

Mineração de Dados aplicada à Detecção de fraudes nas redes de distribuição de energia elétrica

Diogo Viana Rocha (UTFPR) diogovianarocha@gmail.com
Gloria Patricia López Sepúlveda (UTFPR) gloriap@utfpr.edu.br
Hugo Andres Ruiz Florez (UTFPR) hugoflorez@utfpr.edu.br

Resumo:

Neste artigo é utilizada a inteligência computacional mediante o uso de um modelo de mineração de dados baseado no processo de descoberta de conhecimento em base de dados, também conhecido como KDD. Este modelo tem como objetivo auxiliar na resolução de um grande problema enfrentado pelas empresas concessionárias de energia elétrica em todo Brasil, como é o caso da identificação de usuários fraudulentos nas redes de distribuição de energia elétrica. Neste trabalho, o objetivo é realizar a identificação de usuários que fazem algum tipo de fraude nas redes de distribuição de energia elétrica, visando reduzir as perdas de energias oriundas de fraudes e diminuir a quantia dos gastos com inspeções em campo. No desenvolvimento desta pesquisa, foram estudadas várias técnicas e algoritmos de mineração de dados, dentre eles foram escolhidos os algoritmos de máquinas de vetor de suporte e árvores de decisão, os quais foram gerenciados através do software de mineração de dados WEKA, este software permite realizar o treinamento e validação dos algoritmos de mineração de dados para serem utilizados no processo de identificação dos usuários que devem ser inspecionados. Neste artigo são descritas as técnicas de mineração de dados que foram usadas e, é descrito o estudo de caso de uma empresa concessionária de energia elétrica, onde foi praticada a seleção de usuários a serem inspecionados por suspeita de fraude segundo a indicação dos modelos de mineração de dados utilizados durante a pesquisa. Os resultados obtidos demonstraram ser satisfatórios quando avaliados com as inspeções realizadas em campo.

Palavras chave: Árvores de Decisão, Inteligência Computacional, Fraudes, Máquina de Vetor de Suporte, Mineração de Dados.

Applied Data Mining to fraud detection in electricity energy distribution networks

Abstract

In This Article is utilized computational intelligence through the use of a data mining model based on knowledge discovery in database, also known as KDD. This model aims to assist in resolving a huge problem faced by electric companies through Brazil, such as the case of identifying fraudulent users in electricity energy distribution networks. In this work, the objective is to identify users who do some type of fraud in the electricity energy distribution networks, helping to reduce electricity losses arising from fraud and reduce the amount of expenses with inspection in the field. During the development of this research, were studied various techniques and algorithms of data mining, between them were used the support vector machine and decision trees algorithms, which were guided through WEKA data mining free software, this program enabled the training and validation of data mining algorithms be use in the process of identifying user to be inspected. In this article are described the data mining techniques and the case study of an electricity energy company, where was practiced the selection of user to be inspected on suspicion of fraud according to the indication of the data mining models used during the search. The obtained results demonstrated be satisfactory when data mining is evaluated with inspections in the field.

Key-words: Decision Tree, Computational Intelligence, Fraud, Support Vector Machine, Data Mining.

1. Introdução

As perdas de energia elétrica são encaradas como um problema mundial que prejudica a sociedade, pois além do prejuízo financeiro para as empresas concessionárias de energia elétrica as ligações irregulares provocam aumentos nas tarifas que são cobradas aos clientes das diferentes empresas distribuidoras. Além das perdas econômicas, as realizações de fraudes nas redes de distribuição de energia elétrica (RDEE) representam um risco para a segurança pública, dado que podem causar acidentes graves como descargas elétricas e incêndios que podem levar a morte de pessoas (DELGADO, 2010). Isto ocorre devido à alteração das características da rede, portanto as fraudes são um problema que vão além das perdas financeiras de uma empresa, o que faz com que este problema seja de grande interesse não só na área acadêmica.

As fraudes feitas pelos usuários das distribuidoras de energia elétrica têm crescido de forma gradativa nos últimos anos, uma das causas deste crescimento é a crise econômica atual que está sendo vivenciadas pelos brasileiros. Esta crise tem feito com que consumidores que tem dificuldades financeiras optem por realizarem fraudes nas RDEE com o objetivo de economizar dinheiro. O problema de fraudes nas redes de distribuição de energia elétrica tem se mostrado difícil de solucionar, e resulta em grandes perdas de receita para as empresas concessionárias distribuidoras de energia elétrica (DELAIBA ET AL, 2004).

No processo de distribuição de energia elétrica existem dois tipos de perdas: as perdas técnicas e as perdas comerciais. As perdas técnicas ocorrem naturalmente no sistema de distribuição de energia elétrica, e estão relacionados às próprias características do sistema. Enquanto as perdas comerciais ocorrem devido a diversos tipos de fraudes, ligações irregulares, furtos, falhas nas medições e erros de leituras (COPEL, 2014). Na figura 1 são apresentados os diferentes tipos de perdas.

