TRANSFORMANDO DADOS EM INFORMAÇÕES PARA APOIO À DECISÃO: USO DE DATA MINING NO COMBATE ÀS PERDAS COMERCIAIS DE ENERGIA ELÉTRICA

André Luiz Medeiros
Universidade Federal de Itajubá
Av. BPS, 1303 – Pinheirinho – CEP: 37.500-903 – Itajubá/MG
andremedeiros@unifei.edu.br

Ricardo Pereira Reis
Universidade Federal de Lavras
Campus Universitário – Caixa Postal 3037 – Lavras/MG
ricpereis@dae.ufla.br

RESUMO

Nas últimas décadas, empresas como as distribuidoras de energia elétrica, armazenaram milhares de gigabits em dados (cadastro, localização, consumo etc.). Entretanto, elas enfrentam dificuldades na solução de problemas por não conseguirem transformar esses dados em informações relevantes para decisão. Nesse sentido, o combate às perdas comerciais de energia é um exemplo. Medidas simples, como inspecionar instalações suspeitas de consumo irregular de energia, poderia resultar em uma decisão técnica e economicamente inviável. Pois muitos consumidores são apenas suspeitos de irregularidade no consumo de energia. Considerando esse problema, este trabalho tem como objetivo utilizar a Mineração de Dados como fundamento para gerar informações que apoiem a decisão de seleção de instalações com suspeita de irregularidade no consumo de energia, incrementando, consequentemente, o Índice de Acerto (IA) das inspeções realizadas. Os resultados mostraram que o fundamento usado é válido, pois ao usar as informações geradas, o IA poderia ser aumentado em 30%.

PALAVARAS CHAVE. Perdas comerciais, Mineração de dados, Modelo de decisão.

EN - OR in Energy // ESN - IO en Energía // EN - PO na Área de Energia

ABSTRACT

In recent decades companies such as electricity distribution companies, stored thousands of gigabits of data (registration, location, consumption etc.). However, they face difficulties in solving problems because they cannot transform these data into information relevant for decision. In this sense, the commercial losses of power is an example. Simple measures such as inspecting facilities suspected irregular energy consumption could result in a technically and economically feasible decision. For lot of consumers are just a suspected irregularity in energy consumption. Considering this problem, this paper aims to address this question of irregular electrical consumption investigation by presenting a methodology based in data mining, which selects and validates the variables to be used by a decision-making model in order to improve inspection quality. Results analysis showed that it would be possible to improve the inspections by up to 30%, considering the same time span and the same quantity of inspection teams, thus enabling a more economically effective means of combating electrical energy theft.

KEYWORDS. Commercial losses, Data Mining. Decision-making model.

EN - OR in Energy // ESN - IO en Energía // EN - PO na Área de Energia

1. Introdução

Nas últimas décadas, as empresas investiram milhares de dólares em tecnologia da informação. Isso possibilitou que, ao longo dos anos, elas acumulassem dados sobre clientes e processos/operações internas. Entretanto, até pouco tempo, esses dados dificilmente eram transformados em informações para apoiar decisões ou auxiliar na solução de problemas.

A situação descrita não é exclusividade de pequenas e ou médias empresas. Grandes empresas, como as distribuidoras de energia elétrica também enfrentam o mesmo problema. Mesmo tendo à disposição os mais diversos dados sobre os clientes como: cadastro, localização, consumo atual, consumo, dentre outros, as distribuidoras possuem dificuldades em transformá-los em informações relevantes para, por exemplo, combater as perdas comerciais de energia.

As perdas comerciais ou não técnicas de energia elétrica resultam de ações de furto ou fraude de energia por parte dos consumidores. Assunto que tem sido matéria prioritária de órgãos reguladores e de concessionárias distribuidoras de eletricidade. A justificativa é que a legislação do setor elétrico brasileiro estipula um valor máximo de repasse das perdas, técnicas e não técnicas (comerciais), à tarifa de energia elétrica dos consumidores. Assim, uma redução das perdas comerciais, permitiria, além de um aumento na receita das concessionárias, uma diminuição da tarifa de energia elétrica beneficiando, portanto, todos os consumidores.

Atualmente as distribuidoras de energia conseguem, usando os bancos de dados de clientes e as regras de consumo, identificar, por exemplo, quais instalações (unidades consumidoras ou clientes) são suspeitas de consumo irregular. Mesmo de posse dessa informação, auditar todas as instalações suspeitas poderia não ser nem técnica e nem economicamente viável.

Nesse contexto, uma das alternativas para solucionar o problema seria melhorar a qualidade da informação disponível quando da realização das inspeções em instalações suspeitas. Ou seja, seria necessário inspecionar instalações que possuem a maior probabilidade de irregularidade no consumo de energia.

A justificativa para o presente estudo fundamenta-se no questionamento de que bases de dados com variáveis pouco relevantes podem gerar informações de baixa significância, o que compromete a decisão gerencial das empresas. Por isso, selecionar variáveis capazes de gerar informações de qualidade, além de acelerar o processamento de dados, melhora a qualidade da decisão a ser tomada e o retorno que ela pode gerar.

Assim, por meio deste estudo, busca-se avaliar a mineração de dados (*data mining*) como fundamento para gerar informações que apoie a decisão de seleção das instalações suspeitas de irregularidade no consumo de energia, incrementando o Índice de Acerto (IA) das inspeções realizadas. Especificamente, pretende-se usar a mineração de dados (*data mining*) para: 1) selecionar as variáveis com maior potencial de gerar informações e que apoiem as inspeções de campo; 2) validar as variáveis selecionadas; e 3) testar se as variáveis selecionadas geram melhorias no Índice de Acerto (IA) das inspeções.

2. Perdas não técnicas ou comerciais

A perda global de energia de uma distribuidora pode ser definida como a diferença entre a energia fornecida a uma determinada rede elétrica e a energia faturada a essa mesma rede. No entanto, essa perda pode ser dividida em técnica (quando relacionada aos materiais e equipamentos utilizados) e em comercial (perda não técnica, quando relacionada, principalmente, à inadimplência, ao furto de energia e à fraude no consumo de eletricidade).

