciona-se a dúbia legislação para regularização de consumidores encontrados em situações de fraude, que estabelece procedimentos para materialização da fraude difíceis de serem seguidos por muitas concessionárias, ou, passíveis de serem desvirtuados por consumidores de má-fé. Igualmente, pode ser citada a influência de instâncias judiciais inferiores para o crescimento da chamada “indústria de liminares”, o que dificulta o processo de recuperação de receitas (PENIN, 2008).

2.4.2.4 Formas de prevenção e combate

Há duas formas de se lidar com a problemática das anormalidades no consumo de energia elétrica: a prevenção e a intervenção, ou combate. Inicialmente, listam-se a seguir algumas das principais medidas para prevenção de perdas comerciais, de acordo com as alternativas propostas por Penin (2008), Araújo (2007) e Calili (2005).

- **Inovações Tecnológicas**: investimentos em novas modalidades de medição interna e externa às unidades de consumo; novas formas de transmissão e padrões tecnológicos para entrega da energia ao consumidor final; uso de estratégias que visem agilizar o processo de estratificação e localização das perdas comerciais;
- **Medidas de apelo social:** campanhas educativas no sentido de conscientizar a população acerca do uso eficiente da energia elétrica e as conseqüências advindas da prática de fraude e furto; estímulo a ações de denúncia de suspeitos de irregularidade;
- **Aperfeiçoamento dos processos de gestão:** medidas relacionadas tanto à modernização dos processos administrativos empresarias para uma melhor efetividade quanto à melhoria das relações com os consumidores. Podem ser mencionados também investimentos para a manutenção de informações consistentes de faturamento, cadastro e topologia da rede, por exemplo, além da identificação de áreas críticas e aperfeiçoamento dos sistemas de faturamento.

- **Medidas nas esferas jurídica e penal:** parcerias com o Poder Público para as atividades de autuação e normalização; melhoramento da eficácia dos processos de recuperação de energia; atribuição de sanções justas aos consumidores fraudulentos, etc.

Já o combate à fraude e irregularidades pelas concessionárias de energia tem como principal frente os procedimentos de inspeção nas instalações dos consumidores (geralmente medidores). Cometti (2005) cita três estratégias usadas pelas concessionárias para direcionamento das inspeções de campo: varredura, denúncias e observação de variações de consumo. O pesquisador menciona em seu trabalho os baixos índices de assertividade decorrente das duas primeiras estratégias citadas anteriormente. Já a terceira alternativa, embora aliada, em algumas concessionárias, a alarmes nos aparelhos de leitura indicando desvios de consumo, também não representa um bom desempenho, dada à quantidade ínfima de perfis de consumo a serem analisados.

Os custos de deslocamento, capacitação e remuneração das equipes de campo são relativamente altos, requerendo certa sistemática para que haja viabilidade nas ações, no sentido de haver um equilíbrio entre o custo e o benefício dos investimentos, como mostrado na Figura 2.4.3. Nota-se que, dada a baixa capacidade das tarifas vigentes em remunerar os investimentos em de detecção de fraudes e ao crescente nível de inadimplência, muitas concessionárias não tem logrado êxito em atingir o limite econômico de sustentabilidade do processo.

Figura 2.4.3 – Curva custo-benefício de investimentos na redução de perdas comerciais

As diversas formas possíveis de se cometer fraudes e a constante evolução e disseminação do fenômeno são um fator adicional que exigem constante treinamento das

equipes e responsabilidade dos que conduzem o processo. Vieiralves (2005) relata, inclusive, a existência de verdadeiras “indústrias de fraude” em algumas concessionárias nacionais, inclusive com a participação ou conivência de funcionários internos. Nesse sentido, há que se ponderar a importância do leiturista para uma eficaz política de prevenção à fraude e ao furto de energia, onde é aconselhável que o mesmo possua treinamento suficiente para deduzir prováveis casos ou indícios de fraude e furto de energia, que esteja sintonizado com as metas da concessionária, e que seja, de preferência, não terceirizado (VIERALVES, 2005). O mesmo vale para profissionais das atividades de corte e religação.

Outro ponto a se destacar é a necessidade da imparcialidade daqueles que combatem fraudes. Ao contrário do que parece, o segmento de baixa renda é responsável por apenas um terço das perdas comerciais contabilizadas na maioria das concessionárias (NETO, 2008). No Estado do Rio de Janeiro, por exemplo, relata-se até a construção de condomínios de luxo com tecnologia que possibilite fraude de energia. Segundo a AMPLA, uma das distribuidoras de energia da região, “*até organizações ou pessoas, teoricamente insuspeitas, fazem furto de energia, como igrejas católicas, igrejas evangélicas, polícia, escolas, residência de um juiz, residência de um prefeito, restaurantes, hotéis, padarias, condomínios horizontais de classe alta*” (BRASIL, 2007, p.39);

Recentemente, o processo de inspeção de medidores e instalações em busca de fraudes tem se tornado viável economicamente, com o uso de estratégias inteligentes de inferência e descoberta de padrões anômalos. A tarefa de identificação computacional de irregularidades envolve o monitoramento dos perfis de uso de serviços ou bens para se estimar, detectar ou prevenir comportamentos indesejáveis (KOU, 2004). Este processo demanda a busca por dados históricos que possam indicar certa correlação com padrões irregulares ou fraudulentos, possibilitando direcionar as inspeções técnicas para os elementos pré-identificados como mais suspeitos, acarretando maiores chances de sucesso e, conseqüentemente, maiores retornos financeiros. Em 2008, por exemplo, a ELETROPAULO recuperou cerca de R$ 43 milhões em receitas e 296 GWh em energia decorrentes do combate a fraudes. Este bom desempenho foi atribuído ao uso de “evoluções tecnológicas, uso de softwares de identificação de irregularidades e de equipamentos eletrônicos para medição e monitoramento cada vez mais inteligentes e eficientes” (ELETROPAULO, 2009).

3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

3.1 INTRODUÇÃO

Qualquer metodologia inteligente para descoberta de perfis irregulares, como a proposta neste trabalho, envolve a manipulação de grandes volumes de dados. É necessário, portanto, um tratamento científico adequado para o desenvolvimento de tais sistemas, atribuição pertinente ao ramo computacional chamado de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Data Bases* – KDD).

Neste Capítulo discorre-se sobre as peculiaridades e etapas operacionais de KDD, com ênfase na *mineração de dados* e seus métodos de grande potencial à aplicação no problema de detecção de anormalidades. Dedica-se uma seção especial às técnicas de clusterização ou agrupamento de dados, particularmente aos métodos de partição *Fuzzy*, como o algoritmo *Fuzzy C-Means e o GK*, onde também se mencionam as principais estratégias usadas para validação de agrupamentos. Por fim, efetua-se uma breve revisão sobre as metodologias para identificação de fraudes e anormalidades no consumo de energia elétrica desenvolvidas sob o âmbito de KDD.

3.2 CONCEITO E ETAPAS OPERACIONAIS DE KDD

Segundo Fayyad *et al.* (1996, p.), KDD *"é um processo composto de várias etapas, não-trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados"*. A expressão "várias etapas"' refere-se tanto à complexidade interna do processo quanto a externa, envolvendo várias áreas de conhecimento, como banco de dados, inteligência artificial, aprendizado de

máquina, entre outras. Já os termos “interativo” e “iterativo” referem-se respectivamente à dependência do fator humano e ao uso de algoritmos iterativos no processo.

A busca de conhecimento em bases de dados muitas vezes é confundida com o termo “Mineração de Dados”, sendo este último processo, na verdade, uma da Descoberta do Conhecimento.

O processo de KDD pode ser resumido em três etapas operacionais: pré-processamento, mineração de dados e pós-processamento. Nas próximas seções, apresenta-se uma breve descrição de cada uma delas (GOLDSCHMIDT, 2005; HAN and KAMBER, 2006; WITTEN and FRANK, 2005).

3.2.1 PRÉ-PROCESSAMENTO

Esta etapa reúne atividades de tratamento e formatação dos dados para posterior uso em algoritmos de mineração. Nesta etapa, podem ser aplicadas as seguintes tarefas:

3.2.1.1 Seleção

A tarefa de seleção consiste na escolha de atributos e casos relevantes para nortear o processo de mineração de dados. Assumindo-se que os dados estejam dispostos em tabelas, organizados em uma seqüência de *m* registros, agrupados em *n* atributos, os seguintes tipos de tratamento de dados podem ser aplicados:

- **Redução Vertical de Dados**: consiste na eliminação ou substituição de atributos, de forma a encontrar o menor subconjunto que preserve as informações originais. Uma escolha coerente do conjunto de atributos traz duas principais vantagens: (i) um modelo de conhecimento mais conciso e de maior precisão; e (ii) redução do tempo de processamento. As estratégias mais usadas para redução de atributos são: eliminação direta, análise de componentes principais (PCA), seleção seqüencial para frente (*Foward Selecion*), seleção seqüencial para trás (*Backward Selection*), algoritmos genéticos e árvores de decisão.
- **Redução horizontal de dados**: caracteriza-se pela escolha de casos ou registros relevantes ao processo de descoberta de conhecimento. No linguajar de Banco de Dados, pode-se dizer que esta tarefa objetiva manter as tuplas mais representativas. A redução pode ser efetuada utilizando-se as seguintes estratégias: segmentação do banco de dados,

eliminação direta, agregação de informações ou amostragem aleatória (simples, por agrupamentos ou estratificada).

3.2.1.2 Limpeza

Limpeza de dados é uma tarefa de pré-processamento essencial para a manutenção da consistência do processo de descoberta de conhecimento, consistindo na correção de dados faltosos, redundantes, inconsistentes ou discrepantes. As técnicas mais comuns para o preenchimento de dados incoerentes são: preenchimento manual; exclusão; preenchimento com valor aleatório ou valor global comum e preenchimento com a média dos atributos ou valor mais provável.

3.2.1.3 Geração de Atributos

Representa a criação de novos atributos que expressam um inter-relacionamento entre os atributos já existentes. Neste processo, é muito comum substituir-se os atributos originais pelos novos atributos (derivados), o que acarreta a simplificação dos algoritmos de Mineração de Dados. O inter-relacionamento entre atributos é efetuado geralmente com o uso de operadores aritméticos tradicionais (+, -, ×, /) ou medidas estatísticas, como média, variância, desvio padrão, entre outras.

3.2.1.4 Normalização

Normalizar um conjunto de atributos significa estabelecer para eles uma faixa comum de valores, por exemplo, entre zero e um. Esta tarefa é de particular utilidade em estratégias de classificação que envolvam redes neurais ou técnicas de clusterização. Em redes neurais *Backpropagation*, a normalização geralmente acelera o processo de aprendizado, enquanto que em algoritmos de clusterização, a tarefa impede que atributos de maiores grandezas direcionem o processo em detrimento dos atributos de menor valor. As técnicas de normalização mais comumente utilizadas são:

- **Normalização Linear**: também conhecida como normalização *min-max*. Executa uma transformação linear dos dados originais, sendo recomendado nos casos onde exista a certeza que os valores do atributo estejam entre os valores mínimo e máximo considerados. Suponha que $\min_A$ e $\max_A$ sejam os valores mínimo e máximo de um

dado atributo A. A normalização *min-max* mapeia um valor, *v*, de *A* no valor *v'*, no intervalo $[\min_A, \max_A]$, computado de acordo com a equação:

$$v' = \frac{v - \min_A}{\max_A - \min_A} \tag{3.2.1}$$

- **Normalização por Desvio Padrão:** útil quando os valores mínimo e máximo do atributo são desconhecidos. Nesse método, um valor *v* de um atributo *A* será mapeado no valor v', da seguinte forma:

$$v' = \frac{v - \overline{A}}{\sigma A} \tag{3.2.2}$$

Onde $\overline{A}$ e σA são, respectivamente, a média e o desvio padrão do atributo *A*.

- **Normalização por Escala Decimal:** consiste no deslocamento do ponto decimal dos valores do atributo a ser normalizado, onde um dado valor *v*, de um atributo *A*, será mapeado no valor v', da seguinte forma:

$$v' = \frac{v}{10^j} \tag{3.2.3}$$

Onde *j* é o menor inteiro tal que o maior valor absoluto normalizado seja inferior a 1, ou seja, *v'<1*.

- **Normalização pela Soma dos elementos:** similar à normalização por escala decimal, exceto que o valor normalizado será divido pelo somatório de todos os valores do atributo, da seguinte forma:

$$v' = \frac{v}{\sum v} \tag{3.2.4}$$

- **Normalização pelo Máximo dos elementos:** neste caso, o valor do atributo será dividido pelo maior valor dentre todos os valores do atributo, de acordo com a expressão:

$$v' = \frac{v}{\max v} \tag{3.2.5}$$

3.2.2 MINERAÇÃO DE DADOS

A tarefa de mineração de dados é a principal etapa do processo de KDD, onde são executados algoritmos em busca da extração de informações implícitas contidas em uma base de dados. Basicamente, nesta fase são executados algoritmos que objetivam produzir um modelo de conhecimento[1] válido a partir do conjunto de dados utilizados.

Toda a etapa de mineração de dados é direcionada a partir dos objetivos estabelecidos no início do processo de KDD, onde são definidos o tipo e a forma de validação do modelo de conhecimento a ser gerado. Havendo sido definido o modelo de conhecimento, é possível, então, escolher a *estratégia de mineração de dados* que melhor se adapte ao problema, bem como a técnica de mineração para dar suporte do processo. Redes Neurais, Algoritmos Genéticos, Modelos Estatísticos e Probabilísticos são exemplos de técnicas que podem ser usadas na etapa de Mineração. Já o termo "estratégias de mineração de dados" refere-se a modelos clássicos de representação de conhecimento em bancos de dados, que serão abordados na Seção 3.3.

3.2.3 PÓS-PROCESSAMENTO

A tarefa final do processo de KDD congrega atividades relacionadas à formatação dos resultados da etapa de mineração, para uma adequada análise, visualização e interpretação, quando necessário. Como exemplo, pode ser citada a tarefa de simplificação de modelos de conhecimento baseados em regras, através do corte de regras por meio de medidas de qualidade. Cita-se ainda a tarefa de transformação de modelo de conhecimento, por exemplo, a conversão de árvores de decisão em regras ou vice-versa.

3.3 ESTRATÉGIAS PARA MINERAÇÃO DE DADOS

Estratégias para mineração de dados, também denominadas *tarefas de KDD*, referem-se a modelos convencionais de representação de conhecimento que podem ser utilizados na etapa de mineração de dados. Algumas das mais importantes são: descoberta de associações, análise de discriminantes, sumarização, análise de grandezas discrepantes (ou detecção de

[1] Refere-se a qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva um conjunto de dados (FAYYAD *et al.*, 1996).

desvios), previsão de séries temporais, regressão, classificação e clusterização, que podem ser vistas em detalhes em Goldschmidt (2005) e Han and Kamber (2006). Nesta seção, são abordadas brevemente as tarefas de classificação, clusterização e detecção de desvios, que apresentam grande potencial de aplicabilidade ao problema objeto deste trabalho.

3.3.1 CLASSIFICAÇÃO

É uma tarefa que consiste em associar os elementos de um banco de dados a rótulos pré-definidos, ou seja, em obter uma função de mapeamento de cada registro $\mathbf{X_i}$ à uma determinada classe $\mathbf{Y_i}$. Em outras palavras, representa o aprendizado supervisionado de um conjunto de exemplos, constituído de pares ordenados da forma $\{(x_1, y_1),\ldots,(x_n, y_n)\}$, que são utilizados para o treinamento de um modelo classificador. Posteriormente, a função já mapeada $(h: X \rightarrow Y)$ é utilizada para associar outros elementos às classes envolvidas.

As técnicas mais comumente usadas para classificação são: *árvores de decisão, redes bayesianas, sistemas baseados em regras, máquina de vetor de suporte* e *redes neurais backpropagation* (HAN & KAMBER, 2006). Há ainda inúmeras outras metodologias fundamentadas em *rough sets*, *Fuzzy sets* e *algoritmos genéticos*, cada qual apropriada a um determinado tipo de problema em particular, onde o conhecimento é representado também de forma específica, como ilustrado na Figura 3.3.1.

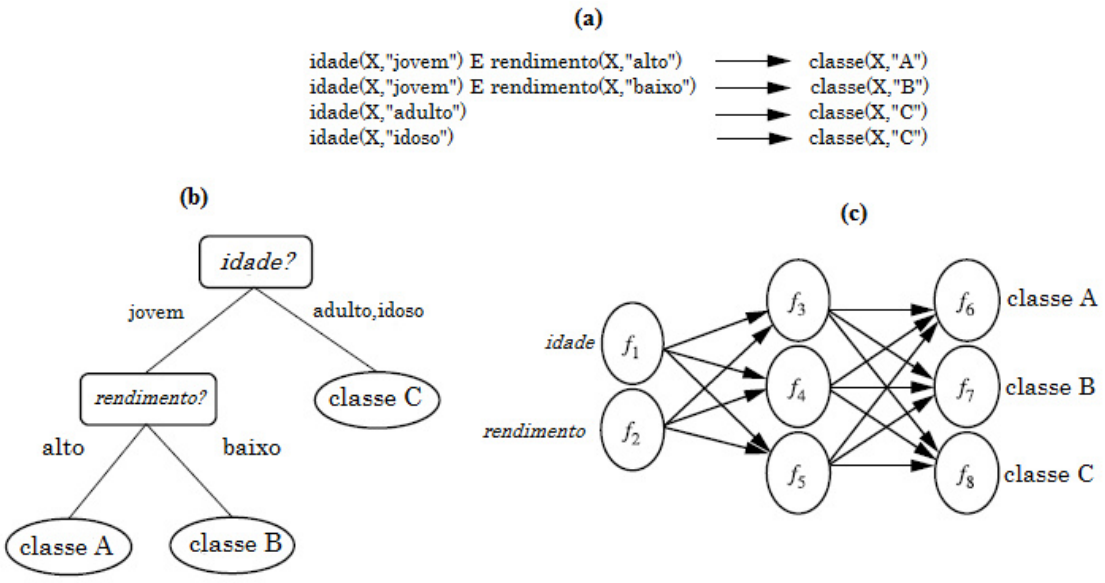


Figura 3.3.1 – Formas de representação de um modelo classificador: **(a)** regras do tipo SE-ENTÃO; **(b)** árvore de decisão; e **(c)** rede neural

3.3.1.1 Métricas para avaliação de sistemas classificadores

A eficiência de um modelo classificador é avaliada pela sua *acurácia,* isto é, pela porcentagem de casos corretamente classificados pelo modelo. Uma melhor análise de desempenho pode ser efetuada através da chamada *matriz de confusão*, que mostra o relacionamento entre classes preditas e reais, de onde também se originam os conceitos de *falsos-positivos, falsos-negativos, positivos-verdadeiros* e *negativos-verdadeiros* (HAN & KAMBER, 2006). A representação da referida matriz para duas classes – FALSO (F) e VERDADEIRO (V) – está mostrada na Tabela 3.3.1, onde se observa, por exemplo, que *falsos-positivos* são casos comprovadamente falsos, mas incorretamente classificados como verdadeiros, enquanto que *falsos-negativos* são os casos legítimos incorretamente classificados como falsos.