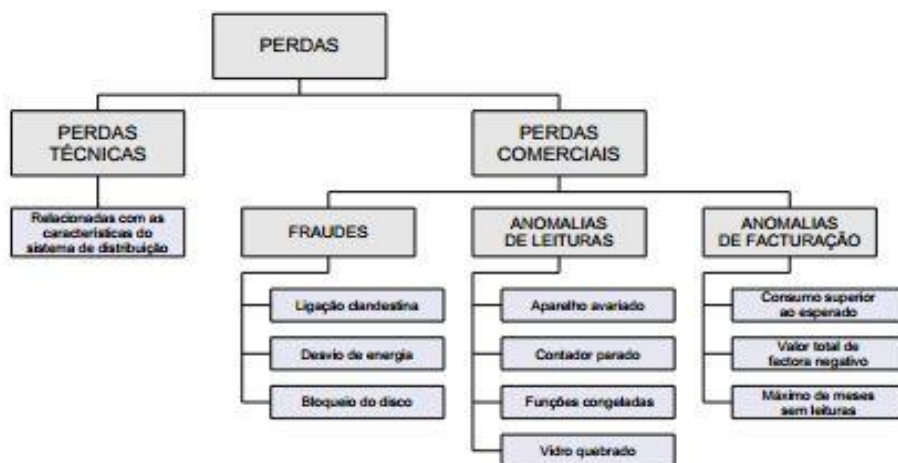


Figura 1 – Tipos de Perdas
Fonte: (Franco; López; Queiroz, 2015)

Algumas das estratégias utilizadas para a detecção de ligações clandestinas são: realização de inspeções nas unidades consumidoras, fiscalização de denúncias feitas por consumidores, realização de estatísticas de consumo de forma manual baseada no consumo dos usuários. Porém essas estratégias acarreta em mais gastos para as empresas concessionárias distribuidoras de energia elétrica (ARAUJO, 2006).

Outra estratégia empregada pelas empresas distribuidoras de energia elétrica tem sido desenvolver diferentes equipamentos que dificultam a violação do medidor de energia elétrica. Além dos equipamentos sofisticados utilizados pelas empresas, pode-se constatar que atualmente diversas empresas estão investindo em pesquisas e no desenvolvimento de novos métodos que auxiliem na detecção de fraudes através do uso de sistemas inteligentes. Estes novos métodos são baseados na inteligência computacional, e permite a análise de grandes bases de dados dos consumidores, nas quais se tem informação de consumo, características dos clientes, características da área em que reside, entre outras. (FRANCO; LÓPEZ; QUEIROZ, 2015).

Considerando o crescimento da população e, por conseguinte dos clientes das empresas concessionárias distribuidoras de energia elétrica, deve ser considerado que a base de dados a ser analisada é de grande porte, portanto foi necessário para o desenvolvimento deste artigo, escolher uma linha de pesquisa da inteligência computacional que permitisse a manipulação de grandes quantidades de dados de forma eficiente e rápida.

Com o intuito de realizar a toma de decisões baseadas em análises obtidos a partir do uso de ferramentas computacionais foram escolhidos algoritmos de mineração de dados (MD), estes algoritmos permitem analisar grandes bases de dados levando a detecção de suspeitos de fraudes nas redes de distribuição de energia elétrica.

A MD pode ser aplicada a grandes quantidades de dados através do uso de diversos algoritmos, dentre os quais destacam-se: árvores de decisão (AD's), redes neurais, máquinas de vetor de suporte (MVS), árvores de regressão, entre outros. Durante o desenvolvimento desta pesquisa, foram utilizados os algoritmos de MVS e de AD's.

Para o gerencialmente destes algoritmos foi utilizado o software WEKA, versão 3.8 na análise dos dados provenientes de um arquivo de texto no formato CSV, que foi gerado após o pré-processamento dos dados usando um software desenvolvido na linguagem de programação JAVA e alguns filtros da ferramenta WEKA.

Os resultados gerados pela mineração de dados auxiliam na tomada de decisões, uma vez que extrai informações a partir de grandes volumes de dados e permite a realização de uma análise estatística dos mesmos, possibilitando determinar o perfil de cada consumidor em relação aos diferentes tipos de anomalias, ou melhor, classificar os clientes conforme sua confiabilidade.

Os dados utilizados para realizar os testes e as avaliações dos algoritmos propostos, são dados reais que foram entregues por uma empresa concessionária de energia elétrica, este banco de dados contém informação de consumo de cada cliente, identificação dos clientes, entre outras características relevantes para aplicar a MD.