Segundo o Instituto Acende Brasil (2007), as perdas não técnicas de energia, foco desse trabalho, geraram prejuízos da ordem de R$ 6 bilhões ao ano, com importantes reflexos sobre o valor da tarifa de energia e sobre a eficiência econômica do país. Apesar disso, ressalta-se que os órgãos reguladores das distribuidoras não fornecem incentivos adequados para que as empresas combatam eficientemente essas perdas. De acordo com Brasil (2008), são raros os estudos sobre perdas não técnicas, sendo que a maioria das pesquisas foca as perdas técnicas e suas formas de mensuração. Mesmo assim, alguns trabalhos discutem e avaliam, de forma geral, as perdas não técnicas, como: Dick (1995); Smith (2004); Steadman (2010); Onat (2010); Calili (2005);

Queiroga (2005); Reis (2005); Vieiralves (2005); Ortega (2008); Penin (2008); Bernardes (2010); dentre outros.

O furto e a fraude de energia, de modo geral, caracterizam-se pelo uso irregular de energia causado, principalmente, pela ação de terceiros ou por equipamentos defeituosos. A fraude pode ser caracterizada por ações como (QUEIROGA, 2005; REIS, 2005; VIEIRALVES, 2005; ORTEGA, 2008; PENIN, 2008): a) a violação ou adulteração do medidor de energia, com a intenção de redução ou eliminação de consumo; b) a religação direta à rede após o corte de energia da unidade consumidora; c) problemas técnicos nas instalações elétricas do consumidor; e d) deterioração de equipamentos.

3. Mineração de dados (*Data Mining*)

Descobrir conhecimento em base de dados é um campo de pesquisa em ascensão, e cujo desenvolvimento tem sido dirigido ao benefício de necessidades práticas, sociais, econômicas, entre outras. A justificativa é que essas bases de dados possuem informações valiosas com tendências e padrões que poderiam ser usados para apoiar e melhorar as decisões (REZENDE *et al.*, 2003).

Mas, extrair informações relevantes de bases de dados não é uma tarefa comum. Para Fayyad, Piatetsky-Shapiro e Smyth (1996) e Santos e Ramos (2009), os princípios para a Descoberta de Conhecimento em Bases de Dados (DCBD) ou *Knowledge Discovery in Databases* (KDD) integram teorias, métodos e algoritmos de diversas áreas (inteligência artificial, aprendizagem automática, reconhecimento de padrões, estatística, bases de dados e sistemas de informação). E, de acordo com Tan, Steibach e Kumar (2009), as técnicas tradicionais de análise de dados podem não gerar as informações necessárias devido ao tamanho dos conjuntos de dados que geralmente são processados.

A Mineração de Dados (*Data Mining* – DM) é um dos fundamentos usados para se obter informação para apoiar na decisão, a partir de bases com grande volume de dados. Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996) e Tan, Steibach e Kumar (2009), DM é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis dos dados.

Considerando esse fundamento, algumas pesquisas já foram realizadas na tentativa de estudar o problema das perdas comerciais. No trabalho de Nagi *et al*. (2008), por exemplo, utilizou-se a DM, especificamente o algoritmo *Support Vector Machine* (SVM), para pré-selecionar os clientes a serem inspecionados com base em irregularidades e comportamento de consumo anormal. Como resultado, o trabalho gerou classes para classificação que eram utilizadas para selecionar os suspeitos. Já no trabalho de Queiroga (2005), o objetivo era identificar, por meio do uso de DM, padrões que indicassem a possibilidade de fraude, cujos resultados comparativos mostraram índices de acertos da ordem de 25% a 45%. Reis Filho (2006), também usando DM, teve como objetivo desenvolver um sistema de auxílio à detecção de fraudes em unidades consumidoras e identificar medidores de energia com problemas, sendo que os índices de acerto ficaram em torno de 40%.

4. Material e Método

4.1 Objeto de estudo

Este trabalho teve como objeto de estudo uma das 64 distribuidoras de energia elétrica que atuam no Brasil. Essa empresa faz parte de um grupo empresarial que é reconhecido por sua dimensão e competência técnica, sendo considerada a maior empresa integrada do setor de energia elétrica do Brasil, constituída por 58 empresas e 10 consórcios.

A principal justificativa para a escolha dessa distribuidora é que, além de atuar internacionalmente e em dezenove estados brasileiros mais o Distrito Federal, a organização se comprometeu, mediante contrato de confidencialidade, a disponibilizar as bases de dados necessárias para o desenvolvimento do trabalho.

Apesar da atuação nacional, apenas uma região metropolitana, de um único Estado brasileiro, foi analisada, pois de acordo com a distribuidora a região pesquisada apresentava

grande diversidade de situações. Ao longo do estudo, essa distribuidora será referenciada como **Distribuidora de Energia Elétrica em Estudo (DEEE)**.

4.2 Classificação e metodologia utilizada na pesquisa

Este estudo investigou e buscou respostas para problemas que estão ligados a processos organizacionais. Assim, conforme Bertrand e Frasso (2002), Gil (1999) e Silva e Menezes (2005), este trabalho pode ser classificado como: 1) pesquisa aplicada quanto à natureza; 2) pesquisa normativa e descritiva quanto aos objetivos; 3) pesquisa quantitativa quanto à forma de abordar o problema e 4) pesquisa-ação quanto aos procedimentos técnicos.

A pesquisa-ação aplica-se em casos em que o estudo é concebido e realizado em estreita associação a uma ação, ou para resolver um problema coletivo ou específico. E para atender aos objetivos deste trabalho foi utilizado o modelo proposto por DeLurgio (1998), apresentado no Quadro 1.

Quadro 1. Passos do modelo de DeLurgio (1998), utilizado para transformar dados em informação e auxiliar no combate às perdas comerciais.