Tabela 3.3.1 – Matriz de Confusão para duas classes

Classe Real	**Classe Predita**	
	V	**F**
V	verdadeiro-positivo	falso-negativo
F	falso-positivo	verdadeiro-negativo

A partir da matriz de confusão, a *acurácia* de um sistema classificador é avaliada conforme a Equação 3.3.1.

$$ac = \frac{VP + FP}{VP + FP + VN + FN} \tag{3.3.1}$$

Há ainda outras métricas para aferição do desempenho de sistemas classificadores, particularmente úteis em problemas de identificação de fraudes, onde ocorre o problema de prevalência de classe. São elas: a *sensibilidade* (eq. 3.3.2), que mostra a relação entre casos positivos descobertos e existentes; a *especificidade* (eq.3.3.3), que determina a proporção de casos negativos descobertos dos existentes; e a *assertividade* (eq.3.3.4), que mede a proporção de casos positivos corretamente classificados.

$$sens = \frac{VP}{P} \tag{3.3.2}$$

Onde P é o total de casos positivos.

$$esp = \frac{VN}{N} \tag{3.3.3}$$

Onde N é o total de casos negativos.

$$ass = \frac{VP}{VP + FP} \tag{3.3.4}$$

3.3.2 CLUSTERIZAÇÃO

É um processo de agrupamento de dados conforme semelhanças naturais, onde não é conhecida *a priori* a distribuição dos grupos ou classes. Esta técnica, também conhecida como análise de agrupamentos, é geralmente utilizada como ferramenta para a obtenção de uma noção geral da distribuição de dados a partir da análise das características dos agrupamentos, ou como uma etapa de pré-processamento para algoritmos de classificação baseados nos grupos detectados. Uma análise mais aprofundada sobre os algoritmos de clusterização é efetuada na Seção 3.4.

3.3.3 DETECÇÃO DE DESVIOS

Em algumas aplicações, como detecção de fraudes ou detecção de invasões[2], a identificação de eventos raros pode ser mais útil do que eventos comuns. Em mineração de dados, a análise de grandezas discrepantes – que se desviam substancialmente da média de modo a levantar suspeitas sobre suas autenticidades – é conhecida como *detecção de desvios* (GOLDSCHMIDT, 2005) ou detecção de *outliers*.

A identificação de grandezas discrepantes pode ser efetuada utilizando **distribuições estatísticas –** que associam ao conjunto de dados uma distribuição ou modelo probabilístico e detecta os elementos *outliers* utilizando um teste de discordância – ou pelo uso de **medidas de distância**, que identificam os elementos que apresentam as maiores distâncias entre os demais. Podem ser usados também métodos baseados **em medidas de densidade**, útil em distribuições de dados com diferentes concentrações, como também estratégias baseadas em **medidas de desvio,** onde se estabelece as características principais de um grupo e identificam-se os elementos que mais se distinguem das características definidas (HAN & KAMBER, 2006).

3.4 TÉCNICAS DE CLUSTERIZAÇÃO DE DADOS

Clusterização refere-se ao agrupamento de registros, observações ou amostras em classes de objetos similares. Dada a grande quantidade de informação que compõe a maioria dos sistemas atuais de bancos de dados, em mineração de dados é freqüentemente recomendada a aplicação inicial de processos de clusterização, de forma a reduzir o espaço de busca de algoritmos posteriores de mineração (LAROSE, 2005). O elemento chave na estratégia de clusterização é o chamado **cluster** – um conjunto cujos elementos internos são os mais similares possíveis dentre toda a distribuição de dados analisada, e onde os elementos pertencentes a outros clusters são os mais dissimilares possíveis. A estratégia de clusterização difere da classificação no sentido que, nas técnicas de agrupamento, não existem variáveis ou rótulos pré-definidos para nortearem o processo.

[2] Do inglês, *intrusion detection*: sistemas computacionais para detecção de intrusos.

3.4.1 TIPOS BÁSICOS DE DADOS EM CLUSTERIZAÇÃO

Os principais algoritmos de clusterização operam sobre as seguintes estruturas de dados:

- **Matriz de Dados Linear**: representa *n* objetos com *p* variáveis (ou atributos) na forma de uma tabela relacional ou de uma matriz de dimensão *n* x *p*. O vetor de *n* objetos (ou *tuplas)* é representado da seguinte forma:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \tag{3.4.1}$$

- **Matriz de Dissimilaridades:** armazena medidas de proximidade relacionadas a qualquer par dentre *n* objetos de um conjunto de dados, na forma de uma matriz de dimensão *n* x *n*, como mostrado a seguir.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix} \tag{3.4.2}$$

3.4.2 MEDIDAS DE DISTÂNCIA

Um imediato questionamento que surge ao se tratar técnicas de clusterização é sobre que tipo de padrão usar para se definir o termo “proximidade”, ou seja, como se determinar a similaridade entre um par de objetos, entre um objeto e um cluster ou entre um par de clusters. Somente através do cálculo da similaridade ou, mais especificamente, da dissimilaridade (distância) é possível se estabelecer a pertinência ou não a um cluster.

Formalmente, a dissimilaridade $d(\mathbf{x},\mathbf{y})$ entre dois vetores $\mathbf{x}$ e $\mathbf{y}$ é definida como uma função satisfazendo as seguintes condições:

$$\begin{aligned} &d(\mathbf{x},\mathbf{y}) \geq 0, \quad \forall\ \mathbf{x} \text{ e } \mathbf{y} \\ &d(\mathbf{x},\mathbf{x}) = 0, \quad \forall\ \mathbf{x} \\ &d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}) \end{aligned} \tag{3.4.3}$$

Já a métrica *distância* é um conceito mais restritivo, requerendo a satisfação do teorema da desigualdade triangular; ou seja, para qualquer padrão $\mathbf{x}$, $\mathbf{y}$ e $\mathbf{z}$, tem-se:

$$d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y}) \tag{3.4.4}$$

No caso de atributos contínuos, uma variedade de funções de distância pode ser usada (vide Tabela 3.4.1), cada qual apropriada a um determinado tipo de problema.

Tabela 3.4.1 – Funções de distância entre dois padrões x e y

Função de Distância	Expressão
Euclidiana	$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Hamming (city block)	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\lvert x_i - y_i\rvert$
Tchebyschev	$d(\mathbf{x},\mathbf{y}) = max_{i=1,2,..,n}\lvert x_i - y_i\rvert$
Minkowski	$d(\mathbf{x},\mathbf{y}) = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p}$, $p > 0$
Canberra	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{x_i + y_i}$, x_i e y_i positivos
Separação Angular	$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\left[\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2\right]^{1/2}}$

Cada medida de distância está vinculada também a um determinado padrão geométrico, que pode ser visualizado e interpretado facilmente para o caso bidimensional – um vetor com dois atributos, $\mathbf{x}=[x_1 x_2]^T$ – sendo calculada sua distância à origem. Os contornos das curvas mostradas na Figura 3.4.1 revelam que tipo de construção geométrica é enfatizado com cada medida de distância. Percebe-se, por exemplo, que a distância euclidiana favorece formas circulares de clusters, enquanto a distância de Tchebyschev ressalta distribuições de agrupamentos quadráticas.

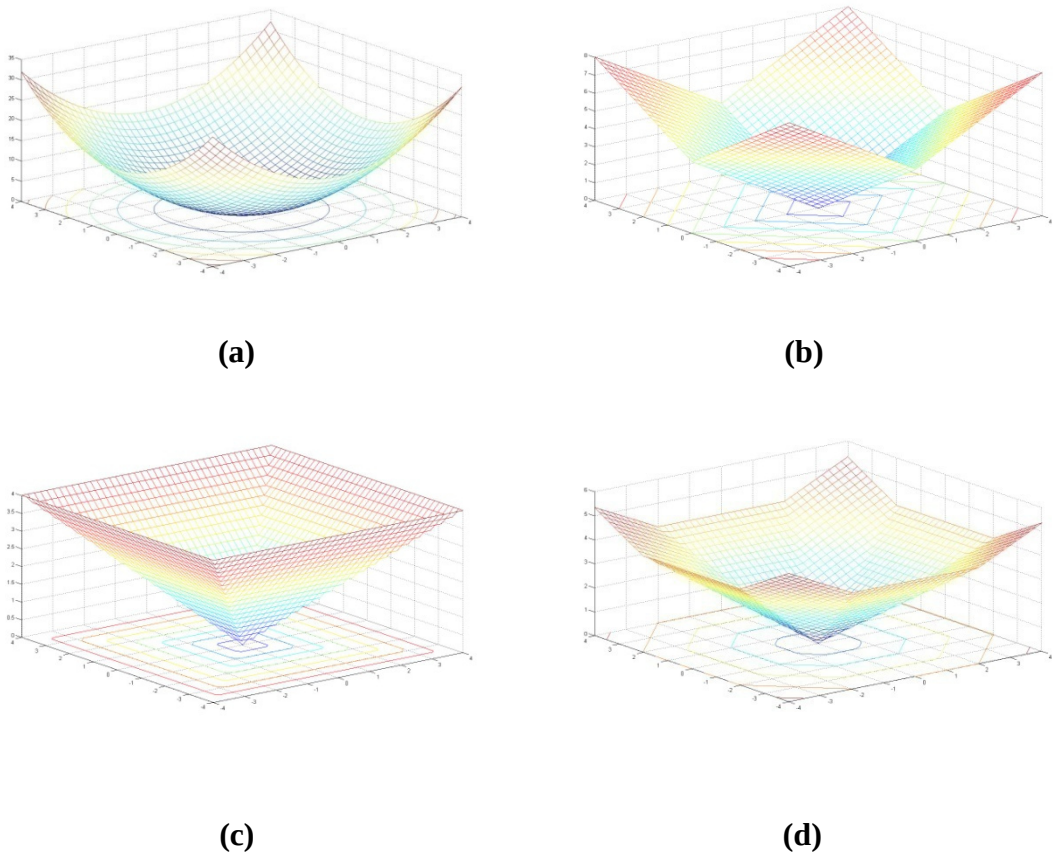


Figura 3.4.1 – Exemplos de funções de distância, gráficos tridimensionais e curvas de nível: **(a)** Euclidiana, **(b)** Hamming, **(c)** Tchebyschev, **(d)** função combinada de distância: max (2/3 Hamming, Tchebyschev).

3.4.3 MÉTODOS MAIS COMUNS DE CLUSTERIZAÇÃO

3.4.3.1 Métodos hierárquicos

Nos métodos hierárquicos é estabelecida uma decomposição hierárquica do conjunto de objetos usados – uma estrutura de agrupamentos semelhante a uma árvore, denominada *dendograma*, criada através de sucessivas partições (métodos divisivos) ou sucessivas combinações (métodos aglomerativos).

Nas estratégias aglomerativas (do inglês, *bottom-up)* trata-se cada elemento como um cluster, sendo que a cada passo os grupos mais próximos são combinados. Já nos métodos divisivos (do inglês, *top-down),* trata-se inicialmente todo o conjunto de *n*-tuplas como um grande cluster, sendo o mesmo fragmentado a cada iteração. Dada à natureza do processo, os métodos hierárquicos oferecem, geralmente, um grande custo computacional (PEDRYCZ, 2005).

Um importante conceito nos métodos hierárquicos é a distância entre clusters. Sejam A e B dois agrupamentos; a distância entre eles pode ser medida das seguintes formas:

- **Ligação simples (*Single-Link*)**: a distância $d(A,B)$ é associada à menor distância entre os elementos pertencentes a A e a B, da seguinte forma:

$$d(A,B) = \min_{x \in A, y \in B} d(\mathbf{x},\mathbf{y}) \tag{3.4.5}$$

- **Ligação Completa (*Complete-Link*):** a distância $d(A,B)$ é associada à **maior** distância entre os elementos pertencentes a A e a B, da seguinte forma:

$$d(A,B) = \max_{x \in A, y \in B} d(\mathbf{x},\mathbf{y}) \tag{3.4.6}$$

- **Ligação Média (*Average-Link*):** a distância $d(A,B)$ é associada à **média** de todas as distância entre os elementos pertencentes a A e a B, da seguinte forma:

$$d(A,B) = \frac{1}{card(A)card(B)} \sum_{x \in A, y \in B} d(\mathbf{x},\mathbf{y}) \tag{3.4.7}$$

Onde *card* representa a cardinalidade, ou seja, a quantidade de elementos pertencentes ao dado cluster.

3.4.3.2 Métodos particionais

Os métodos de agrupamento particionais têm por objetivo segmentar o conjunto de dados analisado em um número pré-definido de partições, através da otimização de uma função objetivo, sendo que, ao final do processo, cada partição representará um cluster. Em essência, a segmentação de *N* padrões em *c* clusters é um problema não-trivial. O número total de partições possíveis (vide Equação 3.4.8) possui complexidade exponencial, o que torna a estratégia de busca exaustiva inviável.

$$\frac{1}{c!}\sum_{j=1}^{c}(-1)^{c-j}\binom{c}{j}j^{N} \tag{3.4.8}$$

Este problema de otimização é melhor direcionado através do uso de algoritmos heurísticos para busca da partição ótima , onde um dos maiores desafios consiste em se formular uma função objetivo apropriada ao problema tratado. Um dos critérios mais usados é o somatório dos erros quadráticos, que fundamenta um dos principais algoritmos particionais, o *C-Means*, tratado a seguir.

- **O algoritmo *C-Means***

Neste algoritmo, cada *tupla* $\mathbf{x_i}$ em um conjunto de dados $\mathrm{X}=\{\mathbf{x_1},\ldots,\mathbf{x_n}\}, \mathrm{X}\in \square^{p}$ é associada a exatamente um cluster. Cada grupo U_j é um subconjunto do conjunto de dados, ou seja, $U_j \in \mathrm{X}$, sendo $U=\{U_1,\ldots,U_c\}$, portanto, uma partição exaustiva de X em *c* subconjuntos disjuntos e não-vazios, tal que $1<c<n$. A partição é dita "ótima" quando a soma do quadrado das distâncias entre os clusters e os elementos associados a eles é mínima, o que representa, em outras palavras, a minimização da seguinte função objetivo:

$$J_h(X,U_h,C)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}d_{ij}^{2} \tag{3.4.9}$$

Onde $C=\{C_1,\ldots,C_c\}$ é o conjunto de protótipos[3], d_{ij} a matriz de distâncias entre $\mathbf{x_i}$ e o centro de cluster $\mathbf{c_j}$, e U é uma matriz binária de dimensão c x n denominada *matriz de partição.* Os elementos individuais $u_{ij}\in\{0,1\}$ indicam a pertinência dos elementos nos

[3] *Protótipo* é o elemento que representa todo o cluster onde se encontra, podendo ser a média dos elementos (*C-Means*), o elemento mais central (*c-medoids*) ou algum outro.

clusters, sendo que $u_{ij} = 1$ se $\mathbf{x_i} \in U_j$ e $u_{ij} = 0$ caso contrário. Cada elemento deve ser associado a somente um grupo, o que impõe também a seguinte restrição:

$$\sum_{j=1}^{c} \mu_{ij} = 1, \quad \forall i \in \{1,\ldots,\mathrm{n}\} \tag{3.4.10}$$

Esta restrição força o uso de partições exaustivas, como também evita a solução trivial ao se minimizar J_h, ou seja, nenhum elemento associado a nenhum cluster: $u_{ij} = 0, \forall i, j$. É possível também que elementos sejam associados a um ou mais clusters restando ainda outros clusters vazios. Para evitar isso, é aplicada a seguinte restrição:

$$\sum_{i=1}^{n} \mu_{ij} > 0, \quad \forall j \in \{1,\ldots,\mathrm{c}\} \tag{3.4.11}$$

A função objetivo (Eq. 3.4.9) é minimizada através de um esquema de otimização alternada, realizado em duas fases: inicialmente, a matriz $\mathbf{C}$ é otimizada tomando-se a matriz $\mathbf{U}$ constante; posteriormente, o processo inverso é efetuado. Sendo assim, o algoritmo *C-means* compõe as seguintes etapas: (i) escolha aleatória de *c* objetos (tuplas), representando os *c* clusters; (ii) associação de cada elemento do conjunto de dados ao cluster mais próximo, baseando-se na distância do elemento ao centro de cada cluster; (iii) cálculo do novo centro de cada cluster (elemento médio); (iv) repetição dos passos (ii)-(iii) até que uma determinada condição de parada seja satisfeita, por exemplo, até que não haja mais mudanças nos clusters.

Algumas restrições do algoritmo *C-Means*, como a não-garantia de convergência ao ponto ótimo global e a sensibilidade a ruídos e *outliers* levaram ao desenvolvimento de muitas variantes do algoritmo. Citam-se, por exemplo, os seguintes algoritmos: *C-Modes*, *C-Prototypes*, *C-Medoids*, *ISODATA* e *PAM* (GOLDSCHMIDT, 2005).

3.4.4 CLUSTERIZAÇÃO *FUZZY*

Os métodos de clusterização descritos nas seções anteriores geralmente não fornecem uma representação convincente da estrutura dos dados nos clusters, como em situações de pertinência mínima ou parcial em clusters, elementos de bordas ou padrões discrepantes. Por meio da chamada clusterização *Fuzzy* é possível contornar estas restrições, através do uso de graus de pertinência nos clusters – valores no intervalo [0,1]. Esta grande vantagem permite a

modelagem de elementos que podem pertencer a mais de um cluster, oferecendo um maior grau de detalhamento ao modelo.

O conceito de graus de pertinência está baseado na definição e interpretação dos *conjuntos Fuzzy*, propostos por Lotfi Zadeh, em 1965. Enquanto que nos métodos clássicos de partição os protótipos dos clusters são representados por conjuntos clássicos, na clusterização *Fuzzy* eles são representados por conjuntos *Fuzzy*. Formalmente, a cada elemento $\mathbf{x_i}$ do conjunto de dados será associado um vetor de partição $\boldsymbol{\mu}_\mathbf{i} = (\mu_{i1}, \ldots, \mu_{ij}, \ldots, \mu_{ic})$ onde μ_{ij} representa o grau de pertinência da tupla $\mathbf{x_i}$ no cluster U_j, isto é, $\mu_{ij} = \mu_{U_j}(\mathbf{x_i}) \in [0,1]$.

Para a definição dos tipos de graus de pertinência e o conjunto de partições *Fuzzy* a ser usado, várias abordagens tem sido propostas, sendo a principal a estratégia probabilística. Mais especificamente, seja $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ uma matriz de dados a ser clusterizada e *c* a quantidade especificada de clusters $(1 < c < n)$, representados por conjuntos *Fuzzy* $\mu_{\Gamma j}, (j = 1, \ldots, c)$; denomina-se $U_f = (\mu_{ij}) = (\mu_{\Gamma j}(\mathbf{x}_i))$ uma partição de agrupamento probabilístico de X se ela atender às seguintes restrições:

$$\sum_{i=1}^{n} \mu_{ij} > 0, \quad \forall j \in \{1, \ldots, c\}, \tag{3.4.12}$$

e

$$\sum_{j=1}^{c} \mu_{ij} = 1, \quad \forall i \in \{1, \ldots, n\} \tag{3.4.13}$$

A restrição (3.4.12) garante que nenhum cluster seja vazio, já a condição (3.4.13) impõe que cada grau de pertinência deva ter o mesmo peso dos seus elementos vizinhos, o que efetua a normalização dos graus de pertinência por cada elemento.

- **O algoritmo *Fuzzy C-Means (FCM)***

O método *Fuzzy C-Means* é o principal representante dos algoritmos probabilísticos *Fuzzy*. Sejam dadas as matrizes $X\,(n \times p)$, $C(c \times p)$ e U_f $(n \times c)$, denominadas, respectivamente *matriz de dados*, *matriz de centros* e *matriz de partição Fuzzy*. O algoritmo FCM baseia-se na minimização da seguinte função de custo:

$$J_f(X, U_f, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m d_{ij}^2 \tag{3.4.14}$$

sujeita às restrições 3.4.12 e 3.4.13, sendo n é o número de elementos, p a quantidade de atributos associados a cada elemento, c o número de clusters, m um fator denominado expoente *fuzzificador*, μ_{ij} o grau de pertinência do elemento i no cluster j e d_{ij} uma medida de distância.