2. Metodologia

A metodologia utilizada no desenvolvimento deste trabalho é baseada em uma técnica de inteligência computacional, conhecida como MD. Neste contexto, são utilizados sistemas computacionais inteligentes, estes sistemas podem variar de um simples sistema de controle, até refinados *Data Warehouses*, *Data Marts*, ou outras formas de processamento analítico de descoberta de informações (FRANCO; LÓPEZ; QUEIROZ, 2015). A principal característica que cada um desses sistemas possui em comum é que todos eles buscam apresentar informações relevantes e precisas de forma rápida. É necessário saber que a escolha e configuração cada um desses sistemas devem ser realizadas de acordo com os objetivos de quem as utiliza, pois devem assimilar o conhecimento da área de negócio, neste contexto é utilizado o processo de descoberta de conhecimento em base de dados, visando realizar a toma de decisões de forma rápida e eficiente.

2.1 Descoberta de Conhecimento em Bases de Dados

O KDD é um processo interativo e iterativo, que envolve vários passos com diversas decisões a serem tomadas pelo usuário [FAYYAD et al, 1996 a]. Esses passos podem ser ilustrados na Figura 2.

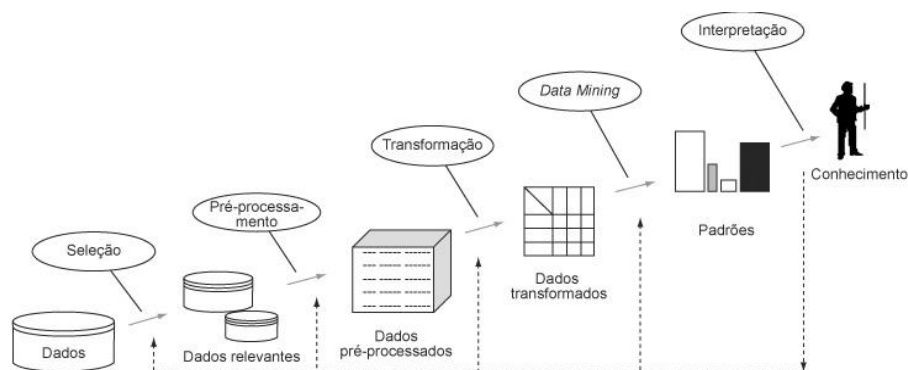


Figura 2 - Processo de KDD
Fonte: FAYYAD et al (1996 a)

De acordo com BRACHMAN e ANAND (1996), os passos básicos para uma boa execução do processo de KDD podem ser descritos da seguinte maneira:

1. Identificar o objetivo do processo de KDD seguindo o ponto de vista do cliente;
2. Criar uma base de dados, fazer um agrupamento dos dados relevantes que sejam relacionados ao objetivo do KDD;
3. Limpar os dados e realizar o pré-processamento: essa é a etapa onde são feitas operações básicas como a remoção do ruído e execução de uma estratégia que lide com campos vazios. Essa etapa tem o objetivo de assegurar a qualidade dos dados selecionados;
4. Transformar os dados: Os dados pré-processados devem passar por outra transformação, que os armazenem em outra forma, para facilitar a execução das técnicas de MD;
5. Escolher o método para MD: é necessário escolher um ou mais métodos que combine com o objetivo do KDD, alguns dos possíveis métodos são: clustering, regressão, classificação e sumarização;
6. Escolher o algoritmo para MD: é a etapa onde se escolhe o algoritmo a ser utilizado na busca de padrões na base de dados. Essa etapa também inclui decidir quais modelos e parâmetros são adequados, e garantir que o formato dos dados seja compatível com o algoritmo, como que o algoritmo seja compatível com o objetivo do KDD;
7. Minerar os dados: aplicar os algoritmos de MD na base de dados em busca de padrões de interesse em uma forma representacional, ou criar representações em formas de regras ou árvores de classificação, regressão, clustering e assim por diante. O usuário pode auxiliar significativamente na eficiência do algoritmo de MD, executando corretamente as etapas anteriores;
8. Interpretar os padrões minerados, e possivelmente retornar a qualquer um dos passos 1-7 para mais uma iteração. Este passo também pode envolver a visualização dos padrões extraídos, ou a visualização dos dados extraídos através do modelo;
9. Consolidar o conhecimento descoberto, incorporando esse conhecimento a outro

sistema, ou simplesmente documentá-lo e relatá-lo às partes envolvidas. Isso também inclui a verificação de informações obtidas com o que se acreditava anteriormente, assim podendo resolver potenciais conflitos com o conhecimento que se acreditava anteriormente.

Cada uma das etapas do processo de KDD representam tarefas de grande importância, porém a etapa de MD por ser considerada a mais relevante, dado que esta etapa permite encontrar padrões dentro das bases de dados mediante o uso de diversos algoritmos.

2.2 Mineração de Dados

O processo de MD é responsável por identificar padrões novos, valiosos, potencialmente úteis e compreensíveis (FAYYAD; PIATETAKY-SHAPIRO; SMYTH, 1996 b). Porém é preciso ressaltar que os algoritmos atuais, mesmo sendo capazes de descobrir novos padrões, ainda não conseguem descobrir se os mesmos são valiosos (NAVEGA, 2002). Por este motivo a mineração de dados requer a interação de analistas, responsáveis por determinar a importância dos padrões encontrados.