Passos	Características
1. Definição do problema	Necessidade de resolver um problema, explicar algum fenômeno, planejar e ou prever um evento futuro.
2. Coletar dados/informações	Obtenção de informações sobre o comportamento de um sistema em que o problema ou o fenômeno se encontra.
3: Hipótese/teoria/formulação do modelo	A partir das informações e observações coletadas no passo 2, formulam-se as hipóteses ou um modelo teórico para descrever os fatos importantes que influenciam o problema ou o fenômeno.
4. Escolha e ajuste do modelo	Com o auxílio de ferramentas estatísticas/matemáticas, seleciona-se os modelos e prepara-se experimentos para testar as hipóteses e as teorias.
5. Execução do experimento	Após os ajustes dos dados, o experimento deve ser desenvolvido e executado.
6: Análise dos resultados	Os resultados do experimento devem ser analisados de forma a aceitar ou rejeitar as hipóteses ou o modelo.
7. Validação	Se os resultados apresentados no passo anterior forem válidos, deve-se manter o modelo. Caso contrário, volta-se ao Passo 1.
8. Continuando a manutenção e verificação	Garantir que o modelo ou a teoria sejam válidos e efetivos. Mesmo após o modelo ter sido validado, algumas interações podem ser convergidas para obter um modelo melhor.
9. Sistema sob controle	Se, mesmo após a manutenção e a verificação do modelo, ele tiver problemas, deve-se voltar ao Passo 1 para checar sua consistência.
10. Continuidade do uso	Caso o modelo não apresente problemas, deve-se continuar o uso.

Fonte: Adaptado de DeLurgio (1998).

O desenvolvimento do trabalho seguiu os passos apresentados no Quadro 1 e destaca-se a parceria entre a DEEE e os pesquisadores na execução de todos os passos apresentados.

5. Desenvolvimento da pesquisa

5.1 Passo 1: Definição do problema, objetivo da pesquisa

Para a DEEE, um dos principais problemas enfrentado pela distribuidora é a perda não técnica de energia elétrica causada, principalmente, por fraude nos medidores de energia e por irregularidades técnicas nos equipamentos de medição. Como padrão, foi definido que tanto a fraude quanto a irregularidade técnica seriam tratadas apenas como fraude.

De posse dessa informação, os pesquisadores e a DEEE definiram que o objetivo desse estudo seria usar a Mineração de Dados como fundamento para gerar informações que apoiassem a decisão de seleção das instalações suspeitas de irregularidade no consumo de energia, incrementando, consequentemente, o Índice de Acerto (IA) das inspeções realizadas. Sendo que o Índice de Acerto é a razão entre a quantidade de inspeções com resultado PROCEDENTE (instalações em que há fraude energia elétrica) e o total de inspeções realizadas, ou seja, a soma entre resultado PROCEDENTE e IMPROCEDENTE (instalações sem fraude).

5.2 Passo 2: Coleta de dados/informações da pesquisa

Os dados foram disponibilizados e coletados diretamente na DEEE, reforçando a participação ativa da distribuidora no estudo. Duas bases de dados distintas foram disponibilizadas pela DEEE: 1) uma que continha além das características das instalações, o

resultado das inspeções já realizadas (PROCEDENTE e IMPROCEDENTE), chamada de BASE DE DADOS R; e 2) outra com as características das instalações suspeitas de irregularidades, chamada de BASE DE DADOS S (com inspeções a serem realizada).

Tanto na Base R quanto na S, os dados ficaram limitados a uma região específica da distribuidora (região metropolitana) referente ao período de seis meses. A justificativa para essa escolha foi que, no período considerado, havia grande e relevante volume de dados, com características distintas.

5.3 Passo 3: Hipótese/Teoria/Formulação do problema da pesquisa

Os pesquisadores e a DEEE definiram duas hipóteses para dar sustentação ao objetivo desse trabalho: 1) as variáveis que abrigam os dados da Base de Dados R, possuíam potencial para gerar informações de apoio à decisão; e 2) As informações geradas, a partir das variáveis da Base de Dados R, conseguiriam melhorar o Índice de Acerto (IA) das inspeções a serem realizadas nas instalações pertencentes à Base de Dados S.

Após definir as hipóteses, considerou-se, a priori, que as dezesseis variáveis obtidas na Base de Dados R deveriam ser testadas para medir o potencial de serem utilizadas para atingir o objetivo do trabalho. As variáveis da Base de Dados R, com suas respectivas características e siglas são apresentadas no Quadro 2.

5.4 Passo 4: Escolha e ajuste do modelo a ser utilizado na pesquisa

Para formular o modelo de mineração foi utilizado o aplicativo *Weka: Data Mining Software in Java* (versão 3.6.1). Esse aplicativo foi escolhido por disponibilizar diversos algoritmos e por ser desenvolvido na plataforma de *software* livre (WITTEN e FRANK, 2005). Após a escolha do aplicativo, os dados da Base R passaram por um pré-processamento, seguindo a sequência proposta por Witten e Frank (2005):

1. Resumo dos dados: a primeira etapa foi reunir os dados em um conjunto de exemplos (*instances*). Os exemplos são as informações das instalações inspecionadas da Base R, distribuídas em cada uma das 16 variáveis relacionadas no Quadro 2. Nessa etapa foi necessário agrupar (*merge*) alguns exemplos e limpar dados de variáveis;
2. Dados escassos: a segunda etapa consistiu na análise dos exemplos que, muitas vezes, são compostos de valores iguais a zero ou por amplo conjunto de texto. Esses tipos de dados dificultam a mineração de dados, pois o aplicativo Weka lê e operacionaliza os dados na forma de matriz (com *n* linhas e *m* colunas);
3. Tipo de atributo: a terceira etapa consistiu na avaliação do tipo mais indicado de variável. Os arquivos processados pelo Weka são do tipo *Attribute-Relation File Format* (ARFF) arquivo em formato relação-atributo, e por isso acomodam dois tipos básicos de dados: nominal e numérico. Assim, foi necessário identificar qual o tipo de dado que cada variável deveria assumir;
4. Valores desconhecidos: variáveis com valores desconhecidos (inexistentes) foram substituídas por um ponto de interrogação (*?*), para que o aplicativo processasse os exemplos;
5. Valores imprecisos: é importante verificar os arquivos de mineração de dados cuidadosamente para evitar variáveis e valores não desejáveis;
6. Padronização ARFF: a última etapa consistiu na padronização dos exemplos em um formato cujo conjunto de dados fosse independente, desordenado e sem qualquer relação entre os exemplos.