A minimização é efetuada através de um algoritmo de otimização alternada. Em uma primeira etapa, os centros são mantidos constantes, sendo otimizados os graus de pertinência. Posteriormente, os centros são otimizados mantendo-se os graus de pertinência constantes. Os valores ótimos são encontrados igualando-se a zero as derivadas parciais da Equação 3.4.14 em relação ao centro e aos graus de pertinência, sendo, respectivamente:

$$c_{jk} = \frac{\sum_{i=1}^{n} \mu_{ij}^m \, x_{ik}}{\sum_{i=1}^{n} \mu_{ij}^m} \tag{3.4.15}$$

$$\mu_{ij} = \frac{d_{ij}^{-\frac{2}{m-1}}}{\sum_{l=1}^{c} d_{il}^{-\frac{2}{m-1}}} \tag{3.4.16}$$

Em suma, o algoritmo FCM é composto pelas seguintes etapas:

1. Definição dos parâmetros c, m e da tolerância ε;
2. Inicialização randômica da matriz de centros **C** (q × p) e da matriz de partição fuzzy **U** (n × q);
3. Cálculo da matriz distância euclidiana **D** (n × q)
4. Atualização dos graus de pertinência nos clusters (Eq. 3.4.16);
5. Atualização da matriz de centros (Eq. 3.4.15);
6. Execução do teste de parada, como por exemplo: $\max |\mu_{ik}^{it)} - \mu_{ik}^{(it-1)}| \leq \varepsilon$

   Se não for satisfeito, voltar ao passo (3).

Recomenda-se utilizar valores do expoente fuzzificador (*m)* no intervalo de 1,1 a 2,0. Valores acima de 2,0 diminuem o caráter fuzzy (pertinências aproximadas) das partições clusterizadas (HWANG and THILL, 2007).

- **O algoritmo *Fuzzy GK (Gustafson-Kessel)***

O algoritmo de clusterização *fuzzy* proposto por Gustafson & Kessel (1979) substitui a distância euclidiana pela distância de *Mahalanobis*, que forma clusters elipsoidais que se adaptam a uma maior gama de distribuição de dados. A distância de Mahalanobis associada a um cluster *j* é definida da seguinte forma:

$$d^2(x_i, C_j) = (\mathbf{x}_i - \mathbf{c}_j)^T \mathbf{A}(\mathbf{x}_i - \mathbf{c}_j) \tag{3.4.17}$$

onde **A** é a matriz covariância do cluster. Para o caso em que $\mathbf{A} = \mathbf{I}$, a Equação anterior se resume à distância euclidiana, que é apta somente à detecção de clusters esféricos.

No método GK o protótipo de cada cluster é representado pela tupla $C_i = (\mathbf{c}_i, A_i)$, possuindo, portanto, dois parâmetros de aprendizado. Entretanto, a função objetivo, as equações de atualização dos centros e dos graus de pertinência são as mesmas do Fuzzy *C-Means*, respectivamente Equações (3.4.14), (3.14.15) e (3.14.16). Já a equação de atualização para as matrizes de covariância é dada por:

$$\mathbf{A}_j = \frac{\mathbf{A}_j^*}{\sqrt[p]{\det(\mathbf{A}_j^*)}}, \text{ com } \mathbf{A}_j^* = \frac{\sum_{k=1}^{N} \mu_{ij}(x_k - c_j)(x_k - c_j)^T}{\sum_{k=1}^{N} \mu_{ij}} \tag{3.4.18}$$

Embora o algoritmo Gustafson-Kessel consiga extrair uma maior quantidade de informação quando comparado a algoritmos baseados em distância euclidiana, ele é mais sensível aos parâmetros de inicialização. Nesse sentido, recomenda-se inicializá-lo com algumas iterações do FCM ou do *C-Means* tradicional.

3.4.5 ESTRATÉGIAS DE VALIDAÇÃO DE CLUSTERS

A validade dos resultados obtidos através de clusterização depende essencialmente dos parâmetros do algoritmo utilizado. A influência de parâmetros como a quantidade de clusters

(*c*) e o expoente fuzzificador (*m*) na qualidade dos resultados ainda não é algo completamente conhecido (OLIVEIRA & PEDRYCZ, 2007).

Três formas principais têm sido utilizadas para a determinação da quantidade ótima de clusters e outros parâmetros em uma distribuição de dados. A primeira estratégia é efetuar o processo de clusterização utilizando diferentes combinações e aferir a qualidade da partição por meio de índices de desempenho. A segunda consiste em escolher as melhores configurações de parâmetros através de técnicas de visualização. Já a terceira, menos utilizada, baseia-se em iniciar a clusterização com um grande número de clusters e, sucessivamente, unir os clusters mais similares, segundo um critério pré-definido.

A seguir, comentam-se as duas primeiras estratégias.

3.4.5.1 Índices de validação de clusters

A maioria dos índices para validação de clusters tem o foco em duas propriedades: *compacidade* e *separabilidade.* A primeira representa uma medida do grau de variação dos dados em um cluster, já a separabilidade é usada para estimar propriedades inter-clusters.

O objetivo principal de um índice de validação de cluster é minimizar a compacidade e maximizar a separabilidade. Entretanto, dada às inúmeras formas de se definir tais grandezas, existem também inúmeros índices de validação, com características e aplicabilidades particulares.

Todavia, a seguir citam-se os índices de desempenho mais robustos, que também serão aplicados na validação da metodologia desenvolvida no presente trabalho.

- **Índice de Partição (SC):** calculado pela razão da soma da compacidade e da separação entre os clusters, objetivando minimizar a seguinte expressão:

$$SC(c) = \sum_{j=1}^{c} \frac{\sum_{i=1}^{n} (\mu_{ij})^m d_{ij}^2}{\sum_{i=1}^{n} \mu_{ij} \sum_{l=1}^{c} v_{lj}^2} \tag{3.4.19}$$

  O índice SC é útil na comparação de diferentes partições portando o mesmo número de clusters.

- **Índice de Separação (S):** utiliza uma métrica de mínima separação entre clusters, calculado da seguinte forma:

$$S(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}(\mu_{ij})^2 d_{ij}^2}{\mathrm{n}\ min_{j,l}(v_{lj}^2)} \tag{3.4.20}$$

Uma quantidade ótima de clusters é atingida ao se minimizar esta equação.

- **Índice *Xie-Beni* (XB):** objetiva mensurar a razão da total variação entre clusters e a separação entre eles, dado pela minimização da seguinte função objetivo:

$$XB(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}(\mu_{ij})^m d_{ij}^2}{\mathrm{n}\ min_{j,i}} \tag{3.4.21}$$

3.4.5.2 Visualização de resultados

Uma das grandes dificuldades enfrentadas em processos de clusterização é se avaliar não somente a similaridade ou dissimilaridade entre clusters, mas o quanto os grupos formados são compatíveis com a distribuição de dados usada. Para solucionar este problema, é necessário se extrair informações adicionais do processo, dado que as métricas convencionais de desempenho de clusters fornecem somente medidas escalares.

Nesse sentido, podem ser usadas técnicas de representação gráfica dos clusters de forma a permitir uma apreciação visual dos resultados e, conseqüentemente, se estimar o melhor algoritmo ou a melhor combinação de parâmetros. Algumas das metodologias de visualização gráfica mais importantes são (OLIVEIRA & PEDRYCZ, 2007): *Avaliação Visual de Tendência de Clusters* (do inglês, VAT), *Validação Visual de Cluster* (do inglês, VCV), *Análise de Componentes Principais* (do inglês, PCA) e o Mapeamento de Sammon, utilizado neste trabalho, descrito mais detalhadamente a seguir.

- **Mapeamento de Sammon**

Representa uma técnica não-linear de mapeamento de dados de um espaço *n*-dimensional superior a um espaço *q*-dimensional inferior, de forma que as distâncias

$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ entre os pontos do espaço de dimensão *n* correspondam às distâncias $d^*_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$. O mapeamento consiste, portanto, em minimizar um critério de erro denominado *Sammon's stress*, dado a seguir:

$$E = \frac{1}{\lambda} \sum_{i=1}^{N-1} \sum_{j=1}^{N} \frac{(d_{ij} - d^*_{ij})^2}{d_{ij}} \tag{3.4.22}$$

onde $\lambda = \Sigma_{i<j} d_{ij} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}$.

A minimização de E é um problema de otimização em $N * q$ variáveis y_{il}, $i = 1, 2, \ldots, N$ e $l = 1, 2, \ldots, q$, com $\mathbf{y_i} = [y_{i1}, \ldots, y_{yq}]^{\mathrm{T}}$. Aplicando-se o método *Steepest Descent*, a aproximação de y_{il} na $t-$th iteração será dada por:

$$y_{il}(t+1) = y_{il}(t) - \alpha \left[\frac{\frac{\partial E(t)}{\partial y_{il}(t)}}{\frac{\partial^2 E(t)}{\partial y_{il}^2(t)^2}} \right] \tag{3.4.23}$$

onde α é um escalar não-negativo, sendo as derivadas primeira e segunda dadas por:

$$\frac{\partial E(t)}{\partial y_{il}(t)} = -\frac{2}{\lambda} \sum_{k=1, k\neq i}^{N} \left[\frac{d_{ki} - d^*_{ki}}{d_{ki} d^*_{ki}} \right] (y_{il} - y_{kl}) \tag{3.4.24}$$

$$\frac{\partial E(t)}{\partial y_{il}(t)} = -\frac{2}{\lambda} \sum_{k=1, k\neq i}^{N} \frac{1}{d_{ki} d^*_{ki}} \left[(d_{ki} - d^*_{ki}) - \left(\frac{(y_{il} - y_{kl})^2}{d^*_{ki}} \right) \left(1 + \frac{d_{ki} - d^*_{ki}}{d_{ki}} \right) \right] \tag{3.4.25}$$

- **Mapeamento de Sammon Modificado (FUZZSAM)**

No mapeamento clássico de Sammon, a projeção de *N* pontos do espaço *n*-dimensional no espaço de dimensão *q* requer, a cada iteração, o cálculo de $N(N-1)/2$ distâncias, o que torna o algoritmo com elevado custo computacional. Focando nesse problema e buscando também a aplicabilidade a clusterização fuzzy, Abonyi & Babuska (2004) propuseram uma alteração do algoritmo inicial de Sammon para se calcular somente $N \times c$ distâncias a cada iteração, fundamentando-se na seguinte função objetivo:

$$E_{fuzz} = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ki})^m (d(\mathbf{x_k}, \mathbf{v_i}) - d^*_{ki})^2 \quad (3.4.26)$$

Onde $d(\mathbf{x_k}, \mathbf{v_i})$ representa a distância entre o ponto $\mathbf{x_k}$ e o centro de cluster $\mathbf{v_i}$, situados no espaço *n*-dimensional original, enquanto que $d^*_{ki} = d^*(\mathbf{y_k}, \mathbf{z_i})$ representa a distância euclidiana entre os centros projetados $\mathbf{z_i}$ e os dados projetados $\mathbf{y_k}$.

O algoritmo resultante é similar ao mapeamento clássico de Sammon, com a diferença que, após a atualização das coordenadas originais a cada iteração, os centros projetados são recalculados baseando-se na fórmula da média ponderada usada nos algoritmos fuzzy (Eq. 3.4.16). Os graus de pertinência dos dados projetados podem ser calculados de forma similar, da seguinte forma:

$$\mu^*_{ki} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d^*(\mathbf{x_k}, \eta_i)}{d^*(\mathbf{x_k}, \mathbf{v_j})}\right)^{\frac{2}{m-1}}} \quad (3.4.27)$$

Sendo $\mathbf{U}^* = [\mu^*_{ki}]$ a matriz de partição contendo os graus de pertinência recalculados.

3.5 METODOLOGIAS APLICADAS A DETECÇÃO DE ANORMALIDADES NO CONSUMO DE ENERGIA ELÉTRICA BASEADAS EM MINERAÇÃO DE DADOS

Nesta seção, citam-se, por fim, alguns dos trabalhos mais relevantes sobre o processo de identificação de fraudes e perfis anômalos no consumo de energia elétrica, desenvolvidos no ramo de Descoberta de Conhecimento em Bases de Dados.

Em Calili (2005) é proposta uma metodologia para detecção de fraudes em consumidores de baixa tensão, onde classificam-se consumidores como fraudulentos, normais e inadimplentes. O trabalho, que relata uma assertividade de 68.92% para clientes fraudulentos, é baseado em dados cadastrais e variáveis socioeconômicas, totalizando 18 variáveis de entrada. Inicialmente é feita a clusterização do banco de dados disponibilizado pela distribuidora de energia elétrica utilizando-se um mapa auto-organizável de *Kohonen*. Posteriormente, os dados sócio-econômicos e de consumo geram distribuições estatísticas que são a base das funções de pertinência *Fuzzy* em um modelo estatístico.

Em Cabral (2005), desenvolveu-se um sistema para detecção de fraudes em consumidores de alta tensão, fundamentado na teoria de *Rough Sets* para o processo de seleção de regras válidas. A partir da definição de perfis de comportamentos diários, classificam-se os consumidores em normais ou anormais, sendo relatada, neste trabalho, uma assertividade de 64.3%, com a utilização de 19 atributos.

Sobre os trabalhos que utilizam o algoritmo *Fuzzy C-Means*, pode ser citada a metodologia de Rocha (2003), que utiliza o FCM e um sistema *neuro-fuzzy* para a classificação de fraudes e extração de regras em clientes comerciais e industriais, utilizando 24 atributos. Cita-se, igualmente, o trabalho de Lazo *el AL* (2005), onde desenvolve-se um sistema baseado em clusterização *Fuzzy* e classificação baseado na matriz de dissimilaridade. Nesse último trabalho, utilizam-se 20 atributos originados de clientes de alta tensão, não sendo informada nenhuma métrica de assertividade.

DOS-ANGELOS *et al.* (2007) propõe uma abordagem para detecção de fraudes que correlaciona a influência de possíveis fatores de perdas comerciais em uma dada região a um índice final, baseando-se nos princípios do raciocínio aproximado modelados pela Lógica *Fuzzy*. No sistema em questão, cinco variáveis *Fuzzy* foram definidas como entradas – *consumo*, *ocupação*, *perfil_região*, *peso_conta* e *satisfação* – sendo elaborada como saída do sistema a variável *indice_fraude.* Por fim, um processo de inferência *Fuzzy* baseado em um sistema do tipo Mamdani avalia a chance de um dado consumidor estar fraudando energia, através de um índice no intervalo unitário, sendo obtidos bons resultados em testes em uma base de testes fictícia.

Este último trabalho apresenta certo ponto conflitante: a dificuldade em se elaborar regras que possam modelar corretamente o maior número possível de padrões associados a fraudes. E mesmo com a geração bem-sucedida de um conjunto de regras para uma dada área de concessão, algumas delas podem não ser compatíveis com uma outra região de concessão. Esta é uma dificuldade comum dos métodos supervisionados, restrição que o modelo a ser desenvolvido procurou contornar.

Finalmente, cabe ainda mencionar as seguintes metodologias: Jiang *et al.* (2002) – um sistema baseado em *wavelets* e múltiplos classificadores para detecção de mudanças bruscas de consumo, obtendo assertividade de 70%; Eller (2003), uma arquitetura de informações baseada em Redes Neurais, para inferência de clientes suspeitos; Ferreira (2007), um sistema

especialista de apoio à tomada de decisão para seleção de alvos – estratégia amplamente utilizada em várias concessionárias (sem testes de assertividade); Filho *et al.* (2004), metodologia para detecção de fraudes baseada em árvore de decisão (assertividade de 40%) e; Yap *et al.* (2007), abordagem baseada em algoritmos genéticos e máquina de vetor de suporte híbrido. Há que se ressaltar a inexistência de métricas padrões de desempenho em várias das metodologias citadas, como neste último trabalho citado.

4 METODOLOGIA PARA CLASSIFICAÇÃO DE ANORMALIDADES BASEADA EM CLUSTERIZAÇÃO *FUZZY*

4.1 INTRODUÇÃO

Considerando as metodologias abordadas, pode-se observar que elas enquadram-se em uma ou ambas das seguintes situações: (i) uso de um grande número de atributos; ou (ii) aplicabilidade restrita a sistemas de média ou alta tensão. Tais situações são indesejáveis, dada à indisponibilidade, por parte de muitas concessionárias, de vastas informações sobre consumidores, além daquelas relacionadas a faturamentos e dados cadastrais. Isto é notório na classe residencial, responsável por mais da metade das perdas comerciais (ARAÚJO, 2007).

A metodologia desenvolvida neste trabalho, assunto que é o tema deste capítulo, representa uma estimativa para contornar as restrições mencionadas, em uma abordagem simplificada e de fácil aplicação. É apresentada, a seguir, a formulação da proposta sob um ponto de vista de Descoberta de Conhecimento em Bases de Dados, sendo então relatadas as etapas e requisitos para sua aplicação. Ao final do capítulo é apresentado um exemplo numérico de aplicação da metodologia.

4.2 DEFINIÇÃO DE HORIZONTES PARA O PROCESSO DE DESCOBERTA DE CONHECIMENTO

Seguindo a linha de raciocínio defendida por Goldschmidt (2005), um problema desenvolvido no âmbito da Descoberta de Conhecimento em Bases de Dados pode ser visto sob dois aspectos: o diretivo e o estrutural. O caráter diretivo refere-se ao objetivo que

norteará todo o processo de Descoberta, bem como ao conjunto de dados usado, os recursos disponíveis para o desenvolvimento das tarefas e os resultados esperados. Já o aspecto estrutural diz respeito à caracterização e implementação propriamente dita das etapas de KDD.

Fundamentando-se nestas premissas, formula-se a seguir a metodologia proposta neste trabalho, também descrita de forma simplificada em Dos-Angelos (2009).

4.2.1 ASPECTOS DIRETIVOS

A metodologia foi desenvolvida objetivando a classificação de perfis anormais de consumo de energia a partir da análise de dados de faturamento mensais, bem como informações sobre o status da medição dos clientes avaliados por leituristas em campo. Estabeleceu-se a premissa de se usar o menor conjunto de atributos possível originado de dados que pudessem ser encontrados sem muitas dificuldades em qualquer concessionária de energia. Buscou-se um modelo que pudesse representar um perfil de consumo histórico "padrão" para cada cliente, de forma que a partir dele se pudesse mensurar a variação do perfil do consumidor tendo decorrido um dado intervalo de tempo, onde quanto maior a variação, maior seria a chance da existência de fraudes. Em relação às ferramentas para implantação da metodologia, buscaram-se utilizar aquelas de licença livre e que pudessem ser facilmente integradas ao domínio de aplicação. Estipulou-se, por fim, a assertividade mínima do modelo baseando-se na média das assertividades das metodologias analisadas na bibliografia, que foi de 70%, sendo que o modelo será testado em consumidores com status de fiscalização (autuado ou não autuado) já conhecido.

4.2.2 ASPECTO ESTRUTURAL: DESCRIÇÃO DA METODOLOGIA

Considerando as diretrizes estabelecidas, desenvolveu-se um modelo seguindo as seguintes etapas: (i) seleção de um vetor de atributos que representará um padrão geral de uso da energia; (ii) após o estabelecimento de uma data de referência, determinação do perfil de consumo histórico de cada consumidor e determinação de grupos de clientes com perfis semelhantes; e (iii) classificação dos perfis anormais, após decorrido um dado intervalo de tempo $\boldsymbol{k}$, considerando-se o grau de desvio do perfil histórico. A seguir, o detalhamento das etapas:

4.2.2.1 Seleção de variáveis relevantes

Foram definidas cinco variáveis para caracterizar um perfil individual de consumo – *M6, MAX6, DEV6, MS6 e APT6* – todas provenientes (exceto a última) de operações realizadas sobre o consumo mensal faturado (em kWh) de cada cliente. Em qualquer mês de referência, portanto, o "perfil" do cliente no dado mês será representado por um vetor com os cinco atributos mencionados. Para o cálculo de cada atributo, será considerado o próprio mês de referência e os cinco meses anteriores a ele. Segue a descrição das variáveis selecionadas:

- **M6**: representa a média de consumo (em kWh) nos seis meses considerados;
- **MAX6:** representa o pico de consumo (em kWh) nos seis meses considerados;
- **DEV6:** representa o desvio padrão dos consumos faturados no período;
- **MS6:** avalia a média dos valores faturados durante os seis meses (em kWh) em um dado setor (qualquer divisão lógica que agrupe um conjunto de consumidores, podendo ser uma quadra, bairro, distrito ou divisão similar);
- **APT6:** esta variável foi definida baseando-se no conceito de "apontamento". Um apontamento refere-se a uma notificação efetuada por um leiturista em campo, acerca da situação do fornecimento de energia do consumidor. Típicos apontamentos usados em muitas concessionárias são: *medidor com defeito*, *medidor ligado direto*, *casa desocupada*, *impedimento de leitura*. No modelo desenvolvido, buscou-se uma forma de combinar todos os apontamentos de um cliente durante um semestre em um índice no intervalo unitário, de forma que quanto mais próximo o índice da unidade, maior será o grau de anormalidade do cliente. Para isso, o cálculo da variável *APT6*, esquematizado na Tabela 4.2.1, passa pelos seguintes procedimentos: (i) levantamento de todos os apontamentos utilizados na área de concessão e associação de um valor no intervalo unitário a cada um; (ii) verificação, para cada cliente, dos apontamentos registrados no semestre considerado, e cálculo do somatório dos valores correspondentes.