- a) De acordo com NAVEGA (2002), é possível distinguir dois tipos de objetivos para a MD:
- b) Verificação, onde o sistema é limitado a verificar as hipóteses do usuário;

Descobrimiento, onde o sistema de forma automática encontra novos padrões.

Ainda é possível subdividir o objetivo de descobrimiento em duas partes:

- a) Predição, onde o sistema busca padrões com o objetivo de prever futuros comportamentos de algumas entidades;
- b) Descrição: onde o sistema busca padrões com o objetivo de apresentá-los de maneira compreensível ao ser humano.

A MD utiliza diversos algoritmos na procura de padrões em bases de dados, alguns destes algoritmos são: Redes Neurais Artificiais, MVS, AD, Regressão Linear, Regressão Logística, *Clustering*, Apriori, K-Means. Como já foi dito anteriormente, nesta pesquisa foram utilizados algoritmos de AD's e de MVS que serão descritas a seguir.

2.2.1 Árvores de Decisão

Uma AD representa uma função que toma como entrada um vetor com valores de atributos e retorna um valor de saída único (RUSSELL e NORVING, 2010). As AD consistem em nós que representam os atributos; em arcos provenientes destes nós que recebem os possíveis valores para estes atributos; e nós folhas que representam as diferentes classes de um conjunto de treinamento (INARGOIA, 1996).

No contexto da MD, as AD são algoritmos de classificação de dados, que alcança sua decisão executando uma sequência de testes. Os algoritmos que utilizam as AD usam uma estratégia chamada dividir-para-conquistar, em outras palavras, um problema complexo é subdividido em vários subproblemas menores, que podem ser resolvidos de maneira recursiva (LEMOS; NIMEOLAS; STEINER, 2005).

Existem vários algoritmos que utilizam a estratégia de AD, os principais são: UserClassifier, SimpleCart, RandomTree, DecisionStump e J48. Nesse projeto é usado o algoritmo J48, que surgiu da necessidade de recodificar o algoritmo C45, que foi implementado na linguagem C para JAVA (QUINLAN, 1993). O algoritmo J48 gera um modelo de AD baseada em uma base de dados de treinamento, e usa esse modelo para classificar futuras entradas de dados.

A utilização de algoritmos de AD apresenta as seguintes vantagens: não assumem nenhuma distribuição particular para os dados; os dados podem ser categóricos (qualitativos) ou numéricos (quantitativos); permite a construção de modelos para qualquer função desde que o conjunto de treinamento tenha um grande número de exemplos; e possui um elevado grau de compreensão (LEMOS; NIMEOLAS; STEINER, 2005).

Após a construção de uma AD é importante avaliá-la. Esta avaliação consiste em utilizar a AD em um conjunto de dados que não tenha sido usado no seu treinamento. Esta estratégia permite julgar como a árvore se adapta a novas situações, além de revelar os erros e acertos ocorrido em sua construção. (BRADZIL, 1999).

2.2.2 Maquinas de Vetor De Suporte

As MVS conseguem resolver problemas de regressão e classificação através de sua capacidade de generalização. MVS são capazes de separar instancias das suas classes a partir de uma função, que é obtida a partir dos dados usados na etapa de treinamento. Seu objetivo é criar um classificador que funcione adequadamente em dados que não foram usados em seu treinamento, desta maneira sendo capaz de prever as saídas de futuras entradas de dados (JUNIOR, 2010).

De acordo com RUSSEL e NORVING (2010), existem três propriedades que tornam as MVS atraentes:

1. As MVS constroem um separador de margem máxima com um limite de decisão com maior distância possível a pontos de exemplo. Isso os ajuda a generalizar futuras entradas de dados.
2. As MVS criam uma separação linear em hiperplano (separam as classes que estão em um plano cartesiano utilizando uma reta), mas tem a capacidade de adicionar mais um espaço de dimensão, usando um algoritmo de kernel.
3. As MVS são um método não paramétrico, elas mantêm exemplos de treinamento e podem precisar armazenar todos eles. Mas na pratica, acabam mantendo uma pequena porção do número de exemplos. Assim as MVS têm flexibilidade para representar funções complexas, e são resistentes a super adaptação.

Uma MVS cria sua função (classificador) a partir dos padrões encontrados na sua base de dados de treinamento. Considerando o exemplo ilustrado na Figura 3, existem classificadores lineares que separam as duas classes, porem apenas uma que possui a mesma distância para os elementos de ambas as classes, o hiperplano com margem máxima recebe o nome de hiperplano ótimo (GUNN, 1998).

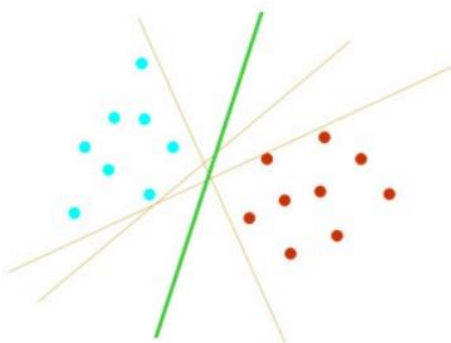


Figura 3 - Possíveis hiperplanos de separação e hiperplano ótimo
fonte: Oliveira Junior (2010)

Porém existem casos em que é impossível separar as classes usando apenas uma reta, nestes casos são necessários executar funções de maior complexidade para tentar contornar o problema.