Feito o pré-processamento, iniciou-se o processamento propriamente dito. Nessa etapa, duas formas de selecionar as variáveis foram utilizadas, a manual e a automática (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996; WITTEN e FRANK, 2005). Na seleção manual, as seguintes variáveis foram retiradas da análise, conforme as justificativas da distribuidora: 1) GR_EXEC_INSPECAO: apenas as informações do serviço da própria DEEE seriam processadas (não seriam analisados serviços prestados por terceiros); 2) EXECUTOR_INSPECAO: há um número muito grande de executores das ordens de inspeção; e 3) RETORNO_SVC: desconsiderar as inspeções em que esse atributo/variável apresente como retorno IMPEDIDO.

Quadro 2. Variáveis com potencial de serem utilizadas na transformação de dados em informações e auxiliar no combate às perdas comerciais.

Atributos/Variáveis	Característica	Sigla
1. Hierarquia política	Local em análise	HIERARQUIA_POLITICA
2. Tensão	Tensão instalada na unidade consumidora	TENSAO
3. Dia de leitura	Dado obtido a partir da unidade de leitura	DIA_LEITURA
4. Local de leitura	Dado obtido a partir da unidade de leitura	LOCAL_LEITURA
5. Rota de leitura	Dado obtido a partir da unidade de leitura	ROTA_LEITURA
6. Classe	Classe em que a instalação de enquadra	CLASSE
7. Ramo de atividade	Ramo de Atividade em que a instalação de enquadra	RAMO_ATIVIDADE
8. Número de fases	Monofásico, Bifásico ou Trifásico	NUMERO_FASES
9. Motivo mais antigo	1º motivo que gerou a suspeita da instalação	MOT_ANTIGO
10. Data do motivo mais antigo menos a data de emissão da nota de inspeção	Dado obtido a partir da diferença entre a data do motivo mais antigo e a data de emissão da nota de inspeção (em dias)	DATAMOTIVO_ DATAENOTA_DIAS
11. Data de emissão da nota de inspeção menos a data de inspeção	Dado obtido a partir da diferença entre a data de emissão da nota de inspeção e a data efetiva da inspeção (em dias)	DATAENOTA_ DATAINSPECAO_DIAS
12. Grupo responsável pela execução da inspeção	Número da equipe de campo que realizou a inspeção	GR_EXEC_INSPECAO
13. Colaborador que executou a inspeção	Número de matrícula do colaborador que realizou a inspeção	EXECUTOR_INSPECAO
14. Retorno do serviço de campo	Procedente, Improcedente, Impedido	RETORNO_SVC
15. Cód. serviço de campo	Código do serviço de campo que indica a situação encontrada na instalação suspeita	COD_SVC
16. Número de irregularidades encontradas	Quantidade de irregularidade encontrada na instalação suspeita	NIE

Fonte: Dados/informações da DEEE.

Na seleção automática, foi utilizado o algoritmo *Ranker* (McGREGOR *et al.*, 2004) também presente no aplicativo Weka. Esse algoritmo foi usado por fazer uma avaliação simples das variáveis, ordenando as que apresentam os melhores resultados. O *rank* das variáveis é apresentado no subitem a seguir.

5.5 Passos 5 e 6: Execução de experimentos e Análise de resultados

Ao rodar o algoritmo *Ranker* com os dados selecionados, cinco métodos de busca se mostraram-se mais adequados para classificar as variáveis. O Quadro 3 apresenta o resultado de cada método de busca e a ordem das variáveis que geraram os resultados mais significativos.

Analisando os resultados do Quadro 3, pode-se afirmar que a ordenação das cinco variáveis que apresentaram os melhores resultados é diferente em cada um dos métodos testados (de A1 a A5). Entretanto, algumas variáveis assumem a mesma ordem em todos os métodos testados, como é o caso de: 1) COD_SVC – que assume sempre a primeira posição na classificação em todos os métodos (de A1 a A5); 2) NIE – que assume a segunda posição na classificação de três algoritmos; e 3) RAMO_ATIVIDADE – com a quinta posição na classificação de dois algoritmos (A4 e A5).

Considerando a diversidade de ordenação de cada um dos métodos apresentados, foi difícil, a partir do Quadro 3, elencar quais variáveis que deveriam ser selecionados para analisar as instalações suspeitas. Na tentativa de obter coerência no resultado, optou-se por criar uma ponderação entre os métodos de busca usados pelo algoritmo *Ranker*, que é apresentado no Quadro 4.

Quadro 3. Métodos de busca e classificação das variáveis indicadas para transformar dados em informação e auxiliar no combate às perdas comerciais, de acordo com o algoritmo *Ranker*.

MÉTODOS DE BUSCA / CARACTERÍSTICA	CLASSIFICAÇÃO DAS VARIÁVEIS
A1 *– ReliefFAtrributeEval*	1 - COD_SVC 2 - LOCAL_LEITURA

Avalia o valor de um atributo por amostragem repetitiva em um exemplo e considera o valor do dado para o atributo mais próximo do exemplo da mesma classe e de classes diferentes. Funciona tanto em classe de dados discretos quanto contínuos.	3 - ROTA_LEITURA 4 - DIA_LEITURA 5 - TENSAO
A2 *– InfoGainAttributeEval* Avalia o valor de um atributo pela medição do ganho da informação com relação à classe.	1 - COD_SVC 2 - NIE 3 - LOCAL_LEITURA 4 - DIA_LEITURA 5 - ROTA_LEITURA
A3 *– ChiSquareAttributeEval* Avalia o valor de um atributo calculando o valor da estatística qui-quadrado em relação à classe.	1 - COD_SVC 2 - NIE 3 - LOCAL_LEITURA 4 - DIA_LEITURA 5 - ROTA_LEITURA
A4 *– GainRatioAttributeEval* Avalia o valor de um atributo pela medição da taxa de ganho em relação à classe.	1 - COD_SVC 2 - TENSAO 3 - NIE 4 - CLASSE 5 - RAMO_ATIVIDADE
A5 *– SymmetricalUncertAttributeEval* Avalia o valor de um atributo através da medição da incerteza simétrica em relação à classe.	1 - COD_SVC 2 - NIE 3 - TENSAO 4 - LOCAL_LEITURA 5 - RAMO_ATIVIDADE

Fonte: Resultados da pesquisa.