Tabela 4.2.1 – Exemplo de cálculo da variável APT6 para um consumidor fictício

Apontamentos registrados no semestre	Grau de anormalidade associado
Medidor ligado direto	1
Casa Desocupada	0.2
Medidor com Defeito	0.3
Impedimento de Leitura	0.6
Leitura Normal	0
APT6	**2.1**

4.2.2.2 Etapa de clusterização

Objetiva definir o perfil de consumo padrão (ou histórico) de cada cliente, com base em um mês de referência e um intervalo retroativo de *k* meses pré-estabelecido. À primeira vista, o perfil histórico poderia ser representado diretamente pelo vetor de atributos calculado em uma data retroativa, entretanto, nota-se que, mesmo em um cliente individual, um perfil de consumo não se mantém rigorosamente uniforme ao longo do tempo. Este fato vem motivando a aplicação de técnicas de clusterização a problemas desta natureza, como visto na Seção 3.5, estratégia que também foi adotada na metodologia proposta.

Sendo assim, nesta etapa aplica-se um processo de clusterização englobando todos os perfis de consumidores situados no horizonte de *k* meses antes do mês de referência para análise, como ilustrado na Figura 4.2.1. Cada cliente analisado, representado por sua tupla – um ponto no espaço *n*-dimensional – entrará junto aos demais clientes do mês histórico em um processo de clusterização, com vistas à determinação de grupos de consumidores com características similares.

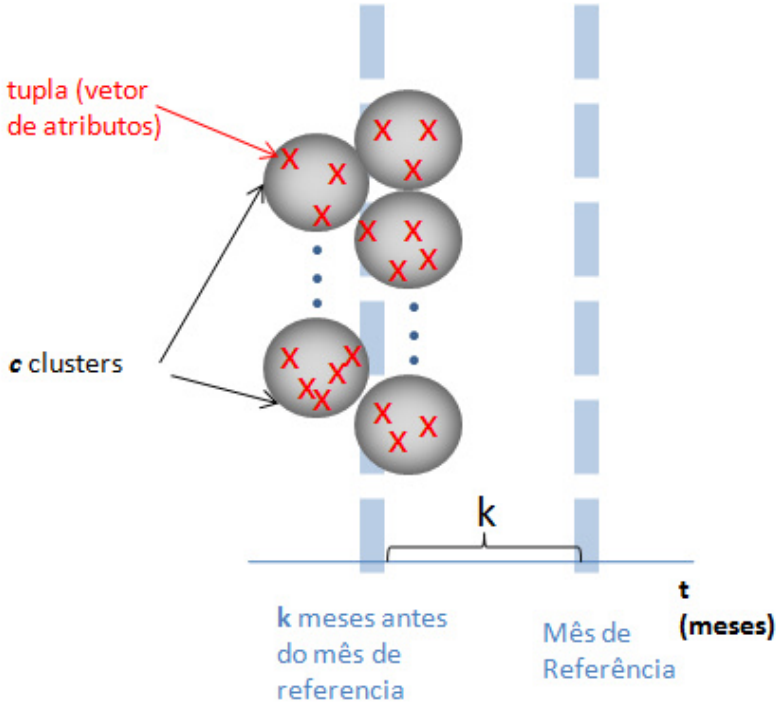


Figura 4.2.1 – Etapa de Clusterização

A metodologia desenvolvida requer a utilização de um algoritmo de clusterização que associe perfis a clusters por meio de graus de pertinência, de maneira que seja possível se aplicar, posteriormente, uma medida de distância aproximada entre os perfis históricos e perfis atuais. Algoritmos possíveis de serem aplicados nesta tarefa são o *Fuzzy C-Means* ou variantes.

A escolha do intervalo retroativo *k* dependerá das informações disponibilizadas à aplicação do modelo. Valores razoáveis seriam 12 ou 18, de forma que o perfil histórico fique bem caracterizado. Em relação à quantidade de clusters *c* e outros parâmetros, a definição dos valores a serem usados estará fortemente vinculada ao algoritmo escolhido. Por fim, ressalta-se a necessidade da normalização do vetor de atributos, de forma a evitar tendências indesejáveis (vide Seção 3.2.1.4).

Após a execução da clusterização, para um total de ***n*** perfis (formados por cinco atributos), tendo sido definidos ***c*** clusters, será gerada então a matriz de centros C (c x 5) e a matriz de partição *Fuzzy* U_0 (n x c), representada a seguir.

$$U_0 = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1j} & \cdots & \mu_{1c} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{i1} & \cdots & \mu_{ij} & \cdots & \mu_{ic} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n1} & \cdots & \mu_{nj} & \cdots & \mu_{nc} \end{bmatrix} \quad (4.2.1)$$

Esta última matriz, como observado na Seção 3.4.4, representa uma distância ponderada entre cada tupla e os centros dos clusters. Define-se, portanto, que cada linha (vetor) da matriz será tida então como o **perfil padrão (histórico)** de cada cliente.

4.2.2.3 Etapa de classificação

Havendo-se definido o perfil histórico de cada consumidor, procede-se nesta etapa com a classificação dos mesmos quanto à presença de anormalidades, partindo-se da hipótese que, quanto maior o desvio em relação ao perfil histórico, maior será a chance da existência de fraudes ou anormalidades em geral. O processo de classificação é desenvolvido da seguinte forma: inicialmente, calcula-se o vetor de atributos de cada consumidor tomando-se como base o próprio mês de referência. Após isso, busca-se então calcular os **perfis atuais**, ou seja, os vetores de partição *Fuzzy* atuais, dados pela Equação 3.4.16, que representa uma medida de distância entre uma tupla $\mathbf{x_i}$ e um centro de cluster $\mathbf{c_j}$. Definem-se como bases para o cálculo destes vetores, nesta etapa, os conjuntos de atributos atuais e os centros convergidos na etapa anterior (clusterização). Por fim, calcula-se a distância[4] entre os perfis (vetores de partição *Fuzzy*) atuais e históricos, que resultará em um vetor de índices que expressará o grau de anormalidade de cada consumidor, denominado $\mathbf{U_v}$ **(**vetor de anormalidade). O vetor é, então, normalizado e ordenado, sendo que os consumidores com índices acima de um determinado valor de corte (0.7, por exemplo) serão considerados altamente suspeitos de estarem irregulares.

Na Figura 4.2.2, ilustra-se o processo global de classificação, que pode ser resumido nos seguintes passos: (i) clusterização, em um horizonte histórico, das tuplas de entrada, que

[4] Refere-se a uma medida qualquer de distância entre dois pontos (vide Seção 3.4.2).

gerará os centros de convergência ($\mathbf{c_1, c_2, c_3, c_4, \ldots, c_n}$) e a matriz de partição *Fuzzy* histórica, que terá como um dos elementos o vetor $\mathbf{u_{i0}}$ (Figura 4.2.2a); (ii) avaliação da situação do consumidor no mês atual (*t*), e cálculo do vetor de partição *Fuzzy* atual (Figura 4.2.2b); (iii) cálculo da distância euclidiana entre os vetores $\mathbf{u_i}$ e $\mathbf{u_{i0}}$, que gerará o índice $\mathbf{U_V}$, representando a chance do mesmo estar em situação anormal (Figura 4.2.2c).

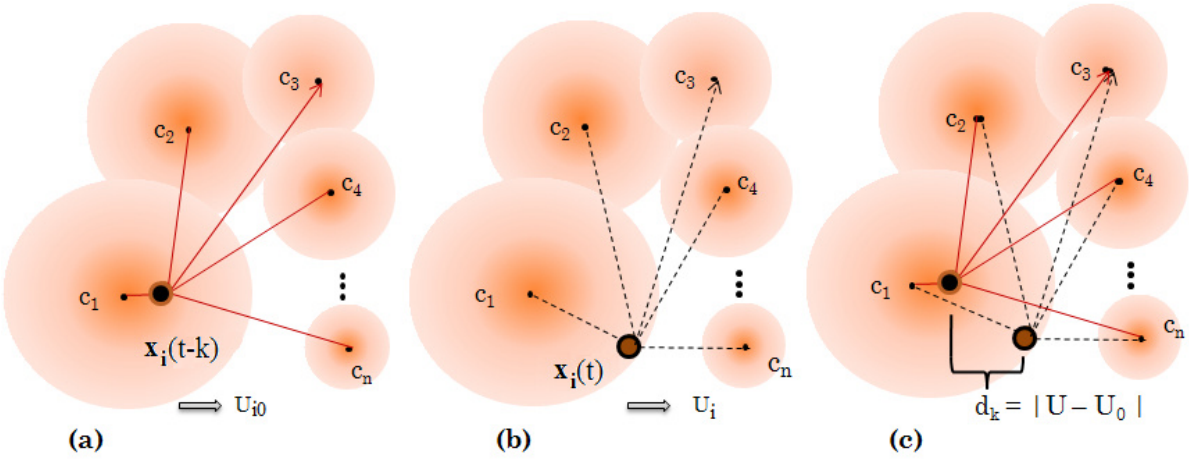


Figura 4.2.2 – Processo geral de classificação: **(a)** clusterização (mês t-*k*), que produzirá o vetor $\mathbf{u_0}$; **(b)** determinação dos perfis atuais de consumo e cálculo do vetor de partição atual **u**; **(c)** cálculo da distância entre os perfis históricos e atual, que resultará no vetor de anormalidade.

Por fim, os diagramas de blocos das três etapas, pré-processamento, clusterização e classificação estão ilustrados, respectivamente, nas Figuras Figura 4.2.3, Figura 4.2.4 e Figura 4.2.5. Na etapa de pré-processamento deverão ser informados inicialmente o mês de referência **t**, o horizonte retroativo **k** e o tipo de consumidor. Na saída do mesmo bloco serão geradas as matrizes de dados para o horizonte histórico $\mathbf{X_{t-k}}$ e para a data atual de classificação $\mathbf{X_t}$.

A matriz de dados $\mathbf{X_{t-k}}$ é o principal elemento de entrada da fase seguinte, clusterização. Nesta etapa deverão ser informados também os parâmetros do algoritmo de clusterização a ser usado, sendo que na figura Figura 4.2.4 está exemplificada a aplicação dos algoritmos FCM e GK ao problema. Ao final desta fase serão geradas a matriz de centros históricos $\mathbf{C_0}$ e a matriz de partição *fuzzy* histórica $\mathbf{U_0}$, que serão utilizadas na etapa posterior de classificação.

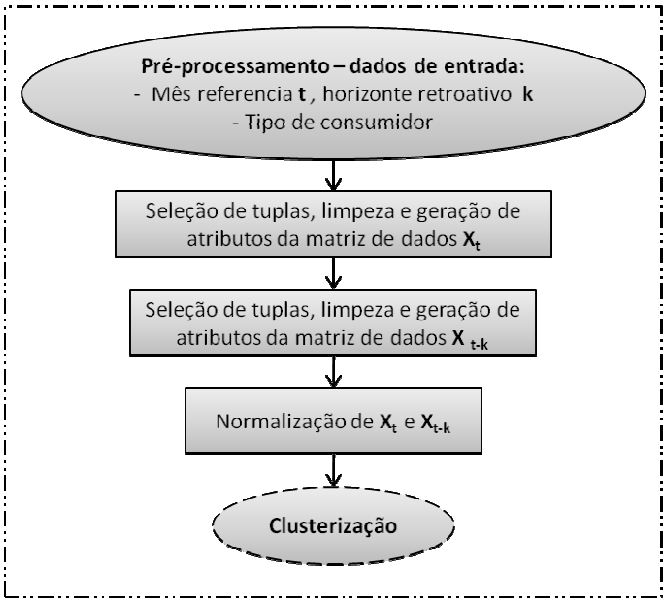


Figura 4.2.3 – Diagrama de blocos da etapa de pré-processamento.

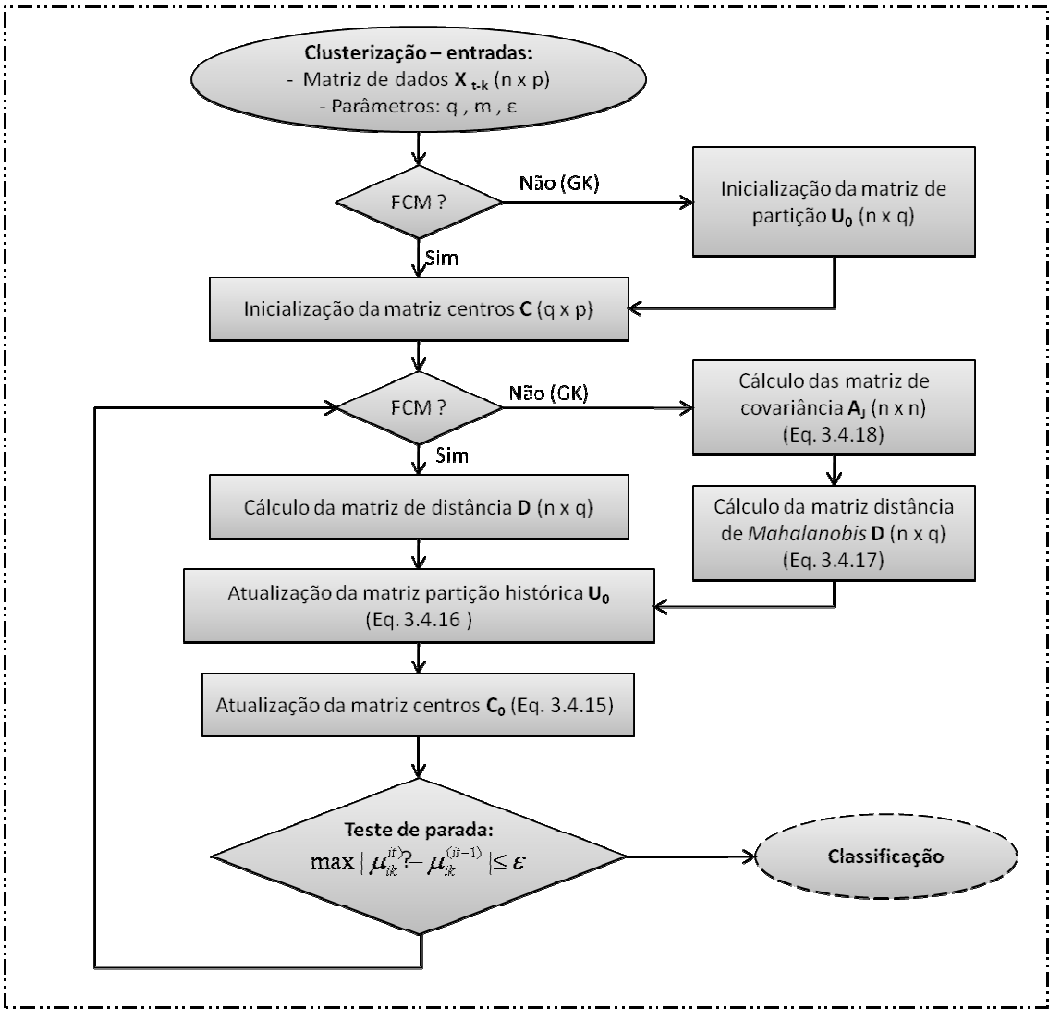


Figura 4.2.4 – Diagrama de blocos da etapa de clusterização

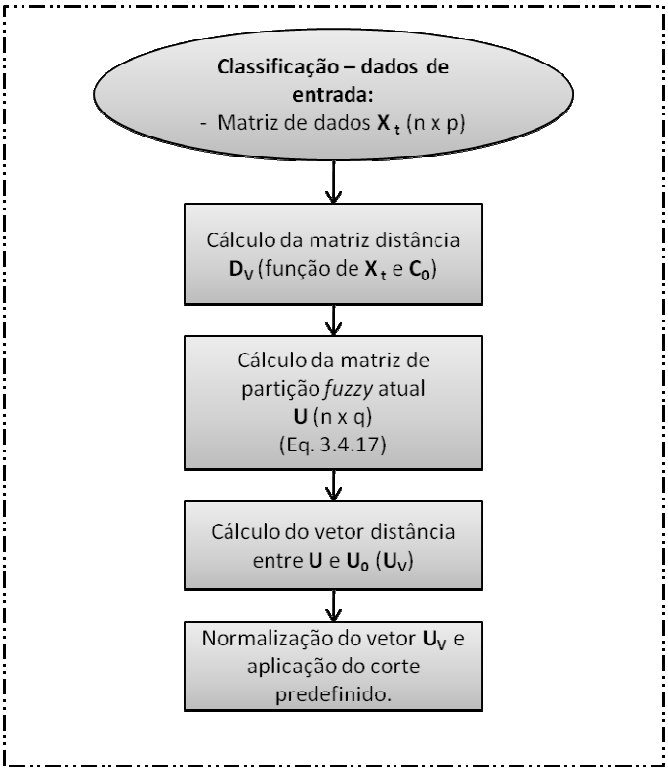


Figura 4.2.5 – Diagrama de blocos da etapa de classificação.

A última etapa do sistema desenvolvido, classificação, está baseada na distância euclidiana entre a matriz de partição histórica $\mathbf{U_0}$ e a atual **U –** calculada em função dos centros de clusterização $\mathbf{C_0}$ e das tuplas no mês de referência, $\mathbf{X_t}$. Será gerado então um vetor de índices de anormalidades ($\mathbf{U_v}$) que representarão o grau de anormalidade de cada cliente analisado, como ilustrado no diagrama da Figura 4.2.5.

4.3 EXEMPLO NUMÉRICO DE APLICAÇÃO

Nesta seção é descrito um exemplo de aplicação da metodologia, para um conjunto de 20 consumidores residenciais.

- **Etapa de Pré-Processamento**

Inicialmente são definidos o mês de referência para fiscalização e o intervalo retroativo para clusterização; ***t** = 01/07/09* e ***k** = 18*. Sendo assim, considerando os seis meses anteriores a **t** (incluindo ele próprio), calculam-se os atributos para a matriz de dados , por exemplo: M6 – média do consumo faturado em kWh entre Jul/09 e Fev/09. Os 20 clientes

selecionados já apresentam todas as condições para a execução do algoritmo de mineração de dados (do contrário, as tarefas de limpezas e correções de valores deveriam ser aplicadas). Esta matriz será utilizada na etapa de classificação.

Semelhantemente, com a definição do intervalo histórico retroativo *k=18*, define-se o mês de referência para clusterização (***t-k*** = Jan/08), a partir do qual são calculados os atributos para a matriz $\mathbf{X_{t-k}}$ a ser utilizada na etapa de clusterização. A matriz resultante está mostrada na Tabela 4.3.1.

Tabela 4.3.1 – Matriz de entrada $\mathbf{X_{t-k}}$ do exemplo numérico.