Essas são as funções de Kernel, que tem a finalidade de projetar os vetores de características de entrada em um espaço de características de alta dimensão para a classificação dos problemas que se encontram em espaços não linearmente separáveis. Pois a medida que se aumenta o espaço da dimensão do problema, também aumenta a probabilidade do problema se tornar linearmente separável. Entretanto, para alcançar uma boa distribuição é necessário um conjunto de treinamento com uma quantidade elevada de instâncias (GONÇAVES, 2011).

A Figura 4 mostra o processo de transformação de um domínio não linearmente separável, em um domínio linearmente separável através do aumento da dimensão onde é feito o mapeamento.

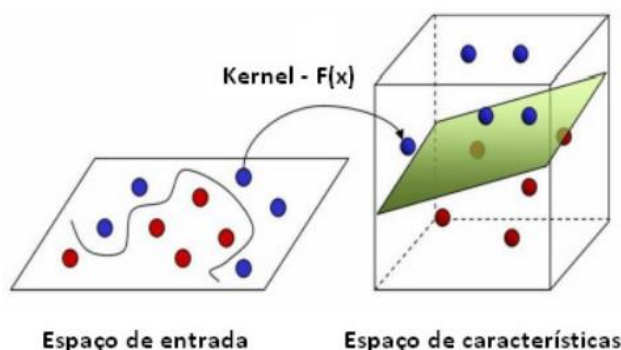


Figura 4 - Processo de transformação de um domínio linearmente não senão sepáravel
Fonte: Oliveira Junior (2010)

3. Estudo de Caso

O estudo de caso analisou o problema de fraudes em um sistema de distribuição de energia elétrica administrado por uma concessionária que pretende diminuir o percentual de perdas comerciais existentes atualmente.

A base de dados fornecida por tal empresa consta de um histórico de 112.431 usuários, para a execução da metodologia de detecção de possíveis fraudes, ela abrange diferentes realidades sociais, como usuários residenciais que pertencem às classes baixa, média e alta, que correspondem a diferentes tipos de consumo e perdas.

Pretende-se analisar os dados repassados pela empresa para realizar o estudo de caso. O banco de dados está composto por informações que são referentes aos registros dos usuários, a leitura dos registrados nos medidores e transformadores, consumo e faturarção mensal dos clientes. Para realizar o processamento dos dados, foi necessário o uso de diferentes ferramentas computacionais, entre elas se utiliza a linguagem de programação JAVA, e o software livre WEKA (versão 3.8).

3.1 Modelo de Extração de Conhecimento

Propõem-se o uso de um modelo de extração de conhecimento que avalia as características dos clientes que são identificados pelo atributo matricula e a partir delas, a criação de regras que serão aplicadas sobre os dados dos clientes. Foram utilizadas quatro regras que avaliam os usuários de acordo com seus históricos de consumo. Estas regras são:

- Classificação do consumo (CLA_CON);

- Risco (RIS);
- Oscilação (OSC);
- Anomalias (ANOM)

Essas regras são processadas de acordo com parâmetros de referência que indicam qual a variação esperada para a classificação dos clientes. A seguir são apresentadas as definições das regras anteriores.

Regra CLA_CON: Os usuários são classificados de acordo com a energia consumida, em consumo baixo (CB), em consumo media (CM), em consumo alto (CA), onde, C_{min} é a menor consumo do usuário, C_{max} é o maior consumo e C_j é o consumo do mês atual. A estrutura de consumo para CB, CM e CA são apresentadas nas expressões (1), (2) e (3) respectivamente.

$$C_{min} \leq C(j) \leq \frac{C_{max} - C_{min}}{3} \quad (1)$$

$$\frac{C_{max} - C_{min}}{3} < C(j) \leq 2 \frac{C_{max} - C_{min}}{3} \quad (2)$$

$$2 \frac{C_{max} - C_{min}}{3} < C(j) \leq C_{max} \quad (3)$$

Regra RIS: Usando um índice de risco, é determinada a possibilidade de fraude de acordo com a comparação do consumo do mês atual, j , e a média dos três meses anteriores. Para os três primeiros meses o índice de risco é igual a zero, caso contrário:

$$IndRis(j) = \frac{C(j) - \frac{(C(j-3)) + (C(j-2)) + (C(j-1))}{3}}{\frac{(C(j-3)) + (C(j-2)) + (C(j-1))}{3}} * C(j) \quad (4)$$