No Quadro 4, cada uma das 16 variáveis foi classificada como “M” (Manter na análise) ou como “E” (Excluir da análise) considerando o critério estabelecido entre os pesquisadores e a DEEE: a) variáveis classificadas como “M” – aquelas que assumiram as oito melhores classificações de acordo com cada método de busca (de A1 a A5); e b) variáveis classificadas como “E” – aquelas que assumiram as oito piores classificações de acordo com os métodos de busca.

Para definir a coluna “Resultado” (Quadro 4), somou-se a quantidade de “M” e “E” que cada variável obteve na classificação. E, adotando o mesmo critério da classificação anterior, as oito variáveis com o maior número de letra “M” foram mantidas na análise; sendo que as demais foram excluídas da análise. Vale ressaltar que o critério de manter ou excluir variáveis da análise também foi estabelecido em comum acordo entre os pesquisadores e a distribuidora.

Avaliando o resultado final apresentado (Quadro 4), a DEEE e os pesquisadores julgaram que as variáveis mantidas na análise realmente deveriam ser selecionadas para gerar informações. Entretanto, a DEEE ponderou que a variável NUMERO_FASES, apesar ter recebido a recomendação de ser excluir da análise, deveria ser mantida por representar a complexidade das inspeções a serem realizadas e determinar, portanto, o tipo de equipe que deverá ser enviado para realizar a inspeção.

Considerando o argumento da distribuidora de energia elétrica e ponderando os benefícios da variável, optou-se por mantê-la na análise. Com isso, ao todo, nove variáveis foram selecionadas, as oito variáveis melhor classificadas pelos métodos de busca e a variável NUMERO_FASES. O passo seguinte foi validar se as variáveis selecionadas eram capazes de gerar informações que apoiassem a decisão de seleção de instalações suspeitas de irregularidade.

5.6 Passo 7: Validação das variáveis selecionadas

Para considerar as variáveis selecionadas válidas, os pesquisadores e a DEEE estabeleceram que os resultados a serem gerados deveriam ser o mais próximo possível do resultado encontrado na Base de Dados R (que apresenta o resultado das inspeções já realizadas, como descrito no subitem 5.2).

Quadro 4. Ponderação dos métodos de busca, para selecionar as variáveis indicadas para transformar dados em informação e auxiliar no combate às perdas comerciais, de acordo com o algoritmo *Ranker*.

Variáveis	A 1	A 2	A 3	A 4	A 5	Total M	Total E	Resultado
HIERARQUIA_POLITICA	M	M	M	M	M	5	0	Manter na análise
TENSAO	M	M	M	M	M	5	0	Manter na análise
DIA_LEITURA	M	M	M	M	M	5	0	Manter na análise
LOCAL_LEITURA	M	M	M	M	M	5	0	Manter na análise
ROTA_LEITURA	M	M	M	E	E	3	2	Manter na análise
CLASSE	M	M	M	M	M	5	0	Manter na análise
RAMO_ATIVIDADE	E	M	M	M	M	4	1	Manter na análise
NUMERO_FASES	E	E	E	M	E	1	4	Excluir da análise
MOT_ANTIGO	E	E	E	E	E	0	5	Excluir da análise
DATAMOTIVO_ DATAENOTA_DIAS	E	E	E	E	E	0	5	Excluir da análise
DATAENOTA_ DATAINSEECAO_DIAS	E	E	E	E	E	0	5	Excluir da análise
COD_SVC	M	M	M	M	M	5	0	Manter na análise
NIE	M	E	E	E	M	2	3	Excluir da análise

Fonte: Resultados da pesquisa.
Nota: M: manter a variável na análise; E: excluir a variável da análise.

Como forma de comparar e padronizar o resultado a ser encontrada a partir das variáveis selecionadas, desenvolveu-se a classificação apresentada na Figura 1.

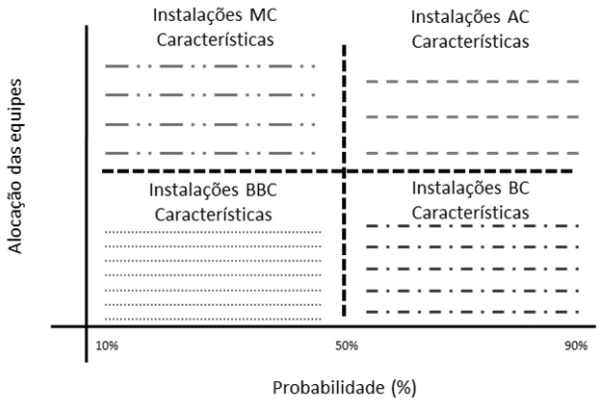

Figura 1. Classificação esperada das instalações suspeitas de irregularidade, a partir das variáveis selecionadas para transformar dados em informação e auxiliar no combate às perdas comerciais.
Fonte: Dados da pesquisa.

A classificação das instalações suspeitas de irregularidade (Figura 1), deve apresentar agrupamentos (*clusters*) de instalações com diferentes níveis de complexidade, sendo que o nível de complexidade das instalações, definido pela DEEE, foi: 1) Alta Complexidade (Instalações AC) – apresentam elevada probabilidade de se encontrar irregularidade e por isso devem receber prioridade na liberação das inspeções; 2) Média Complexidade (Instalações MC) – são aquelas com média complexidade de se encontrar irregularidade e a prioridade na liberação das inspeções deve ser menor em relação à AC; 3) Baixa Complexidade (Instalações BC) – apresentam baixa complexidade de se encontrar irregularidade e deveriam receber pouca prioridade na liberação de inspeção; e 4) Baixíssima Complexidade (Instalações BBC) – são aquelas com mínima prioridade na inspeção das instalações suspeitas.