Cliente	M6 (kWh)	MAX6 (kWh)	DEV6	M6SETOR (kWh)	APT6	Classe
1	77.500	83	6.317	66.050	0	**irregular**
2	51.500	82	15.783	87.347	0	**normal**
3	52.500	62	4.806	59.284	0	**normal**
4	127.667	166	45.284	66.912	0	**normal**
5	108.667	127	15.908	91.750	0	**irregular**
6	68.833	75	4.997	58.688	0	**irregular**
7	61.500	76	10.858	54.667	0	**irregular**
8	72.333	86	11.308	83.626	0	**irregular**
9	152.000	174	13.726	168.314	0	**normal**
10	93.333	102	8.116	172.333	0	**normal**
11	127.000	145	14.142	74.238	0	**normal**
12	104.167	137	27.491	60.921	0	**irregular**
13	30.000	30	0.000	65.017	0.4	**irregular**
14	122.500	137	9.731	92.144	0	**normal**
15	73.833	238	81.224	60.222	0	**irregular**
16	97.000	113	9.695	101.755	0.6	**irregular**
17	31.333	34	1.633	42.788	1	**irregular**
18	38.500	69	15.437	55.778	0	**normal**
19	91.667	104	11.639	77.655	0	**normal**
20	84.333	192	64.593	97.322	0	**normal**

. Por fim, as matrizes de dados deverão ser normalizadas. Por exemplo, para normalizar o atributo M6 do cliente 1 ($M6_1 = 77{,}5$), da matriz $\mathbf{X_{t-k}}$, procede-se da seguinte forma:

(i) Determinação dos valores máximo e mínimo do atributo: $\max(M6) = 152$; $\min(M6) = 30$

Aplicação da normalização linear: $\overline{M6}_1 = \frac{M6_1 - \min(M6)}{\max(M6) - \min(M6)} = \frac{77{,}5-30}{152-30} = 0{,}389$

As matrizes normalizadas $\mathbf{X_t}$ e $\mathbf{X_{t-k}}$ estão mostradas nas Tabelas Tabela 4.3.2 e Tabela 4.3.3.

Tabela 4.3.2 – Matriz de dados $\mathbf{X_{t-k}}$ normalizada (dimensões: $\mathbf{20 \times 5}$).

Cliente	M6 (kWh)	MAX6 (kWh)	DEV6	M6SETOR (kWh)	APT6
1	0.389	0.255	0.078	0.180	0
2	0.176	0.250	0.194	0.344	0
3	0.184	0.154	0.059	0.127	0
4	0.801	0.654	0.558	0.186	0
5	0.645	0.466	0.196	0.378	0
6	0.318	0.216	0.062	0.123	0
7	0.258	0.221	0.134	0.092	0
8	0.347	0.269	0.139	0.315	0
9	1.000	0.692	0.169	0.969	0
10	0.519	0.346	0.100	1.000	0
11	0.795	0.553	0.174	0.243	0
12	0.608	0.514	0.338	0.140	0
13	0.000	0.000	0.000	0.172	0,4
14	0.758	0.514	0.120	0.381	0
15	0.359	1.000	1.000	0.135	0
16	0.549	0.399	0.119	0.455	0,6
17	0.011	0.019	0.020	0.000	1
18	0.070	0.188	0.190	0.100	0
19	0.505	0.356	0.143	0.269	0
20	0.445	0.779	0.795	0.421	0

Tabela 4.3.3 – Matriz de dados $\mathbf{X_t}$ normalizada (dimensões: $\mathbf{20 \times 5}$).

Cliente	M6 (kWh)	MAX6 (kWh)	DEV6	M6SETOR (kWh)	APT6
1	0.000	0.000	0.000	0.061	0.333
2	0.020	0.033	0.063	0.048	0.000
3	0.059	0.058	0.030	0.062	0.000
4	0.120	0.106	0.047	0.075	0.000
5	0.000	0.000	0.000	0.000	0.000
6	0.022	0.048	0.100	0.064	0.000
7	0.029	0.035	0.052	0.064	0.000
8	0.060	0.065	0.063	0.135	0.000
9	0.139	0.122	0.056	0.267	0.000
10	0.070	0.069	0.060	0.287	0.000
11	0.134	0.117	0.039	0.106	0.000
12	0.005	0.013	0.033	0.074	0.000
13	0.000	0.000	0.000	0.079	0.333
14	0.093	0.091	0.056	0.114	0.000
15	0.060	0.054	0.027	0.104	0.250
16	0.020	0.050	0.112	0.170	0.000
17	0.000	0.000	0.000	0.061	0.000
18	0.004	0.019	0.045	0.060	0.167
19	0.073	0.070	0.040	0.110	0.000
20	0.021	0.061	0.129	0.113	0.000

- **Etapa de Clusterização**

Nesta fase foi utilizado o algoritmo de clusterização *Fuzzy C-Means* nos parâmetros $q = 2, m = 2$ e $\varepsilon = 0.01$. A matriz de dados no horizonte histórico $\mathbf{X_{t-k}}$ entrará em um processo de clusterização, que produzirá em um processo iterativo a matriz de centros convergidos $\mathbf{C_0}$ e a matriz de partição Fuzzy $\mathbf{U_0}$, ambas mostradas nas Tabelas.

Tabela 4.3.4 – Matriz de centros convergidos $\mathbf{C_0}$ (dimensões: $\mathbf{2 \times 5}$).

Cluster	M6 (kWh)	MAX6 (kWh)	DEV6	M6SETOR (kWh)	APT6
1	0.650	0.564	0.328	0.378	0.047
2	0.255	0.231	0.129	0.207	0.105

Tabela 4.3.5 – Matriz de partição fuzzy histórica $\mathbf{U_0}$ (dimensões: $\mathbf{20 \times 2}$).

Pertinência Cliente	Cluster 1	Cluster2
1	0.110	0.890
2	0.105	0.895
3	0.060	0.940
4	0.846	0.154
5	0.895	0.105
6	0.068	0.932
7	0.059	0.941
8	0.129	0.871
9	0.725	0.275
10	0.588	0.412
11	0.862	0.138
12	0.807	0.193
13	0.180	0.820
14	0.862	0.138
15	0.636	0.364
16	0.517	0.483
17	0.341	0.659
18	0.098	0.902
19	0.453	0.547
20	0.730	0.270

- **Etapa de Classificação**

Nesta etapa, os perfis de consumo do mês atual são comparados com os perfis determinados no horizonte histórico, da seguinte forma:

(i) Cálculo da matriz $\mathbf{D_t}$ (distância euclidiana entre a matriz de dados atual $\mathbf{X_t}$ e os centros convergidos na etapa de clusterização $\mathbf{C_0}$, relativos aos perfis históricos). Esta matriz está mostrada na Tabela 4.3.6, onde cada consumidor está representado por um vetor de 2 atributos. Os cálculos dos elementos referentes ao cliente 4, sublinhados na tabela, estão descritos a seguir:

$\mathbf{D_t(4,1)}$: distância do perfil atual 4 ao cluster convergido 1:

$$= \sqrt{(0.120 - 0.650)^2 + (0.106 - 0.564)^2 + (0.047 - 0.328)^2 + (0.075 - 0.378)^2 + (0 - 0.047)^2}$$

$$= 0.815$$

$\mathbf{D_t(4,2)}$: distância do perfil atual 4 ao cluster convergido 2:

$$= \sqrt{(0.120 - 0.255)^2 + (0.106 - 0.231)^2 + (0.047 - 0.129)^2 + (0.075 - 0.207)^2 + (0 - 0.105)^2}$$

$$= 0.263$$

Tabela 4.3.6 – Matriz $\mathbf{D_t}$ (dimensões: $\mathbf{20 \times 2}$).

Distância Elemento	Cluster1	Cluster2
1	1.016	0.457
2	0.928	0.368
3	0.893	0.333
4	0.815	0.263
5	0.997	0.435
6	0.903	0.347
7	0.918	0.356
8	0.854	0.294
9	0.739	0.213
10	0.815	0.287
11	0.791	0.239
12	0.950	0.385
13	1.010	0.451
14	0.825	0.266
15	0.903	0.334
16	0.868	0.318
17	0.975	0.409
18	0.954	0.375
19	0.857	0.296
20	0.872	0.322

(ii) Cálculo da matriz de partição fuzzy atual **U,** a partir da matriz $\mathbf{D_t}$, de acordo com a Equação 3.4.17. O cálculo do elemento $U_v(1,1)$ é exemplificado como segue:

$$U_v(1,1) = [D_t(1,1)]^{-2}/[(D_t(1,1)^{-2} + D_t(2,1)]^{-2}$$
$$= 1.016^{-2}/(1.016^{-2} + 0.457^{-2}) = 0.968/(0.968 + 4.788)$$
$$= 0.168$$

A matriz resultante está mostrada na Tabela 4.3.7.

Tabela 4.3.7 – Matriz de partição fuzzy atual, **U** ($\mathbf{20 \times 2}$).

Pertinência Cliente	Cluster 1	Cluster2
1	0.168	0.832
2	0.136	0.864
3	0.122	0.878
4	0.094	0.906
5	0.160	0.840
6	0.129	0.871
7	0.131	0.869
8	0.106	0.894
9	0.077	0.923
10	0.111	0.889
11	0.083	0.917
12	0.141	0.859
13	0.166	0.834
14	0.094	0.906
15	0.120	0.880
16	0.118	0.882
17	0.150	0.850
18	0.134	0.866
19	0.106	0.894
20	0.120	0.880

(iii) Cálculo do vetor de anormalidade $\mathbf{U_v}$, referente à distancia euclidiana entre as matrizes de partição atual (**U**) e histórica ($\mathbf{U_0}$). O vetor de anormalidade original e o normalizado (normalização linear) estão mostrados na Tabela 4.3.8.

Tabela 4.3.8 – Vetor de anormalidade (original $\mathbf{U_V}$ e normalizado $\overline{\mathbf{U_v}}$)

Cliente	U_v	$\overline{U_v}$
1	0.724	0.657
2	0.759	0.780
3	0.821	1.000
4	0.754	0.762
5	0.737	0.703
6	0.805	0.946
7	0.814	0.975
8	0.765	0.802
9	0.678	0.489
10	0.565	0.085
11	0.781	0.857
12	0.668	0.455
13	0.654	0.403
14	0.769	0.815
15	0.571	0.107
16	0.541	0.000
17	0.544	0.011
18	0.769	0.817
19	0.561	0.071
20	0.628	0.313

(iv) Ordenação de $\overline{\mathbf{U_v}}$, aplicação do corte adequado e verificação da assertividade, conforme mostrado na Tabela 4.3.9.

Tabela 4.3.9 – Vetor de anormalidade com indicação de classes e nível de corte utilizado.

Cliente	$\overline{\mathbf{U_v}}$	Classe real
3	1.000	normal
7	0.975	irregular
6	0.946	irregular
11	0.857	normal
18	0.817	normal
14	0.815	normal
8	0.802	irregular
2	0.780	normal
4	0.762	normal
5	0.703	irregular
1	0.657	irregular
9	0.489	normal
12	0.455	irregular
13	0.403	irregular
20	0.313	normal
15	0.107	irregular
10	0.085	normal
19	0.071	normal
17	0.011	irregular
16	0.000	irregular

Neste exemplo teste, acima do corte 0,9 foram encontrados 2 consumidores irregulares e um normal. Sendo assim, o nível de assertividade no dado corte foi:

$$ASS09 = \frac{2}{2+1} = 66{,}66\%$$

O desempenho do sistema em um universo de poucos clientes é inferior em relação a um universo maior, devido à perda de representatividade no processo de clusterização.

5 ESTUDO DE CASO E DISCUSSÃO DOS RESULTADOS

O sistema desenvolvido foi validado utilizando uma base de dados real, compondo uma dada região formada por 376 bairros, totalizando 29382 consumidores, dentre residenciais, comerciais, industriais e poder público, contendo informação de 17 meses de faturamento. Selecionaram-se para as simulações somente os clientes previamente fiscalizados, isto é, que possuíam informação quanto à situação de fraude, ou seja, irregular ou normal. Embora o conhecimento prévio de classes não seja pré-requisito para o processo de clusterização, elas serviram para se avaliar a eficiência da metodologia proposta, dado que, a princípio, não foram efetuados testes de campo.

O estudo de caso foi realizado em três etapas: (i) caracterização dos dados utilizados; (ii) tarefas de pré-processamento aplicadas; e (3) execução do algoritmo proposto e análise dos resultados e da assertividade do modelo. Segue nas próximas seções a pormenorização de cada etapa.

5.1 CARACTERIZAÇÃO DOS ATRIBUTOS UTILIZADOS

O conjunto de atributos utilizado originou-se de três tabelas principais: cadastro, faturamento e fiscalização (ver Figura 5.1.1). Da tabela cadastro, extraíram-se variáveis relacionadas à identificação dos consumidores, como o número individual de registro (aqui denominado unidade consumidora – UC), o nome do cliente e endereço. Na segunda tabela, constavam variáveis associadas ao faturamento mensal de cada cliente, como o valor faturado (em kWh), mês de leitura e apontamentos de campo. Já da tabela fiscalização, foram utilizados dados sobre inspeções de fiscalização efetuadas em cada consumidor, como a data de referência de fiscalização e o tipo de irregularidade, quando encontrado.

A descrição dos atributos utilizados no sistema consta na Tabela 5.1.1. A partir dos atributos primários se derivaram então os 5 atributos mencionados no capítulo anterior, que modelam o perfil de consumo para cada consumidor.

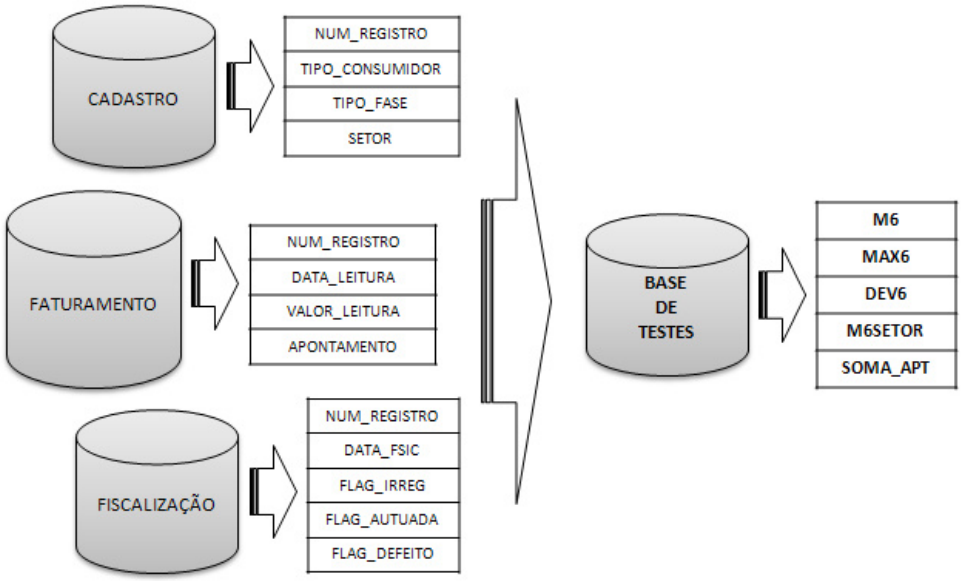


Figura 5.1.1 – Diagrama relacional com as tabelas e atributos utilizados no sistema

Tabela 5.1.1 – Atributos Primários Utilizados na metodologia

TABELA	ATRIBUTO	DESCRIÇÃO
CADASTRO	NUM_REGISTRO	Identificação da unidade consumidora (UC)
CADASTRO	TIPO_CONSUMIDOR	Tipo do Consumidor (residencial, comercial, etc..)
CADASTRO	TIPO_FASE	Fase de Alimentação (Mono, Bi ou Trifásica)
CADASTRO	SETOR	Região onde se encontra a UC
FATURAMENTO	DATA_LEITURA	Data da Leitura do Consumo
FATURAMENTO	VALOR_LEITURA	Valor lido (kWh)
FATURAMENTO	APONTAMENTO	Notificações de campo (ver Seção 4.2.2)
FISCALIZAÇÃO	DATA_FISC	Data de execução de fiscalização
FISCALIZAÇÃO	FLAG_IRREG	Presença ou ausência de irregularidade
FISCALIZAÇÃO	FLAG_AUTUADA	Presença ou ausência de fraude
FISCALIZAÇÃO	FLAG_DEFEITO	Constatação de defeito no medidor

5.2 ETAPA DE PRÉ-PROCESSAMENTO

Para a utilização efetiva e adaptação das informações disponibilizadas à metodologia desenvolvida, foi necessário efetuar o tratamento e formatação dos dados de entrada. Para isso, foram aplicados processos de pré-processamento de dados (Seção 3.2.1) como, por exemplo, a exclusão das UC's com campos nulos ou inválidos, a retirada de UC's fiscalizadas mais de uma vez e a seleção do subconjunto de UC's enquadradas no horizonte de classificação definido.

Com base na quantidade de UC's que atingiram os critérios considerados, foi ainda definido o horizonte das datas de referência de fiscalização — 01/01/2008 a 01/05/2009 — e os tipos de classes de consumidores escolhidos para análise — residenciais e comerciais. Por fim, na etapa de geração de atributos foram calculados, para cada consumidor selecionado, os cinco atributos mencionados na Seção 4.2.2, sendo criada uma tabela à parte (base para testes) contendo todas as informações necessárias para a execução posterior do algoritmo de classificação. As etapas de classificação, pré-processamento, mineração de dados e classificação foram automatizadas com um software próprio desenvolvido em linguagem C++ na plataforma Visual C+ Express 2005 (MICROSOFT, 2005), portando interface *Oracle C++ Call Interface* (OCCI, 2007) para comunicação com sistema de banco de dados Oracle 10G XE (ORACLE, 2007), onde os dados dos clientes estiveram armazenados. As ferramentas usadas são de licença livre.

5.3 APLICAÇÃO E RESULTADOS

Avaliou-se o modelo de classificação de anormalidades utilizando, respectivamente, seis tipos de simulações: (i) domínio de clientes residenciais; (ii) universo de clientes comerciais; (iii) domínio conjunto de consumidores residenciais e comerciais; (iv) diferentes distribuições de classes; (v) distintos níveis de corte para o índice de irregularidade; e (vi) distintos horizontes de tempo.

As simulações foram efetuadas considerando os seguintes valores padrões:

- Mês de referência para classificação (**t**): 01/07/2009;
- Intervalo retroativo (**k**): 18;

- Mês de referência para clusterização (**t-k**): 01/02/2008;
- Tolerância de convergência para clusterização (ε): 0,01;
- Corte padrão para análise de assertividade: 0,8.

Durante a etapa de clusterização, foram aplicados os algoritmos Fuzzy *C-Means* e Fuzzy GK, tendo sido usado também o mapeamento modificado de Sammon (FUZZSAM) como suporte para a visualização dos clusters e graus de pertinência atribuídos. Para cada configuração simulada foram determinados os parâmetros ótimos **c** e **m**, com base nos índices de desempenho SC (índice de partição), S (índice de separação) e XB (índice *Xie-Beni*), descritos na Seção 3.4.5, e principalmente nas métricas ASS08 (assertividade ao corte 0,8) e aSENS08 (sensibilidade ao corte de 0,8), particularmente relacionadas a sistemas de detecção de fraudes. Todas as simulações foram efetuadas com distribuições iguais de classes (exceto na simulação 4), de forma a minimizar problemas de prevalência de classes.

5.3.1 SIMULAÇÃO NO DOMÍNIO DE CLIENTES RESIDENCIAIS

Os clientes residenciais na base de testes portando mês de referência **t** = 01/09/07 foram um total de 1544. As matrizes de dados, nesse caso, foram de dimensões 1544 × 5. Os resultados das simulações com os algoritmos FCM e GK são descritos a seguir.

5.3.1.1 Simulação com o Algoritmo Fuzzy C-Means

Inicialmente procedeu-se a seleção dos parâmetros ótimos (quantidade de clusters – *c* e expoente fuzzificador – *m*) para este algoritmo. Para isso, efetuaram-se várias simulações nos seguintes intervalos: $2 \le c \le 11$ e $1.1 \le m \le 2.5$. Os valores resultantes para os índices XB, SC, ASS08 e SENS08, em cada configuração,estão indicados, respectivamente, nas Tabela 5.3.1 a Tabela 5.3.4, estando sublinhados os três melhores valores de cada índice.