Fazendo o uso do índice de risco é determinado o nível de risco que cada cliente tem de estar cometendo fraude como: alto risco (AR), médio risco (MR), baixo risco (BR) e sem risco (SR), de acordo com as excreções (5) (6) (7) e (8) respectivamente:

$$IndRis(j) < (-0.2 * C(j)) \quad (5)$$

$$(-0.2 * C(j)) \leq IndRis(j) \leq 0 \quad (6)$$

$$0 \leq IndRis(j) \leq (C(j) * 0.2) \quad (7)$$

$$(0.2 * C(j)) \leq IndRis(j) \quad (8)$$

Regra OSC: Usando o índice de oscilação, é determinada a possibilidade de fraude de acordo com a comparação da medição de consumo do mês atual, j , e a medição do mês anterior, $j-1$, levando em conta que para o primeiro mês de consumo do usuário o risco de oscilação é igual a zero, caso contrário.

$$IndOsc(j) = \frac{C(j) - C(j-1)}{C(j-1)} * C(j) \quad (9)$$

Fazendo uso do índice de oscilação é classifica o usuário de acordo com o tipo de variação no consumo como de oscilação descendente (OsD), oscilação normal (OsN) e oscilação ascendente (OsA), de acordo com as excreções (10), (11) e (12) respectivamente.

$$IndOsc(j) < (-0.15 * C(j)) \quad (10)$$

$$(-0.15 * C(j)) \leq IndOsc(j) \leq (0.2 * C(j)) \quad (11)$$

$$(0.2 * C(j)) \leq IndOsc(j) \quad (12)$$

Regra ANOM: Esta regra é utilizada para determinar quais medições são suspeitas e poderiam indicar usuário que devem ser inspecionados. A classificação correspondente a esta regra é determinada através do uso das regras CLA_CON, RIS e OSC. Com esta regra podem ser determinados os valores abaixo.

Possui anomalia (SIM): Se CLA_CON é CB, RIS é AR e OSC é OsD, ou, se CLA_CON é CB, RIS é MR e OSC é OsD, então é preciso inspecionar.

Não possui anomalia (NÃO): Em caso contrário aos descritos acima.

3.2 Construção da Base de Dados

Foram feitas análises na base de dados a fim de selecionar um subconjunto de dados, contendo o que realmente seja necessário. O conjunto de dados foi definido em um modelo que está dividido em três conjuntos: informações do registro do usuário, informações sobre o histórico dos usuários e informações sobre as anomalias de leituras encontradas. Para atender as necessidades da análise, foi necessário considerar as seguintes suposições: um só tipo de usuário (com consumo inferior a 130 kW); usuários com um histórico de medições de no mínimo 12 meses; usuários com todas as medições maiores que zero; e usuários com medições consideradas corretas (sem erro de leitura).

Pré-processamento e transformação: a informação essencial utilizada nessa etapa é o histórico do consumo de cada usuário. Para tal, foi necessária a criação de um software utilizando a linguagem de programação JAVA para o tratamento dos dados fornecidos pela empresa. Este software também é responsável por fazer a leitura dos dados que se encontravam em várias planilhas, a limpeza dos dados, fazer a seleção do conjunto de dados, aplicar as regras do modelo de extração de conhecimento e gerar um documento no formato .CSV com os resultados da aplicação das regras, com o intuito de aplicar posteriormente o processo de MD.

3.4 Aplicando o Algoritmo J48 usando a regra ANOM

Ao aplicar o algoritmo J48 usando a regra ANOM, foi obtido a matriz de confusão da Tabela 2 e a AD da Figura 6. A partir da matriz de confusão é possível observar que, todas as 3.242.541 instâncias do total de 3.242.541 foram classificadas corretamente, para os valores SIM (Possui anomalias) e NÃO (Não possui anomalias), foram classificados corretamente 100.392 e 3.142.149 (100% do total dos dados) respectivamente, enquanto que 0, 5.147 e 613.250 instâncias (19,08% do total dos dados) foram classificadas incorretamente para os mesmos valores. De acordo com as taxas de acertos obtidas pelo modelo, pode-se afirmar que há confiabilidade “Média” nesta classificação.

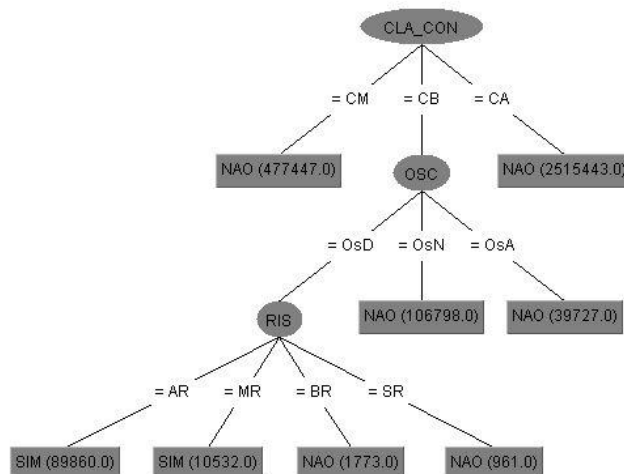


Figura 5 - AD do algoritmo J48 aplicado a regra ANOM

	Não	Sim
Não	3.142.149	0
Sim	0	100.392

Tabela 1 - Matriz de Confusão do algoritmo J48 aplicado a regra ANOM

Observando a AD (Figura 6) é possível fazer as seguintes afirmações:

- Se CLA_CON tem o valor CB e OSC tem o valor OsD e RIS tem o valor AR, então a classificação indica que 89.860 classificações possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsD e RIS tem o valor MR, então a classificação indica que 10.532 classificações possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsD e RIS tem o valor BR, então a classificação indica que 1.773 classificações não possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsD e RIS tem o valor AR, então a classificação indica que 89.860 classificações possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsD e RIS tem o valor SR, então a classificação indica que 961 classificações não possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsN, então a classificação indica que 106.798 classificações não possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsA, então a classificação indica que 39.727 classificações não possuem fraudes.
- Se CLA_CON tem o valor CB e OSC tem o valor OsA, então a classificação indica que 39.727 classificações não possuem fraudes.
- Se CLA_CON tem o valor CM, então a classificação indica que 447.447 classificações não possuem fraudes.
- Se CLA_CON tem o valor CA, então a classificação indica que 2.515.443 classificações não possuem fraudes.

3.5 Aplicando o Algoritmo de MVS suando a regra ANOM

Os treinamentos realizados empregando o algoritmo de MVS permitiram modificações em diversos parâmetros, tipo de MVS, custo, degrau, Kernel, entre outros. Porém as alterações dos parâmetros não mudaram a taxa de acerto, houve mudança apenas nos tempos gastos durante o treinamento e testes de cada modelo, e no tamanho final do modelo.

A porcentagem de acerto obtida na criação dos modelos com aplicação das MVS usando o algoritmo LibSVM foi de 100%, e não foi encontrado nenhum padrão diferente ao mudar as configurações do algoritmo. A tabela 3 apresenta os resultados obtidos nos processos de criação e testes de alguns modelos.

Kernel	Degre	Coef	Custo	Instancia para treinamento	Instancias para teste	Tempo gasto no treinamento	Tempo gasto no teste	Porcentagem de Acerto
Radial Basis Funcion	3	0	1	2.918.287	324.254	71,402	3,869	100%
Linear	3	0	1	2.918.287	324.254	25,928	2,652	100%
Polynomial	3	0	1	2.918.287	324.254	71,542	6,583	100%
Sigmoid	3	0	1	2.918.287	324.254	77,61	4,649	100%
Linear	3	0	5	2.918.287	324.254	24,638	2,467	100%
Linear	3	0	15	2.918.287	324.254	27,586	2,358	100%
Polynomial	3	0	2	2.918.287	324.254	52,245	4,196	100%
Polynomial	3	0	3	2.918.287	324.254	45,926	3,432	100%
Polynomial	3	0	4	2.918.287	324.254	42,339	3,276	100%
Polynomial	3	2	3	2.918.287	324.254	62,166	2,48	100%
Polynomial	3	4	5	2.918.287	324.254	96,83	2,371	100%
Polynomial	1	0	1	2.918.287	324.254	41,333	3,26	100%
Polynomial	3	0	1	2.918.287	324.254	114,598	8,518	100%
Polynomial	5	0	1	2.918.287	324.254	535,269	44,304	100%
Polynomial	1	0	1	2.918.287	324.254	40,065	3,151	100%
Polynomial	5	0	1	2.918.287	324.254	535,269	44,304	100%
Sigmoid	3	0	10	2.918.287	324.255	27,023	1966	100%
Sigmoid	3	0	30	2.918.287	324.254	55,084	2824	100%
Radial Basis Funcion	3	10	1	2.918.287	324254	69,185	2,5175	100%
Radial Basis Funcion	3	30	1	2.918.287	324254	66,6725	2,665	100%
Radial Basis Funcion	10	0	1	2.918.287	324254	66,595	3,11	100%
Radial Basis Funcion	20	0	1	2.918.287	324254	68,0475	2,7875	100%
Radial Basis Funcion	15	8	10	2.918.287	324254	66,2675	2,605	100%
Linear	3	0	25	2.918.287	324.254	25,334	2,469	100%
Linear	3	0	50	2.918.287	324.254	30,807	2,201	100%

Tabela 2 - Modelos MVS aplicados a regra ANOM

3.5 Conclusões

Com base nas informações proporcionada pela empresa distribuidora de energia elétrica, foram classificados os clientes usando as quatro regras propostas neste trabalho: CLA_CON, RIS, OSC e ANOM. Para identificar os valores destes quatro atributos, foi definido um conjunto de regras e índices. A partir destas definições, foi aplicado o processo de MD com o objetivo de auxiliar na resolução do problema proposto, conseguindo padrões de comportamento que assistam na identificação dos usuários suspeitos de estar cometendo algum tipo fraude.