Para atingir os resultados apresentados na Figura 1, as variáveis poderiam ser analisadas por diversas técnicas ou diferentes conjuntos de algoritmos (GOLDBERG, 1989; FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996; WITTEN e FRANK, 2005). Nesse trabalho utilizou-se como fundamento a Mineração de Dados, especificamente os algoritmos de *cluster*. Essa escolha se justifica porque o algoritmo deve ser aplicado quando não existe uma classe específica a ser prevista, mas sim quando os casos devem ser divididos em grupos naturais (WITTEN e FRANK, 2005). Assim, os grupos resultantes desses algoritmos refletem mecanismos de domínio, a partir dos quais é possível encontrar pontos de semelhança. Com isso, a reposta de um *cluster* pode ter a forma de um diagrama que mostra como as instalações suspeitas se agrupariam (nos clusters). Os algoritmos desse tipo tendem, naturalmente, a gerar resultados semelhantes ao

esquema da Figura 1.

Nesse trabalho, o algoritmo de *cluster* usado para validar a seleção das variáveis foi o *SimpleKMeans*. Essa escolha se justifica porque além de ser simples é o mais difundido e utilizado em estudos semelhantes (McGREGOR *et al*., 2004; WITTEN e FRANK, 2005). Ao processar os dados das variáveis selecionadas da Base R, usando o algoritmo *SimpleKMeans*, os seguintes resultados foram encontrados:

1. 37,50% dos casos foram agrupados no cluster em que as variáveis possuíam características como: BT; 13_D; L_8015; R_30; RESIDENCIAL; RA_380; PROCEDENTE. Esse agrupamento pode ser comparado com as Instalações AC, da Figura 1;
2. 11,36% dos casos possuíam variáveis com características como: MT; 44_D; L_0109; R_02; OUTROS_SERVICOS_E_OUTRAS_ATIV; RA_999; PROCEDENTE, agrupamento comparado com as Instalações MC;
3. 5,60% dos casos apresentaram variáveis com as seguintes características: BT; 15_D; L_0530; R_25; RESIDENCIAL; RA_380; PROCEDENTE; agrupamento comparado com as Instalações BC;
4. 45,45% dos casos foram agrupados no cluster em que as variáveis eram caracterizadas por: BT; 03_D; L_0631; R_12; RESIDENCIAL; RA_380; IMPROCEDENTE; agrupamento comparado com as Instalações BBC.

A partir desses resultados, foi possível classificar as instalações da Base R como apresentado na Figura 2.

Analisando os resultados da Figura 2, os pesquisadores e a DEEE concluíram que as instalações foram agrupadas de forma semelhante aos resultados reais obtidos pelas inspeções que constavam na Base de Dados R. Entretanto, um fato que chamou a atenção foi que os percentuais das instalações dos agrupamentos (AC, MC, BC e BBC) eram diferentes dos realmente obtidos com as inspeções. Diferença que pode ser explicada pela característica do algoritmo de *cluster*. Para o algoritmo, o conjunto de variáveis selecionadas possibilitou que uma dada instalação se agrupasse, por exemplo, às Instalações AC, por ter dados comuns a esse grupo. Entretanto, na prática, ela pertence às Instalações MC. Dessa forma, é possível afirmar que, mesmo com o delineamento dos grupos (AC, MC, BC e BBC), há uma área de interseção entre eles. Essa interseção pode ser explicada pelas variáveis que foram retiradas na análise, como apresentado no subitem 5.5.

Apesar dessa diferença, pode-se afirmar que as variáveis geraram informações válidas para a seleção de suspeitos de fraude. Além disso, ela já era esperada, pois o algoritmo está lidando apenas com parte dos dados que representa o mundo real.

5.7 Passo 8: Continuando a manutenção e verificação

Para certificar se as variáveis selecionadas eram realmente capazes de gerar informações válidas para apoiar a decisão, um novo teste foi realizado. Nesse teste, foram analisadas outras 2.499 instalações suspeitas de irregularidade da Base de Dados S (suspeitos). Essas instalações, mesmo pertencendo à Base S, já haviam sido inspecionadas pelo serviço de campo da DEEE, mas os pesquisadores, até então, não tinham conhecimento dos resultados das inspeções, o que permitiu comparar os resultados gerados pela combinação das variáveis selecionadas e mineração de dados com o resultado encontrado pela equipe de campo da DEEE.

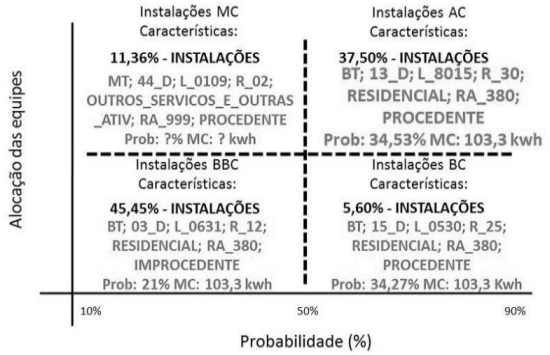

Figura 2. Classificação das instalações da Base de Dados R, a partir das variáveis selecionadas para transformar dados em informação e auxiliar no combate às perdas comerciais.
Fonte: Resultado da pesquisa.

Figura 3. Classificação das instalações suspeitas da Base de Dados S a partir da combinação entre variáveis selecionadas e a mineração de dados, para transformar dados em informação e auxiliar no combate às perdas comerciais.
Fonte: Resultado da pesquisa.

Ao inspecionar as 2.499 instalações a equipe de campo da DEEE apurou que, do total, 1.420 foram classificadas como PROCEDENTE e 1.079 como IMPROCEDENTE, resultando em um Índice de Acerto (IA) de 56,8%. A partir dessas informações, foi possível comparar os resultados gerados pela combinação entre as variáveis selecionadas e mineração de dados, com o IA obtido nas inspeções realizadas pela DEEE. A Figura 3 compara a classificação das instalações da Base de Dados S, a partir da combinação com o resultado obtido pela DEEE.