Tabela 5.3.1 – Índice XB: Resultados para o Algoritmo FCM

m \ c	2	3	4	5	6	7	8	9	10	11
1.1	3.9582	4.2034	6.6956	6.3245	6.9028	6.0358	6.5880	6.5242	7.3983	7.3123
1.2	3.7377	3.8198	6.0277	5.5334	5.9590	5.1228	5.4988	5.3861	6.1301	5.9792
1.3	3.5382	3.4832	5.4584	4.8736	5.1844	4.3838	4.6288	4.4843	5.1331	4.9398
1.4	3.3570	3.1867	4.9696	4.3190	4.5434	3.7809	3.9284	3.7645	4.3417	4.1216
1.5	3.1918	2.9245	4.5474	3.8496	4.0091	3.2855	3.3603	3.1857	3.7077	3.4719
1.6	3.0407	2.6919	4.1803	3.4498	3.5606	2.8755	2.8962	2.7171	3.1953	2.9515
1.7	2.9020	2.4848	3.8595	3.1072	3.1819	2.5339	2.5145	2.3352	2.7778	2.5315
1.8	2.7744	2.2998	3.5776	2.8119	2.8599	2.2476	2.1986	2.0220	2.4348	2.1897
1.9	2.6566	2.1342	3.3288	2.5562	2.5847	2.0060	1.9354	1.7636	2.1508	1.9096
2	2.5477	1.9854	3.1080	2.3336	2.3481	1.8010	1.7149	1.5490	1.9140	1.6783
2.1	2.4467	1.8515	2.9113	2.139	2.1435	1.626	1.5291	1.3697	1.715	1.4859
2.2	2.3529	1.7306	2.7354	1.968	1.9658	1.4758	1.3714	1.219	1.5466	1.3249
2.3	2.2655	1.6211	2.5773	1.8171	1.8105	1.3461	1.237	1.0916	1.4031	1.189
2.4	2.184	1.5219	2.4348	1.6834	1.6742	1.2335	1.1217	0.9832	1.28	1.0737
2.5	2.1079	1.4316	2.3057	1.5646	1.554	1.1353	1.0223	0.8905	1.1737	0.9752

Tabela 5.3.2 – Índice SC: Resultados para o Algoritmo FCM

m \ c	2	3	4	5	6	7	8	9	10	11
1.1	8.8778	6.4394	1.5909	1.4199	1.3346	1.2547	1.1916	1.1897	0.5971	0.5991
1.2	8.2849	5.7916	1.3924	1.2202	1.1260	1.0436	0.9799	0.9660	0.4798	0.4778
1.3	7.7575	5.2290	1.2307	1.0585	0.9598	0.8776	0.8144	0.7925	0.3921	0.3874
1.4	7.2855	4.7377	1.0974	0.9260	0.8260	0.7455	0.6837	0.6565	0.3255	0.3188
1.5	6.8606	4.3063	0.9861	0.8164	0.7170	0.6393	0.5793	0.5490	0.2741	0.2660
1.6	6.4761	3.9259	0.8922	0.7246	0.6274	0.5529	0.4951	0.4631	0.2337	0.2248
1.7	6.1266	3.5890	0.8122	0.6472	0.5531	0.4822	0.4266	0.3939	0.2016	0.1921
1.8	5.8074	3.2895	0.7434	0.5813	0.4909	0.4236	0.3705	0.3378	0.1758	0.1660
1.9	5.5149	3.0223	0.6837	0.5249	0.4384	0.3748	0.3241	0.2919	0.1547	0.1448
2	5.2459	2.7832	0.6316	0.4762	0.3939	0.3338	0.2854	0.2541	0.1373	0.1275
2.1	4.9977	2.5686	0.5858	0.4339	0.3558	0.2992	0.2531	0.2228	0.1229	0.1132
2.2	4.7680	2.3755	0.5452	0.3970	0.3230	0.2696	0.2257	0.1966	0.1107	0.1012
2.3	4.5549	2.2013	0.5091	0.3646	0.2946	0.2443	0.2026	0.1746	0.1005	0.0912
2.4	4.3567	2.0437	0.4768	0.3360	0.2699	0.2225	0.1827	0.1560	0.0917	0.0827
2.5	4.1720	1.9009	0.4478	0.3107	0.2482	0.2035	0.1657	0.1402	0.0841	0.0754

Tabela 5.3.3 – Métrica ASS08: Resultados para o Algoritmo FCM

m \ c	2	3	4	5	6	7	8	9	10	11
1.1	0.5023	0.2593	0.4737	0.4667	0.1250	0.4043	0.1667	1.0000	0.8750	0.0000
1.2	0.4964	0.2535	0.5217	0.5714	0.1111	0.4000	0.1667	1.0000	0.8696	0.0000
1.3	0.4974	0.2576	0.5714	0.5714	0.1111	0.3929	0.1667	1.0000	0.8696	0.0000
1.4	0.4958	0.3000	0.5714	0.6000	0.1250	0.4074	0.2000	1.0000	0.8750	0.0000
1.5	0.5045	0.2927	0.5714	0.5625	0.1667	0.4074	0.2500	1.0000	0.8750	0.0000
1.6	0.5024	0.3158	0.5789	0.5294	0.1667	0.4074	0.2500	1.0000	0.8750	0.0000
1.7	0.4983	0.3243	0.5789	0.5294	0.1667	0.4074	0.2500	1.0000	0.8750	0.0000
1.8	0.4936	0.3636	0.5556	0.5294	0.2000	0.4151	0.2500	1.0000	0.8750	0.0000
1.9	0.4971	0.4063	0.5882	0.5294	0.2000	0.4231	0.2500	1.0000	0.8750	0.0000
2	0.5010	0.4194	0.5556	0.5625	0.2500	0.4286	0.2500	1.0000	0.8750	0.0000
2.1	0.5011	0.4375	0.5556	0.5333	0.2500	0.4167	0.2500	1.0000	0.8750	0.0000
2.2	0.4978	0.4242	0.5556	0.5333	0.2000	0.4130	0.2500	1.0000	0.8750	0.0000
2.3	0.4989	0.4412	0.5882	0.5333	0.2000	0.4130	0.2500	1.0000	0.8750	0.0000
2.4	0.4977	0.4167	0.5882	0.5333	0.2000	0.4130	0.2500	1.0000	0.8750	0.0000
2.5	0.4951	0.4054	0.5882	0.5333	0.2000	0.4130	0.2500	1.0000	0.8750	0.0000

Tabela 5.3.4 – Métrica SENS08: Resultados para o Algoritmo FCM

m \ c	2	3	4	5	6	7	8	9	10	11
1.1	0.5777	0.0272	0.0117	0.0091	0.0013	0.0246	0.0013	0.0104	0.0272	0.0000
1.2	0.5298	0.0233	0.0155	0.0104	0.0013	0.0285	0.0013	0.0104	0.0259	0.0000
1.3	0.4974	0.0220	0.0155	0.0104	0.0013	0.0285	0.0013	0.0104	0.0259	0.0000
1.4	0.4547	0.0194	0.0155	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
1.5	0.4365	0.0155	0.0155	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
1.6	0.4028	0.0155	0.0142	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
1.7	0.3769	0.0155	0.0142	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
1.8	0.3497	0.0155	0.0130	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
1.9	0.3342	0.0168	0.0130	0.0117	0.0013	0.0285	0.0013	0.0104	0.0272	0.0000
2	0.3174	0.0168	0.0130	0.0117	0.0013	0.0272	0.0013	0.0104	0.0272	0.0000
2.1	0.3070	0.0181	0.0130	0.0104	0.0013	0.0259	0.0013	0.0104	0.0272	0.0000
2.2	0.2927	0.0181	0.0130	0.0104	0.0013	0.0246	0.0013	0.0104	0.0272	0.0000
2.3	0.2863	0.0194	0.0130	0.0104	0.0013	0.0246	0.0013	0.0104	0.0272	0.0000
2.4	0.2746	0.0194	0.0130	0.0104	0.0013	0.0246	0.0013	0.0104	0.0272	0.0000
2.5	0.2642	0.0194	0.0130	0.0104	0.0013	0.0246	0.0013	0.0104	0.0272	0.0000

As quatro tabelas foram registradas para se exemplificar o processo de definição dos parâmetros ótimos da etapa de clusterização. Para a escolha da quantidade ótima de clusters (q), procedeu-se da seguinte forma: observou-se inicialmente, segundo a Tabela 5.3.1, que o valor de otimização do índice XB foi $c=11$. Entretanto, notou-se que, para esta quantidade de clusters, a assertividade e a sensibilidade foram nulas (observar Tabelas Tabela 5.3.3 eTabela 5.3.4). O valor que atendeu satisfatoriamente às métricas foi $c=10$, sendo, portanto, escolhido como valor ótimo.

Para a determinação do valor ótimo do expoente fuzzificador *m*, efetuou-se a plotagem das quatro grandezas para diferentes valores de *m*, mantendo-se $c=10$, como ilustrado na Figura 5.3.1. Notou-se no gráfico a tendência de otimização das métricas XB, S e SC para valores de m maiores que 2,5. Verificou-se também, segundo a Tabela 5.3.3, que, para valores de *m* acima de 2, não houve decréscimo da assertividade. Além disso, como visto na Seção 3.4.4, valores acima de 2,0 diminuem o caráter fuzzy dos conjuntos clusterizados. Sendo assim, escolheu-se como expoente fuzzificador ótimo o valor 2,0.

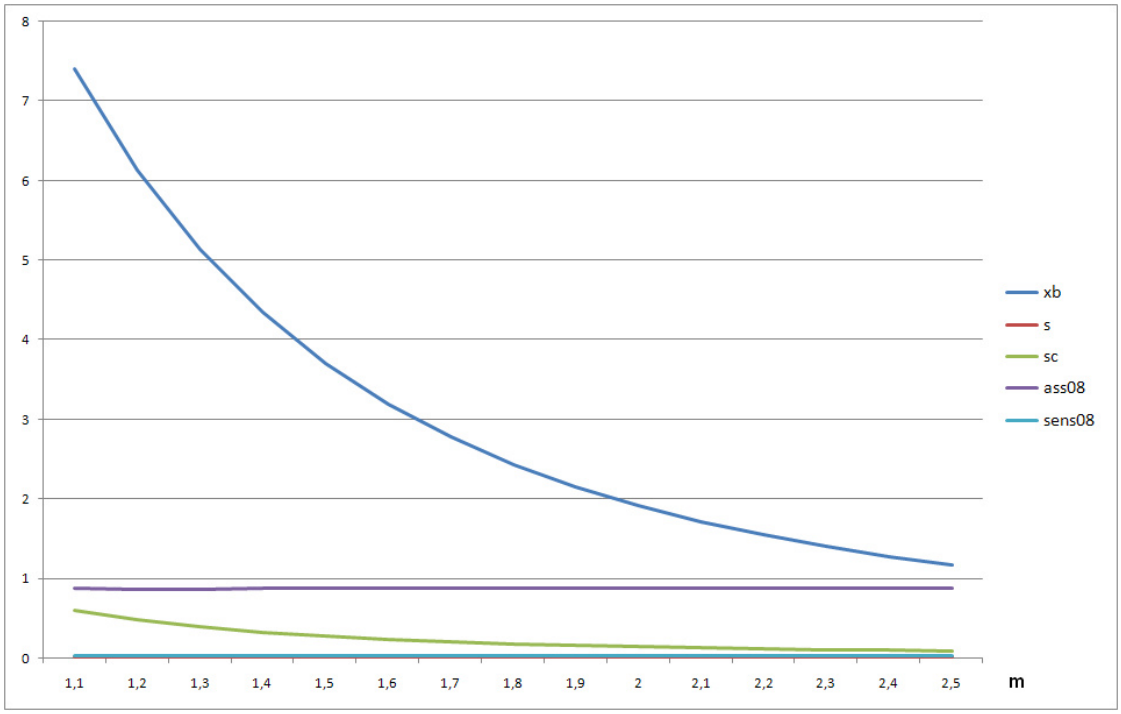


Figura 5.3.1 – Trajetória dos índices de desempenho para o FCM, para diferentes valores de *m*, considerando $c=10$.

- **Mapeamento FUZZSAM da configuração ótima**

Havendo simulado o algoritmo FCM para a configuração ótima, efetuou-se então o mapeamento no espaço bidimensional através da projeção modificada de Sammon, ilustrado na Figura 5.3.2. Os pontos azuis representam os perfis (vetor de 5 atributos) dos clientes analisados (um total de 1544), projetados no espaço bi-dimensional. Os pontos vermelhos representam os centros dos clusters convergidos após a execução do FCM, sendo que os contornos representam os valores de pertinência recalculados. Percebe-se no gráfico a tentativa do Algoritmo FCM em abarcar o maior número de perfis nos clusters. Nota-se também a baixa capacidade do algoritmo em tratar elementos *outliers*, inclusive associando um cluster inteiro a uma gama de pontos *outliers* (cluster superior). Este último fato não representa uma desvantagem na presente metodologia, dado que o caráter discrepante é associado à chance de fraude, na etapa classificatória.

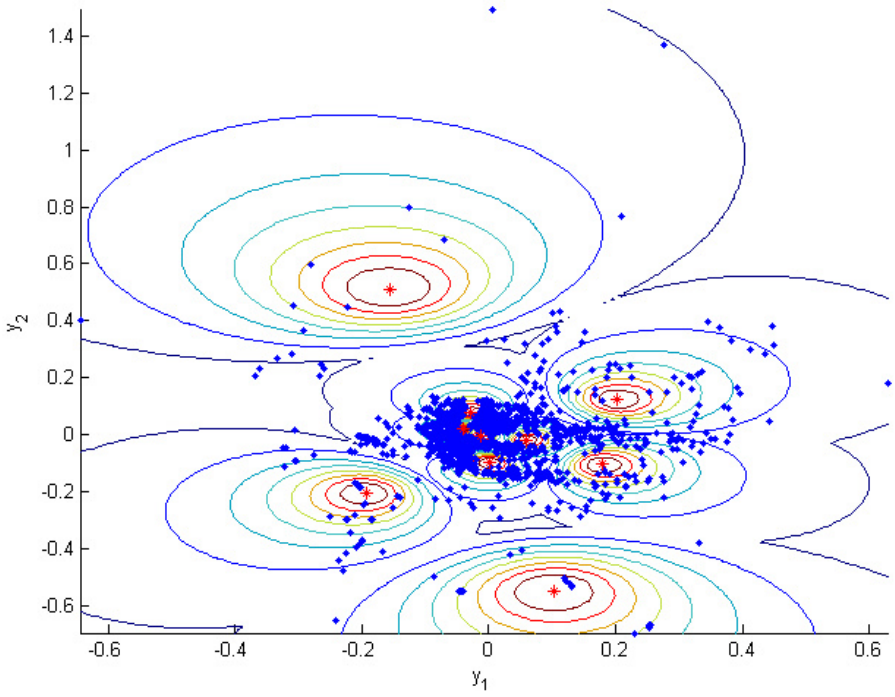


Figura 5.3.2 – Mapeamento FUZZSAM do Algoritmo FCM, para $c=10$ e $m=2$

- **Avaliação da eficiência do modelo classificador**

Após a clusterização dos dados, efetuou-se a classificação dos clientes quanto à presença de anormalidades no consumo de energia elétrica. O desempenho do modelo classificador foi medido com base na assertividade e na sensibilidade da configuração simulada (Tabelas Tabela 5.3.3 e Tabela 5.3.4)

Inicialmente, é importante destacar a existência de valores de assertividades iguais para várias combinações de *c* e *m*, como visto na Tabela 5.3.3. Por exemplo, o valor ASS08 igual a 1 para *c=9* e *m=2* ocorre porque, dos 8 clientes localizados acima do corte de 0,8, todos eram de fatos pré-rotulados como anormais, levando o sistema a 100% de assertividade (ver Tabela 5.3.5). Comportamentos semelhantes foram observados paras os demais valores de *m*, para *c=9.*

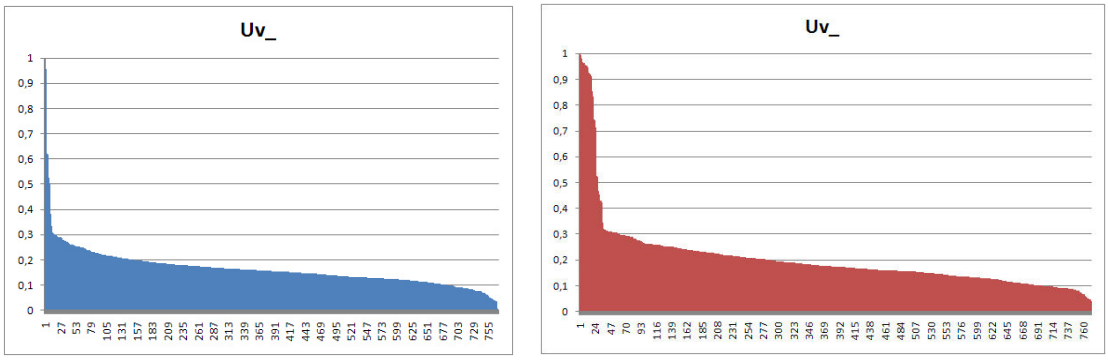
Tabela 5.3.5 – Resultados do sistema para clientes com $\mathbf{U_v} > 0.6$, para c=9, m=2.

Cliente	Classe	Uv
1	Anormal	1
2	Anormal	0.9722
3	Anormal	0.9713
4	Anormal	0.94912
5	Anormal	0.94882
6	Anormal	0.94635
7	Anormal	0.9304
8	Anormal	0.92249
9	Anormal	0.71701
10	Anormal	0.68925
11	Anormal	0.6864
12	Anormal	0.666
13	Anormal	0.66573

Já os valores nulos de assertividade para *c*=11 ocorreram porque acima do corte 0,8, nestes casos, não foram encontrados nenhum cliente pré-rotulado como anormal.

Em linhas gerais, na configuração ótima ($c=10, m=2$) obteve-se uma assertividade de 87,5% e uma sensibilidade de 2,72%. A assertividade se manteve em uma margem significativa, com a existência de somente 3 clientes normais acima do corte selecionado. Já a sensibilidade obtida foi baixa, dado que de um total de 772 irregulares na amostra, o sistema identificou somente 21 deles.

Na Figura 5.3.3 ilustram-se as curvas de classificação da configuração simulada. O objetivo das curvas é avaliar a eficiência do modelo classificador para cada classe distinta, para isso comparando o índice de fraude atribuído a cada cliente (contido no vetor Uv) com sua respectiva classe, pré-conhecida antes da simulação.



(a) **(b)**

Figura 5.3.3 – Curvas de classificação do modelo utilizando o FCM: **(a)** clientes normais; **(b)** clientes irregulares.

Na Figura 5.3.3a foram selecionados os 772 consumidores residenciais pré-rotulados como normais, sendo em seguida plotados no eixo vertical os respectivos índices de anormalidades gerados pelo sistema, ordenados de forma decrescente para uma melhor visualização. Considerando que o objetivo do sistema é atribuir índices os mais próximos possíveis de zero a clientes normais, então em um classificador ideal o aspecto do referido gráfico seria uma reta constante de valor zero. O mesmo vale para o caso dos clientes anormais (Figura 5.3.3b). Um classificador ideal produziria um gráfico com índices **Uv** constantes e iguais a um, para os demais 772 consumidores pré-rotulados como irregulares.

Na prática, o desempenho de um classificador não-discreto (que gera índices lineares) pode ser medido avaliando sua eficiência sob um dado nível de corte. Na presente simulação, acima do corte de 0,8 foram encontrados 21 clientes normais e 3 normais, indicando que o FCM apresentou um bom desempenho de classificação, obtendo assertividade acima da media de outras metodologias estudadas. Sob valores de corte muito baixos (0,2 por exemplo) o desempenho do modelo decai, não estabelecendo uma distinção apropriada entre as classes.