Os resultados obtidos após a aplicação dos algoritmos de MD usando as regras CLA_CON e ANOM indicam distintos graus de precisão na classificação. Observou-se que a escolha da regra a qual o algoritmo de MD é aplicado influencia bastante no resultado final.

Quando o algoritmo J48 é aplicado à regra ANOM foi obtido uma taxa de acerto de 100%, representando uma confiabilidade alta na tomada de decisões. Porém ao aplicá-lo a regra CLA_CON houve uma taxa de acerto de 80%, representando uma confiabilidade menor.

Percebesse que a aplicação das MVS no processo de MD consegue bons resultados, ao aplicar o algoritmo LibSVM a regra ANOM foi obtido uma taxa de acerto de 100%, isso implica que as MVS são eficientes no processo de descoberta de conhecimento.

Verificou-se que por intermédio da MD é possível realizar análises precisas para a detecção de possíveis fraudes nos sistemas de distribuição de energia elétrica. Além de que, evidenciou-se que a escolha da regra a ser usada durante o treinamento do algoritmo de classificação é de grande importância para a obtenção de resultados confiáveis. Observando os resultados dos algoritmos de MD é possível identificar com alto grau de confiança se o consumidor de energia elétrico é suspeito de ser fraudulento ou não.

Foi constatado que o uso de inteligência computacional na toma de decisões que involucram grandes quantidade de dados é uma escolha razoável, válida e eficiente, dado que com os testes realizados foram obtidos resultados satisfatórios.

Referências

A ARAUJO, Antônio Carlos Marques de. *Perdas e Inadimplência na atividade de distribuição de energia elétrica no Brasil.* 2006. Tese (Pós-Graduação em Engenharia) - Programa de pós-graduação de engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

BRACHMAN, R.; ANAND, T. *The Process of Knowledge Discovery in Databases: A Human Centered Approach.* In ADKKM, AAAI/MIT, pag. 37 a 58, 1996.

BRADZIL, P. B. *Construção de modelos de decisão a partir de dados.* 1999. Disponível em: <<http://www.nacc.up.pt/~pbradzil/Ensino/ML/ModDecis.html>>. Acessado em: 15 de março de 2015.

CASTANHELAS, Luciana Gomes. *Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões.* 2008. Dissertação (Pós-Graduação em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

COPEL, *Relatório de Sustentabilidade 2014,* Curitiba, 30 de Abril de 2015.

DELAIBA, A. C.; REIS FILHO, J.; GONTIJO, E. M.; MAZINA, E.; CABRAL, J.E.; PINTO, J. O. P. *Fraud Identification in Electricity Company Costumers Using Decision Tree.* Netherlands, International Conference on System, Man and Cybernetics, outubro de 2004.

DELGADO, Jair José Lopes. *Sistema de Informação de Apoio à Detecção de Perdas de Energia Elétrica – O Caso da Electra.* 2010. Dissertação (Departamento de Eletrônica, Telecomunicações e Informática - UA) Programa para obtenção de grau mestre em Engenharia Eletrônica e Telecomunicação vertente e Sistema de Informação, Universidade de Aveiro 2010.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in knowledge discovery & data mining.* Menlo Park, CA, USA: AAAI/MIT, 1996 a.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. *Knowledge Discovery and Data Mining: Toward a Unifying Framework.* Redmond, WA, USA, AAAI, 1996 b.

GUNN, S. R. *Support Vector Machine for Classification and Regression.* Reporte Técnico, University of Southampton, maio de 1998.

GONÇALVES, André Ricardo. *Otimização em Ambientes Dinâmicos com Variáveis Contínuas empregando algoritmos de estimação de distribuição.* 2011. Dissertação (Mestrado em Engenharia Elétrica) - Departamento de Engenharia de Computação e Automação Industrial, Universidade Estadual de Campinas, Campinas, 2011.

LEMO, E. P; STEINER, M. T. A; NIEVOLA, J. C. *Análise de crédito bancário por meio de árvores de decisão.* Revista Administração USP, São Paulo, p. 255 a 234, setembro de 2005.

NAVEGA, S. *Princípios Essenciais do Data Mining.* Cenadem, Anais da InfoImagem, novembro de 2002.

OLIVEIRA JUNIOR, Gilson Medeiros de. *Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado.* 2010. Trabalho de Graduação (Centro de Informática - UFPE) Graduação em Ciências da Computação, Universidade Federal do Pernambuco, 2010.

FRANCO E. M. C.; LÓPEZ G. P.; QUEIROZ A. S. *Algoritmos de Inteligência Computacional Utilizados na Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica,* XI LATIN AMERICAN CONGRESS ELECTRICITY GENERATION AND TRANSMISSION - CLAGTEE, 2015.

QUINLAN, J.C. C4.5: Programs for machine learning, San Mateo, Morgan Kaufmann, 1993.

RUSSEL, Stuart J; NORVING, Peter. *Inteligência Artificial.* Editora Hall, livro de 2010.