Considerando a Figura 3, verifica-se que das 2.499 instalações suspeitas analisadas, a combinação registrou como PROCEDENTE 1.347 instalações (53,9% – AC e BBC) e como IMPROCEDENTE 1.152 instalações (46,1% – MC e BC).

Do total PROCEDENTE, quase 33,0% foi classificado como Instalações AC e pouco mais de 21,0% classificadas como Instalações BBC. Já as instalações julgadas como IMPROCEDENTE 24,04% foram classificadas como Instalações MC e 22,04% foram classificadas como Instalações BC.

Avaliando o IA resultante dessa combinação, constatou-se que ela ajustou corretamente quase 61,0% das instalações registradas como PROCEDENTE, ou seja, todas as instalações classificadas como AC. Do total registrado como IMPROCEDENTE, o IA da combinação foi de quase 48,0%, ou seja, todas as que foram classificadas como BC. Por outro lado, a combinação não gerou um bom ajuste nas instalações classificadas como MC e BBC.

Operacionalizando outra análise, pode-se afirmar que a combinação gerou um conjunto intermediário de resultados informando que 601 instalações seriam IMPROCEDENTES sendo que, nas inspeções da DEEE, o resultado foi PROCEDENTE. Essa situação pode ser traduzida como uma falha da combinação, pois há fraude na instalação, mas ela não seria inspecionada. De forma semelhante, a combinação também considerou que 528 inspeções teriam resultado PROCEDENTES, sendo que na verdade o resultado de campo da DEEE foi IMPROCEDENTE, ou seja, não há fraude na instalação, mas ela seria inspecionada. Apesar de menos grave do que a situação anterior, ela gera um desperdício de recursos ao inspecionar instalações sem evidências de fraude.

Por fim, para certificar-se da relevância da informação gerada pelas variáveis selecionadas, combinado com o processamento fundamentado na Mineração de Dados, dois outros indicadores foram sugeridos pelos pesquisadores: 1) o Índice de Acerto Total (IA Total); e 2) o Índice de Acerto Ajustado (IA Ajustado).

O IA Total é o percentual obtido pela razão entre a quantidade total instalações inspecionadas com o resultado PROCEDENTE e o total de instalações inspecionadas (com resultado PROCEDENTE e IMPROCEDENTE). Esse índice se aplicaria somente se as informações geradas pela combinação (variáveis selecionadas e mineração de dados) fossem

usadas como base para se realizar as inspeções de campo. Assim, no caso do teste realizado, o IA Total seria de 60,8%, pois ao invés de inspecionar as 2.499 instalações, seriam inspecionadas apenas 1.347, das quais 819 teriam resultado PROCEDENTE e 528 teriam resultado IMPROCEDENTE.

Já o IA Ajustado é o percentual obtido pela razão entre o total de acerto conseguido (resultado PROCEDENTE e IMPROCEDENTE) e o total de instalações que deveriam ser inspecionadas, caso as informações geradas pela combinação fossem usadas como referência. Considerando esse indicador, o IA Ajustado do teste realizado seria de 72,18%, pois:

1) As informações produzidas pela combinação (variáveis selecionadas e Mineração de Dados) possibilitariam acertar o resultado PROCEDENTE de 819 instalações inspecionadas e acertaria também o resultado IMPROCEDENTE de 551 instalações que não seriam inspecionadas. Ou seja, com isso a combinação acertaria o resultado de 1.370 instalações;
2) Considerando que a informação gerada pela combinação fosse usada, apenas 1.898 instalações teriam sido inspecionadas e não 2.499;
3) A razão entre o total de acerto (item 1) e o total de instalações que deveriam ser inspecionadas (item 2), obtêm-se o resultado percentual de 72,18%.

Com base nos resultados apresentados, pode-se afirmar que usar a Mineração de Dados como fundamento para transformar dados em informação pode gerar bons resultados para as distribuidoras no combate à fraude no consumo de energia elétrica. Pois o Índice de Acerto (IA) resultante das informações geradas pela combinação (variáveis selecionadas e mineração de dados) foi sempre superior ao IA que a DEEE conseguiu nas inspeções realizadas. Se a DEEE tivesse usado as informações geradas pela combinação, ela teria conseguido uma melhora de 27,03% no índice de acerto, considerando o mesmo espaço de tempo e sem aumentar o número de equipes de inspeção.

5.8 Passos 9 e 10: Sistema sob controle e Continuidade do uso

Diante dos resultados obtidos, os pesquisadores e a DEEE julgaram que o fundamento da Mineração de Dados é válido para gerar informações que apoiem o combate às fraudes no consumo de energia elétrica. Por isso, novas comparações entre os resultados obtidos pelas equipes de campo da DEEE e as informações geradas por meio da mineração de dados deveriam ser realizadas.

6. Considerações finais

Avaliando os resultados apresentados, pode-se afirmar que o estudo atingiu o objetivo proposto. Pois, se a Mineração de Dados (data mining) tivesse sido usada como fundamento para gerar informações que apoiassem a decisão de seleção das instalações suspeitas de irregularidade no consumo de energia, o Índice de Acerto (IA) da distribuidora poderia ter tido uma melhora de aproximadamente 30,0%, considerando o mesmo horizonte de tempo e a mesma quantidade de equipes de inspeção. Esse resultado aumentaria a viabilidade técnica e a econômica do combate às perdas comerciais.

Além disso, a metodologia usada pode ser considerada válida, pois foi possível selecionar e validar as variáveis mais relevantes a serem utilizadas para gerar informações que apoiassem a decisão da distribuidora, melhorando, consequentemente, o resultado das inspeções realizadas.

Para trabalhos futuros, sugere-se que os procedimentos usados neste trabalho sejam implementados em um sistema de apoio à decisão para que as distribuidoras de energia consigam, a partir das informações a serem geradas, melhorar a decisão de inspeção das instalações suspeitas de fraudar o consumo de energia elétrica e tornar o combate às perdas comerciais mais efetivo.