5.3.1.2 Algoritmo Gustafson-Kessel

- **Etapa de Clusterização**

O algoritmo GK foi simulado nos seguintes intervalos de parâmetros: $2 \leq c \leq 11$ e $1.1 \leq m \leq 2.0$. A quantidade ótima de clusters foi definida de forma semelhante ao algoritmo FCM, que resultou no valor $c = 9$. O mesmo processo foi usado para a definição do expoente fuzzificador ótimo *m*. Na Figura 5.3.4 são ilustradas as trajetórias de minimização dos índices SB, S e SC para diferentes valores de *m*, considerando *c* constante i igual a nove.

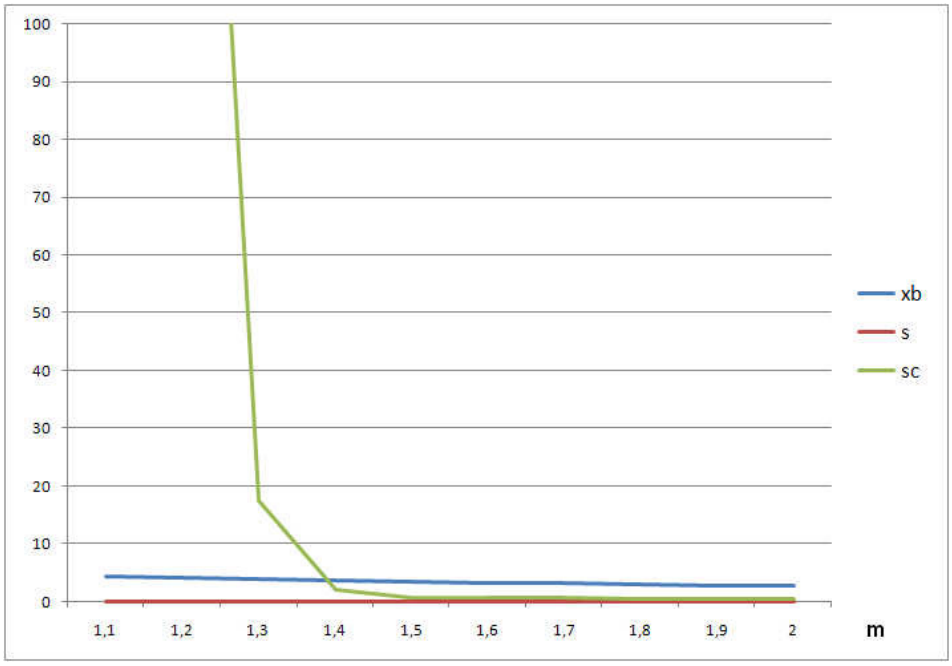


Figura 5.3.4 – Comparação do desempenho do algoritmo GK para diferentes valores de *m*, considerando $c = 9$.

Fundamentando-se na minimização das métricas -mencionadas e, principalmente, tendo em vista uma assertividade de valor satisfatório, o expoente fuzzificador ótimo escolhido foi 2,0.

- **Mapeamento FUZZSAM**

A projeção bidimensional dos dados de entrada, os centros de clusters e as curvas de pertinência do algoritmo Gustafson-Kessel na configuração ótima estão ilustrados na Figura 5.3.5. Observa-se no gráfico que o referido algoritmo produziu clusters mais compatíveis com a distribuição analisada, com características levemente elipsoidais – propriedade peculiar dos agrupamentos gerados pelo algoritmo GK. Percebe-se também que, mesmo diante do número elevado de clusters da configuração (total de 9), o algoritmo convergiu para uma razoável separabilidade entre os grupos, fato que também comprova a eficiência das métricas utilizadas.

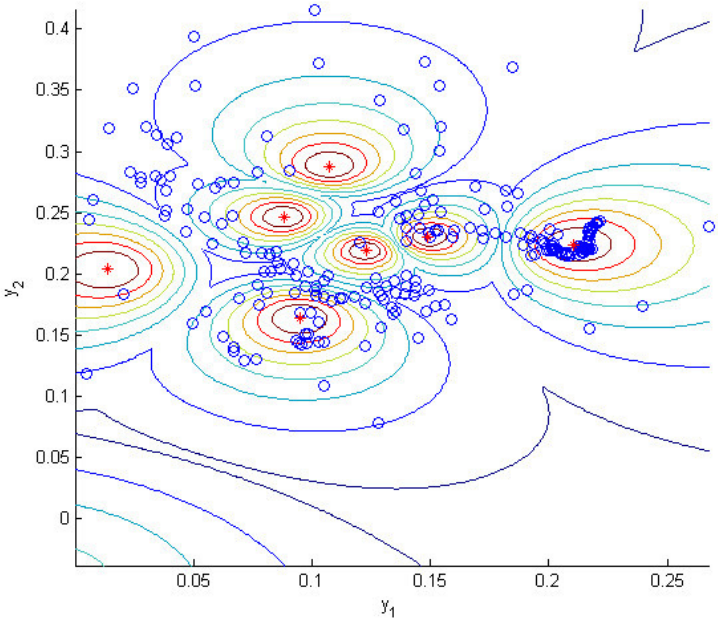


Figura 5.3.5 – Mapeamento FUZZSAM do Algoritmo GK, para $c=9$ e $m=2$

Já na Figura **5.3.6** mostra-se a projeção FUZZSAM para o algoritmo GK nos parâmetros $c=4, m=1.3$. Nota-se, à semelhança da configuração anterior, o grande potencial do algoritmo GK em se adaptar a diferentes configurações de dados, de forma a produzir clusters mais consistentes. É possível verificar também que os 4 clusters da configuração conseguiram cobrir adequadamente a distribuição de dados, mesmo não representando o número ótimo de clusters determinado. Outro ponto a ser destacado é o impacto do baixo valor do expoente fuzzificador ($m=1,3$) nos resultados, que tornou as fronteiras entre clusters menos suaves, como observado na figura.

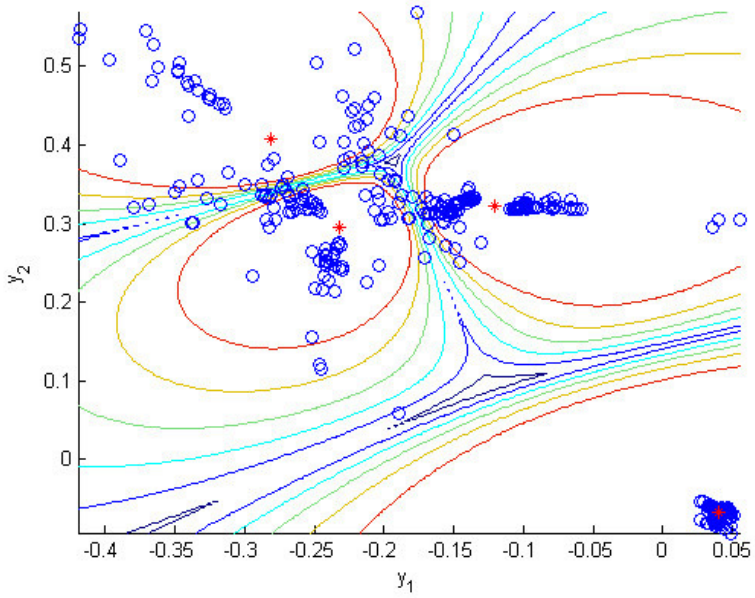


Figura 5.3.6 – **Mapeamento FUZZSAM do Algoritmo FCM, para** $c = 4$ e $m = 1{,}3$

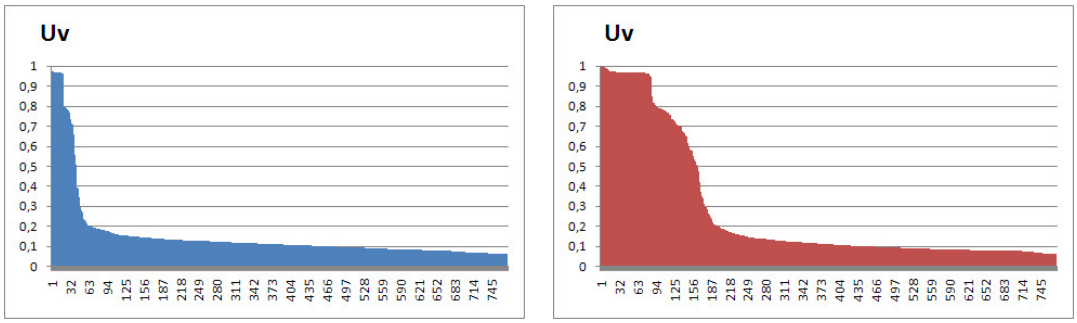
- **Avaliação da eficiência do modelo classificador**

Utilizando o algoritmo GK, nos parâmetros ótimos, o modelo classificador atingiu uma assertividade de 83,83%, isto é, de 99 clientes localizados acima do corte de 0.8, 83 deles eram de fato irregulares. Já a sensibilidade – que representa uma noção da abrangência do modelo classificador – foi de 10,75% neste caso, significando que de um total de 772 clientes fraudulentos analisados, 83 deles foram localizados pelo sistema. De fato, o que se requer de um modelo de detecção de fraudes não é que consiga identificar todos os elementos anormais existentes no espaço de busca – uma exigência impraticável – mas que tenha uma boa precisão de acerto dentro da margem que trabalhar (neste caso, um corte acima de 0,8).

Para a configuração $c = 4, m = 1.3$ atingiu-se uma assertividade de 86,66%, mas com uma sensibilidade (6,73%) reduzida a quase a metade do valor da configuração ótima. Mesmo assim, o desempenho do algoritmo nesta configuração foi ainda melhor do que o FCM, indicando que o método GK é mais robusto para a metodologia desenvolvida.

Na Figura 5.3.7 são mostradas as curvas de assertividade do modelo, ao ser utilizado o GK como algoritmo de clusterização. A assertividade superior do referido algoritmo pode também ser visualizada no gráfico, onde para um mesmo valor de corte (0,8 por exemplo)

mais clientes irregulares foram corretamente classificados (comparar Figuras Figura 5.3.7 e Figura 5.3.3).



(b) **(b)**

Figura 5.3.7 – Curvas de classificação do modelo utilizando o GK: **(a)** clientes normais; **(b)** clientes irregulares.

Por fim, na Tabela 5.3.6 é efetuada uma comparação geral entre os desempenhos do FCM e do GK, onde se incluem também informações de tempo computacional, como o número de iterações efetuadas e o tempo de clusterização. O algoritmo FCM convergiu em um tempo mais rápido (0.6 segundos de diferença) e 42 iterações a menos em relação ao GK. Entretanto, dada a dimensão considerável da mesma matriz de entrada usada (1544 clientes) e a assertividade do sistema obtida com o GK, este último ainda se mostrou mais apropriado à configuração simulada, atingindo uma sensibilidade 5 vezes superior ao FCM. A cobertura do GK para distribuições de dados dispersas, como a usada nesta simulação, se mostrou superior.

Tabela 5.3.6 – Comparação geral de desempenho entre o FCM e o GK, clientes residenciais.

Algoritmo	Parâmetros ótimos		Eficiência				Tempo de clusterização	Iterações efetuadas
	c	m	Ass09	Sens09	Ass08	Sens08		
FCM	10	2	0.9	0.02	0.875	0.02	1.281 s	234
GK	9	2	0.811	0.11	0.824	0.11	1.828 s	192

Uma análise adicional pode ser feita através da denormalização dos centros convergidos no processo de clusterização. Na Tabela 5.3.7 são mostrados os centros de

clusters obtidos após a execução do algoritmo GK, no horizonte histórico. É possível se verificar classes de consumo bem distintas, para cada centro obtido, como por exemplo:

- Cluster 1 – clientes residenciais de baixa-renda com consumo baixo, relativamente constante e com situação de fornecimento normal;
- Cluster 7 – consumidores residenciais com consumo máximo de 153.53 kWh e com várias notificações graves de apontamentos.

Tabela 5.3.7 – Centros convergidos após a clusterização, algoritmo GK.

Centro cluster	M6 (kWh)	MAX6 (kWh)	DEV6	M6SETOR (kWh)	APT6
1	30.65	33.796	15.435	66.121	1.247
2	153.85	239.25	44.223	78.327	0.741
3	82.841	124.22	25.867	63.629	10.119
4	78.811	157.14	51.125	68.432	0.807
5	95.228	122.63	19.775	85.992	0
6	79.578	92.097	97.222	82.061	0.835
7	128.65	153.53	20.333	157.46	14.603
8	109.1	130.96	18.494	85.997	0.394
9	62.686	90.349	23.442	73.902	14.719

5.3.2 SIMULAÇÃO NO DOMÍNIO DE CLIENTES COMERCIAIS

Nesta etapa, aplicou-se o sistema de detecção de anormalidades somente no universo de consumidores comerciais. Os parâmetros utlizados foram:

- Mês de referência para classificação (**t**): 01/07/2009;
- Intervalo retroativo (**k**): 18;
- Dimensões da matriz de entrada (**n** × **p**): 174 × 5;
- Tolerância de convergência para clusterização (ε): 0,01;

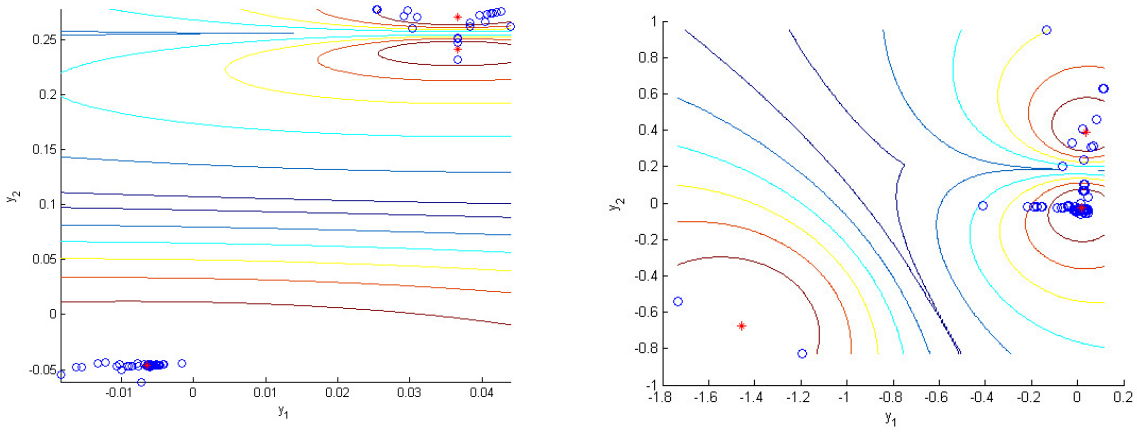
Os resultados para as melhores configurações de clusterização, para os algoritmos FCM e GK estão listados na Tabela 5.3.8.

Tabela 5.3.8 – Eficiência da metodologia para o universo de clientes comerciais

Algoritmo	Parâmetros ótimos		Eficiência				Tempo de clusterização	Iterações efetuadas
	c	m	Ass09	Sens09	Ass08	Sens08		
FCM	3	2	0.7143	0.0575	0.6667	0.0690	16ms	21
GK	3	2	0.6250	0.0575	0.6471	0.1264	94ms	28

Verifica-se na Tabela que a assertividade ao corte de 0,8 foi um pouco melhor no algoritmo FCM. Nota-se, entretanto, que a sensibilidade ao mesmo corte foi superior na simulação com o algoritmo GK. Isto comprova novamente uma maior robustez do algoritmo Gustafson-Kessel também no universo de clientes comerciais. Com relação ao tempo computacional, percebe-se novamente a vantagem do FCM em relação ao GK. Entretanto, dado o tempo de processamento reduzido, este último algoritmo ainda se torna viável, devido a sua sensibilidade superior.

Na Figura 5.3.8 mostram-se as projeções modificadas de Sammon dos algoritmos utilizados. É possível, inicialmente, notar claramente a redução do universo de clientes analisados, em relação à simulação com clientes comerciais. Nota-se também que o algoritmo GK gerou clusters mais adaptados à distribuição de dados, onde se percebe a existência de três grandes grupos de clientes comerciais.



(a) **(b)**

Figura 5.3.8- Mapeamentos FUZZSAM, para $c=3$ e $m=2$: consumidores comerciais: **(a)** algoritmo GK; **(b)** algoritmo FCM.

Por fim, a análise dos centros convergidos na etapa de clusterização (como mostrados na Tabela 5.3.9) é útil para o estabelecimento de grupos de consumidores comerciais analisados, como por exemplo:

- Cluster 1 – clientes comerciais de consumo reduzido, mas com consideráveis quantidades de apontamentos graves;
- Cluster 2 – consumidores comerciais com poucos apontamentos graves e picos médios de consumo de 409.47 kWh.
- Cluster 3 – clientes comerciais com altíssimos consumos com desvios padrões consideráveis.

Tabela 5.3.9 – Centros convergidos após clusterização pelo algoritmo GK

Centro	M6	MAX6	DEV6	M6SETOR	APT6
1	234.5	295.05	43.55	208.38	12.428
2	326.48	409.47	56.71	345.08	0.024
3	17579	24784	7504,9	30815	0.006

5.3.3 SIMULAÇÃO NO DOMÍNIO DE CLIENTES COMERCIAIS E RESIDENCIAIS

Simulou-se também o sistema em um universo contendo tanto clientes residenciais quanto comerciais. As métricas *assertividade* e *sensibilidade* aos cortes 0,8 e 0,9, para diferentes configurações de *c* e *m*, bem como os tempos de clusterização e iterações efetuadas em cada algoritmo estão mostradas na Tabela 5.3.10.

Tabela 5.3.10 – Eficiência do modelo classificador: clientes residenciais e comerciais

Algoritmo	Parâmetros		Eficiência				Tempo de clusterização	Iterações efetuadas
	c	m	Ass09	Sens09	Ass08	Sens08		
GK	3	1.1	0.7674	0.0723	0.7468	0.0861	0.25 s	36
FCM	3	2	0.9167	0.0080	0.9524	0.0146	0.11 s	24
FCM	7	2	0.7100	0.0518	0.7130	0.0562	0.40 s	48

Verifica-se novamente uma melhor correlação entre as métricas ASS08 e SENS08 por parte do algoritmo GK, que atingiu uma assertividade de 74,68% em uma sensibilidade de 8,61% (quase o dobro da sensibilidade do FCM na melhor configuração). O algoritmo FCM apresenta um bom desempenho na assertividade, mas sempre em detrimento da sensibilidade, o que favorece o algoritmo GK, mesmo com superior tempo computacional e iterações.

Os clusters gerados pelo algoritmo Gustafson-Kessel, plotados no espaço bi-dimensional segundo a projeção FUZZSAM estão mostrados na Figura 5.3.9. Percebe-se uma adequada separabilidade entre os clusters, mesmo diante de uma relativa dispersão da distribuição de dados. Nesse caso, como a análise considera tanto clientes residenciais como comerciais, é injustificada a associação de um dado cluster a uma classe comercial com comportamento definido.

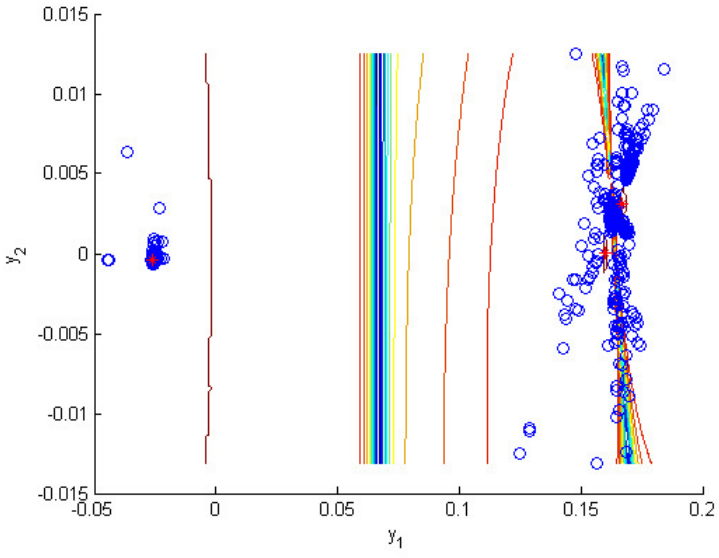


Figura 5.3.9- Mapeamento FUZZSAM do Algoritmo GK, para $c = 3$ e m = 1,1: consumidores residenciais e comerciais.

5.3.4 INFLUÊNCIA DA DISTRIBUIÇÃO DE CLASSES NA ASSERTIVIDADE

Nesta análise, busca-se observar a influência da proporção de clientes anormais existentes na base de testes sobre a eficácia do sistema. Os resultados obtidos para clientes residenciais, utilizando o algoritmo GK com os mesmos parâmetros da Seção 5.3.1 ($c=9$ e $m=2$) estão registrados na Tabela 5.3.11.

Tabela 5.3.11 – Desempenho do sistema para diferentes distribuições de classes (GK)

Métrica	Porcentagem de Casos Anormais				
	90%	75%	50%	25%	10%
ASS08	0.9700	0.8043	0.7879	0.7045	0.3529
SENS08	0.2056	0.0881	0.1684	0.1206	0.0706
ASS09	0.9813	0.7899	0.8039	0.6757	0.2857
SENS09	0.0833	0.0746	0.1062	0.0973	0.0471

Nota-se que, quanto maior a proporção de casos irregulares na amostra, ou, em outras palavras, quanto mais irregular for uma região, maiores serão as taxas de detecção de anormalidades – fato também observado na prática. A proporção de casos anormais na base de testes pode, por outro lado, tornar tendenciosa a análise de resultados do sistema. Justamente por isso, recomenda-se a utilização de proporções iguais das amostras, de forma a obter-se uma noção mais geral sobre a eficácia do sistema.

5.3.5 ANÁLISE DE DIFERENTES NÍVEIS DE CORTE

Esta análise teve por objetivo avaliar o desempenho do modelo em diferentes níveis de corte, tomando-se também como base a configuração do algoritmo GK simulada na Seção 5.3.1. Pelos resultados obtidos, mostrados na Tabela 5.3.12, nota-se, primeiramente, que a sensibilidade mostrou-se inversamente proporcional ao nível de corte, o que é coerente, pois quanto mais baixo o corte, mais casos anormais serão contabilizados.

Tabela 5.3.12 – Desempenho do sistema para diferentes níveis de corte (GK)

Corte	Métrica	
	Assertividade	Sensibilidade
0.9	0.803	0.106
0.8	0.787	0.168
0.7	0.791	0.205
0.6	0.783	0.215
0.5	0.773	0.221
0.4	0.773	0.221

Pode-se observar também que é possível se determinar o nível de corte ideal dependendo do objetivo das metas para o processo de detecção de anormalidades. A depender do nível de corte selecionado, é possível alcançar uma assertividade de até 80,3% (corte 0,9) ou uma sensibilidade de até 22,1% (corte 0.4). Tal escolha é uma relação custo-benefício entre duas alternativas: descobrir muito e acertar pouco ou descobrir pouco e acertar muito.

5.3.6 INFLUÊNCIA DA SAZONALIDADE

Nesta última simulação, procurou-se analisar a robustez do modelo quando executado em horizontes de tempo distintos e seqüenciais — procedimento igualmente realizado por centros reais de identificação de irregularidades. Utilizando uma distribuição de 50% de consumidores residenciais, simulou-se o sistema em 13 meses de referência, sendo os resultados mostrados na Figura 5.3.10.

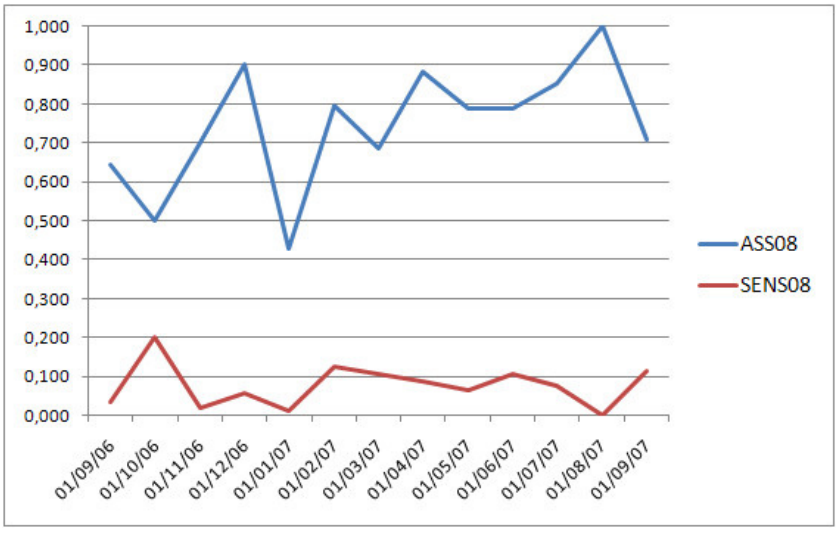


Figura 5.3.10 – Trajetória da assertividade e da sensibilidade do modelo em 13 meses.

Observa-se que o sistema conseguiu manter o desempenho esperado nos treze meses simulados, atingindo uma assertividade média de 0,745 e uma sensibilidade média de 0,07. O mês de assertividade mais alta (01/08/07) foi também aquele de menor sensibilidade, e o mês de uma das menores assertividades (01/10/06) foi aquele de maior sensibilidade (cerca de 20%), indicando uma relação inversamente proporcional entre estas duas grandezas, como era esperado. Uma das explicações para a existência de diferentes taxas de assertividade ao longo dos meses é a mudança da distribuição de dados, que requer o ajuste ocasional dos parâmetros do algoritmo e do nível de corte mais adequado.

6 CONCLUSÃO

Este trabalho apresentou uma metodologia para classificação de anormalidades em perfis de consumo de energia elétrica. O modelo se baseia na técnica de clusterização, amplamente utilizada em várias outras metodologias de identificação de fraudes. A proposta fundamenta-se também em um processo de classificação fuzzy, dado que o problema abordado envolve a estimativa da pertinência em grupos.

Foi realizada uma ampla pesquisa sobre a problemática de fraudes e anormalidades no consumo de energia elétrica, de forma a levantar as variáveis de maior correlação com o fenômeno. Efetuou-se também uma revisão das principais metodologias correlatas, analisando seus pontos fortes e desvantagens, bem como as principais tendências do assunto.

A presente metodologia foi desenvolvida sob o âmbito da Descoberta de Conhecimento em Bases de Dados (KDD), ramo computacional que congrega as técnicas para manipulação, extração e interpretação de informações em bancos de dados. Para garantir a validade científica do modelo desenvolvido foram embutidas no sistema tarefas próprias de KDD, como pré-processamento, normalização de dados, além da estratégia de mineração de dados desenvolvida.

O sistema foi simulado com dados reais, apresentando um desempenho satisfatório na identificação de irregularidades em clientes residenciais e comerciais, sendo avaliado com os algoritmos de clusterização *Fuzzy C-Means* e *Gustafson-Kessel*, em várias distribuições da base de testes, em vários níveis de corte e em distintos horizontes de análise. A qualidade dos agrupamentos formados foi validada com índices de desempenho de clusterização, bem como com uma técnica de visualização gráfica de dados. A eficiência do modelo classificador foi testada utilizando métricas de performance padrões de sistemas de detecção de fraudes.

O objetivo do trabalho foi atingido, com a constatação da viabilidade e aplicabilidade da metodologia, principalmente ao extenso domínio dos clientes residenciais. O fato do sistema utilizar como variáveis de entrada um conjunto reduzido de informações que podem ser encontradas na maioria das concessionárias, aliado ao seu caráter não-supervisionado,

que dispensa a utilização de regras, torna factível sua aplicação em qualquer área de concessão.

Em suma, algumas das contribuições deixadas por este trabalho são:

- Uma proposta de classificação de perfis de consumo de energia elétrica a partir de dados de consumo, apropriada a identificação de fraudes e anormalidades em qualquer classe de consumidor;
- Um sistema particularmente voltado à identificação de anormalidades em consumidores residenciais – segmento incipientemente coberto nas atuais metodologias – dado ao uso de um reduzido conjunto de atributos que podem ser encontrados facilmente na maioria das concessionárias de energia;
- Como um sistema baseado em clusterização, fornece uma alternativa para segmentação natural de perfis de consumo de energia elétrica;
- A curto-médio prazo, possibilita a redução do nível de perdas comerciais, com o aumento da eficácia das inspeções de campo a partir da indicação dos clientes com maiores chances da existência de anormalidades;
- A longo prazo, possibilita a melhoria da qualidade da energia fornecida, redução de investimentos em manutenção e expansão e redução da tarifa de energia.

Como horizontes para futuros trabalhos propõem-se:

- Implantação de informações relativas à previsão de retorno financeiro de cada cliente;
- Implantação de um algoritmo para escolha automática dos parâmetros ótimos do modelo de clusterização e do nível de corte mais adequado à distribuição de dados;
- Teste do sistema em consumidores industriais;
- Utilização de outros algoritmos de clusterização fuzzy, como o *C-Means* possibilístico (PCM);
- Validação da assertividade do modelo em testes de campo.

Faz-se necessário, por fim, ressaltar a necessidade de uma maior integração entre os centros de pesquisa e combate à fraudes. Não obstante já não serem poucos os trabalhos relacionados à questão, é necessário ainda um maior direcionamento a temas importantes do processo, como a viabilidade econômica de inspeções, dentre outros.

REFERÊNCIAS BIBLIOGRÁFICAS

ABONYI, J & BABUSKA, R. **FUZZSAM – Visualization of fuzzy clustering results by modified Sammon mapping.** In JM Zurada (Ed), *FUZZ-IEEE 2004: IEEE International Conference on fuzzy systems* (pp. 1-6). Piscataway: IEEE 2004.

ARAÚJO, Antônio Carlos M. de e SIQUEIRA, Cláudia Aguiar de. **Considerações sobre as perdas na Distribuição de Energia Elétrica no Brasil.** In: XVII Seminário Nacional de Distribuição de Energia Elétrica, 2006, Belo Horizonte. Disponível em http://www.brasilengenharia.com.br/580/Art_Eletrica_Considera__esp112.pdf . Acesso em: 19 dez. 2006.

ANEEL. **Nota Técnica nº 026/2006,** Tratamento regulatório das perdas de energia nas tarifas dos sistemas de distribuição de energia elétrica, SRD/SRC/SRE/ANEEL, 2006.

ARAÚJO, Antônio Carlos M. de. **Perdas e inadimplência na atividade de distribuição de energia elétrica no Brasil**. 2007. 98 f. Tese de doutorado – Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, 2007.

BRASIL. **Relatório sobre Auditoria Operacional Realizada na Agência Nacional de Energia Elétrica (ANEEL)**, TRIBUNAL DE CONTAS DA UNIÃO – TCU, Brasília. 2007.

BOLTON, R. J. and HAND, D. J. **Statistical fraud detection: A review**, Statistical Science, vol. 17, no. 3, pp. 235–249, 2002.

CABRAL, J. E; GONTIJO, E. M; PINTO, J. O. P. E FILHO, J. R. **Fraud detection in electrical energy consumers using rough sets**, *in Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*, pp. 3625–3629. 2005.

CALILI, Rodrigo Flora. **Desenvolvimento de sistema para detecção de perdas comerciais em redes de distribuição de energia elétrica.** Rio de Janeiro, 2005. 157p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

COMETTI, E. S. e VAREJÃO, F. M.. **Melhoramento da identificação de perdas comerciais através da análise computacional inteligente do perfil de consumo e dos dados cadastrais de consumidores**. Relatório final de projeto de P&D, ciclos 2003/2004. Escelsa/Aneel, Vitória/ES, 2005.

DOS-ANGELOS, E. W. S.; SAAVEDRA, O. R. e CORTES, O. A. C, **Sistema Inteligente para Identificação de Fraudes em Redes de Energia Elétrica baseado em Lógica *Fuzzy***. In: VIII SBAI – Simpósio Brasileiro de Automação Inteligente. Florianópolis. 2007.

DOS-ANGELOS, E. W. S.; SAAVEDRA, O. R. e CORTES, O. A. C, **Sistema para Classificação de Anormalidades no Consumo de Energia Elétrica**. In: IX SBAI – Simpósio Brasileiro de Automação Inteligente. Brasília. 2009.

ELETROPAULO. 2009. Relatório de Sustentabilidade 2008. Perdas de Energia Elétrica. Disponível em http://www.grupoaesbrasil.com.br/ Acesso em 22/06/08.

ELLER, N. A. **Arquitetura de Informação para o Gerenciamento de Perdas Comerciais de Energia Elétrica**, Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, Florianópolis/SC, 2003.

FAYYAD, U., PIATETSKY-SHAPIRO, G. e SMYTH, P. **From data mining to knowledge discovery in databases**, *AI Magazine* 17: 37– 54. 1996

FERREIRA, G. G. L. **Desenvolvimento de um sistema baseado em regras para detecção de fraude em unidades consumidoras ligadas em baixa tensão**, Trabalho de graduação, Universidade Federal de Santa Maria – RS, Santa Maria – RS, 2007.

FILHO, J. R. e GONTIJO, Edgar M. *et al.*: **Fraud identification in electricity company customers using decision tree**. Proceedings of the IEEE International Conference on Systems, Man Cybernetics: The Hague, Netherlands, October 2004.3730-3734.

GOLDSCHMIDT, R. e PASSOS, E. **Data Mining: Um guia prático,** Elservier – Campus, Rio de Janeiro, 2005.

GUIMARÃES, Luiz Carlos Silveira – ABRADEE. **Visão da Associação Brasileira de Distribuidores de Energia Elétrica.** In: I Workshop Furto/Fraude de Energia e Roubo de Condutores e Equipamentos. 2004, Curitiba. **Painel 1: Furto / Fraude de Energia.** Disponível em http://www.abradee.org.br/Downloads/1_Worshop_Furtos/abradee.pdf . Acesso em 12 dez. 2006.

GUSTAFSON, E. E. and KESSEL, W. C. '**Fuzzy Clustering with a fuzzy Covariance matrix'** *Proc. of the IEEE Conference on Decision and Control, San Diego, Californien,* pp. 761-766. IEEE Press, Piscataway, NJ, 1979.

HAN, J. e KAMBER, M. **Data Mining, Concepts and Techniques,** 2nd ed, Morgan Kaufmann, 2006.

HAND, D. J. **Discrimination and Classification**. Wiley, Chichester, 1981.

HWANG, S.; THILL, J-C, **Using Fuzzy Clustering Methods For Delineating Urban Housing Submarkets**, In: CMGIS'07, November 7-9, 2007, Seattle, 2007.

INTERNATIONAL ENERGY AGENCY – IEA. Database: Energy Statistics for Electricity / Heat. 2007. Disponível em: http://www.iea.org/Textbase/stats/prodresult.asp?PRODUCT=Electricity/Heat . Acesso em: 15 de Mar. 2007.

JIANG, R., TAGARIS, H., LACHSZ, A. E JEFFREY, M. **Wavelet based feature extraction and multiple classifiers for electricity fraud detection**, IEEE Transmission and Distribution Conference and Exhibition **3**: p.2251– 2256, 2002.

LAROSE, Daniel T. **Discovering Knowledge in data**. Wiley Interscience, 2005.

NETO, Edison; ARANHA, A.C.. **Combate às Perdas Não-Técnicas no Brasil**. Apresentação Power Point. CLADE, 2008.

KOU, Y., LU, C.-T., SIRIRAT, S. and HUANG, Y.-P. **Survey of fraud detection techniques**, IEEE International Conference on Networking, Sensing and Control, pp. 21–23. 2004.

LAROSE, Daniel T. **Discovering Knowledge in data: An Introduction to Data Mining**, Wiley-Interscience, New Jersey, 2005.

LAZO, J. G.; VELLASCO, M. M. R.; PACHECO M. A. C; LEITE, K. T. F.; BARBOSA, C. R. H.; TANSCHEIT, R. e ROCHA, J. E. N. (2005). **Identificação e prevenção de perdas comerciais no faturamento**, Anais do III CITENEL, In: III Congresso de Inovação Tecnológica em Energia Elétrica - CITENEL, Florianópolis – SC.

MCLACHLAN, G. J. **Discriminant Analysis and Statistical Pattern Recognition**. Wiley, New York. 1992.

MICROSOFT VISUAL C++ 2005 Express Edition, *Microsoft Corporation*, 2005, Disponível em: http://msdn.microsoft.com/en-us/vstudio/bb984878.aspx. Acesso em 27 Ago 2009.

OCCI, *Oracle C++ Call Interface*, v.10.2.0.3.0 *Patch 13*, Oracle Corporation, 2007. Disponível http://www.oracle.com/technology/tech/oci/occi/occidownloads.html. Acesso em 27 Ago 2009.

OLIVEIRA, J. V. de; PEDRYCZ, Witold. **Advances in Fuzzy Clustering and Its Applications,** John Willey & Sons Ltda, London, 2007.

ORACLE Database 10G Express Edition, v.10g, Release 2 (10.2.0.3), Oracle Corporation, 2007, Disponível em http://www.oracle.com/technology/products/database/xe/index.html. Acesso em 27 Ago 2009.

PATHAK, J.; VIDYARTHI, N. and SUMMERS, S. L.. **A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims**, *Managerial Auditing Journal* 20(6): 632–644, 2005.

PHUONG, N. H.; SANTIPRABHOB, P.; THANG, C. and MASAYUKI, A. **A fuzzy consultation system for computer configurations**, *International Conference InTech/VJFuzzy*, pp. 137–142, 2002.

THANG, C., TOAN, P. Q., COOPER, E. W. and Kamei, K. **Application of soft computing to tax fraud detection in small businesses**, *IEEE International Conference on Communications and Electronics, pp. 402–407*, 2006.

PEDRYCZ, Witold. **Knowledge-Based Clustering: From Data to Information Granules,** Wiley-Interscience, New Jersey, 2005

PENIN, Carlos Alexandre de Souza. **Combate, prevenção e otimização das Perdas Comerciais de Energia Elétrica**. 2008. 194 p. Tese de doutorado – Universidade de São Paulo, Escola Politécnica, São Paulo, 2008.

REIS, Claudia Zuccolotto. Lustosa, Leonardo Junqueira. **Eficácia de solução tecnológica para redução de furtos de energia elétrica em empresas distribuidoras: Estudo de caso.** Rio de Janeiro, 2005. 81p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro. 2005.

ROCHA, José Eduardo Nunes da. **Sistemas inteligentes no estudo de pedras comerciais do setor de energia elétrica**. Rio de Janeiro, 2003. 198f. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

SACIC, Bernardo – AMPLA. **Perdas Não-Técnicas.** Audiência 004/2007 ANEEL. Apresentação Power Point. Disponível em: http://www.aneel.gov.br/aplicacoes/audiencia/arquivo/2007/004/apresentacao/bernardo_sacic_-_ampla.pdf Acesso em 15 de Agosto de 2009. 2007.

SUNDARAM, A. **An Introduction to Intrusion Detection**. ACM Crossroads, Vol. 2, No. 4, 1996.

VIEIRALVES, Eduardo de Xerez. **Proposta de uma Metodologia para Avaliação das Perdas Comerciais dos Sistemas Elétricos.** O Caso Manaus. 136p. Dissertação de Mestrado. Campinas: Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas. 2005.

WITTEN, Ian. H; FRANK, Eibe. **Data mining: Practical Machine Learning Tools and Techniques,** Second Edition, Morgan Kaufmann Publishers, 2005.

YAP, Keem Siah *et al.*, **Abnormalities and fraud electric meter detection using hybrid support vector machine & genetic algorithm**, Proceedings of the third conference on IASTED International Conference: Advances in Computer Science and Technology, p.388-392, Phuket, Thailand, April 02-04, 2007.