Agradecimentos

Merece destaque e agradecimento a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) que apoia o grupo de pesquisa no qual o projeto foi realizado. Além disso, deve-se reconhecer o importante papel desempenhado pela CAPES na expansão e consolidação da pós-graduação.

Referências

BERNARDES, M. L. (2010) *Proposta de um plano estruturado de ação para atenuação de perdas não técnicas de distribuição de energia elétrica em uma empresa do Rio Grande do Sul*. Dissertação de Mestrado, Universidade do Vale do Rio dos Sinos, Programa de Pós-graduação em Engenharia de Produção e Sistemas. 2010.

BERTRAND, J. W. M.; FRASSO, J. C. (2002) Modelling and simulation. Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, v. 22. n. 2. p. 241-264. 2002.

BRASIL. (2008) Agência Nacional de Energia Elétrica. Superintendência de Regulação dos Serviços de Distribuição. **Nota técnica nº 342/2008**. Brasília: SER/ANEEL, 2008.

CALILI, R. F. (2005) *Desenvolvimento de sistema para detecção de perdas comerciais em redes de distribuição de energia elétrica*. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro, 2005, 2.2.

DeLURGIO, S. A. (1998) Forecasting principles and applications. 1st Edition. Singapore: McGraw-Hill. 802p. 1998

DICK, A. J. (1995) Theft of electricity – How UK electricity companies detect and deter. *Security and Detection, 1995, European Convention on*, pp.90-95, 16-18 May 1995. doi: 10.1049/cp:19950476.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996) From Data Mining to Knowledge Discovery in Databases. *AI MAGAZINE* [S.I.], v. 738, n. 4602, p. 18, 1996.

GIL, A. C. (1999) *Métodos e técnicas de pesquisa social*. São Paulo: Atlas. 1999.

GOLDBERG, D. E. (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley. 1989.

INSTITUTO ACENDE BRASIL. (2007) Perdas e inadimplência no setor elétrico. In: *Cadernos de Política Tarifária*. Análise do processo de revisão tarifária e da regulação por incentivos. São Paulo: Instituto Acende Brasil, 2007. Disponível em: <http://www.acendebrasil.com.br/archives/files/estudos/Caderno_05_Perdas_e_ Inadimplencia.pdf >. Acesso em: 29 jun. 2011.

MCGREGOR, A.; HALL, M.; LORIER, P.; BRUNSKILL, J. (2004) Flow Clustering Using Machine Learning Techniques. *In:* BARAKAT, C.; PRATT, I. Passive and Active Network Measurement. Lecture Notes in Computer Science. Springer Berlin: Heidelberg. 2004. p.205-214

NAGI, J.; MOHAMMAD, A. M.; YAP, K. S.; TIONG, S. K.; AHMED, S. K. (2008) Non-tecnical loss analysis for detection of electricity theft using Support Vector Machines. In: 2nd. IEEE International Conference on Power and Energy (PECon 08). December 1-3, 2008, Johor Bahru. Malaysia.

ONAT, N. (2010) Transmission and distribution losses of Turkey´s power system. *In*: Advances in Energy Planning, Environmental Education and Renewable Energy Sources. Tunisia. 2010. ISSN: 1790-5095, ISBN: 978-960-474-187-8.

ORTEGA, G. V. C. (2008) Redes neurais na identificação de perdas comerciais do setor elétrico. Rio de Janeiro. 2008. 184p. Dissertação (Mestrado em Engenharia Elétrica). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

PENIN, C. A. de S. (2008) Combate, prevenção e otimização das perdas comerciais de energia elétrica. São Paulo. 2008. 214p. Tese (Doutorado) – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Energia e Automação Elétricas. São Paulo.

QUEIROGA, R. M. (2005) Uso de Técnicas de Data Mining para Detecção de Fraudes em Energia Elétrica. Dissertação de Mestrado, Universidade Federal do Espírito Santo. 2005.

REIS FILHO, J. (2006) Sistema Inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição de energia elétrica. Uberlândia, 2006. Dissertação (Mestrado) – Universidade Federal de Uberlândia. 2006.

REIS, C. Z. (2005) Eficácia de solução tecnológica para redução de furtos de energia em empresas distribuidoras: Estudo de Caso. Rio de Janeiro, 2005. Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro. 2005.

REZENDE, S. O.; PUGLIESE, J. B.; MELANDA, E. A.; PAULA, M. F. (2003) Mineração de dados. In: REZENDE, S. O. (ed.) *Sistemas inteligentes – fundamentos e aplicações*. Barueri: Editora Manole, 2003. Cap.12. p.307-335.

SANTOS, M. Y.; RAMOS, I. (2009) *Business intelligence – Tecnologia de informação na gestão de conhecimento*. FCA Editora de Informática, 2009.

SILVA, E. L. da; MENEZES, E. M. (2005) *Metodologia da pesquisa e elaboração de dissertação*. 4ªed. Florianópolis: UFSC. 138p. 2005.

SMITH, T. B. (2004) *Electricity theft*: a comparative analysis. Energy Policy, Guildford, v.32, n.18, p. 2067-2076, December 2004. doi:10.1016/S0301-4215(03)00182-4

STEADMAN, K. (2010) *Electricity theft in Jamaica*. New York: University of New York at Binghamton, v.1, p., 2010.

TAN, P. N; STEIBACH, M.; KUMAR, (2009) V. *Introdução ao data mining mineração de dados*. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009. 900p.

VIEIRALVES, E. de X. (2005) Proposta de uma metodologia para avaliação das perdas comerciais dos sistemas elétricos: o caso Manaus. Campinas, 2005. Dissertação (mestrado). Universidade Estadual de Campinas, Faculdade de Engenharia Mecânica. Campinas. 2005.

WITTEN, I. H.; FRANK, E. (2005) *Data Mining: practical machine learning tools and techniques*, 2nd Ed., San Francisco: Morgan Kaufmann, 2005.