Aplicação dos Algoritmos K-Means e FP-Growth em Perfis de Consumo para Identificação de Perdas Não-Técnicas de Energia em Consumidores Comerciais

Jorge Gustavo Sandoval Simão, Raimundo Celeste Ghizoni Teive

Universidade do Vale do Itajaí

jorge.sandoval@edu.univali.br, rteive@univali.br

Abstract. *One of the main problems faced by electric power concessionaires is the occurrence of power loss in distribution networks, caused by electricity theft. This study presents an analysis made with the application K-Means and FP-Growth algorithms in order to identify patterns of consumption that may be associated with fraud in the electricity distribution network. Clustering techniques and association rules were applied to a real database, considering commercial consumers, and the preliminary results show that the proposed methodology is promising in the identification of fraud in the distribution network.*

Resumo. *Um dos principais problemas que enfrentam as empresas concessionárias de energia elétrica é a ocorrência de perda de energia nas redes de distribuição, causados por fraudes de furtos de energia elétrica. Este estudo apresenta uma análise feita com a aplicação dos algoritmos K-Means e FP-Growth a fim de identificar padrões de consumo que possam estar associados a fraudes na rede de distribuição de energia elétrica. As técnicas de clusterização e regras de associação foram aplicadas a um banco de dados real, considerando consumidores comerciais, e os resultados preliminares demonstram que a metodologia proposta é promissora na identificação de fraudes na rede de distribuição.*

1. Introdução

O crescimento das concessionárias e o aumento dos prejuízos financeiros causados pelas perdas de energia elétrica, consequentes das anomalias nas redes de distribuição, têm incrementado a busca por novas tecnologias de detecção de fraudes [Queiroz *et al*. 2016].

As perdas de energia elétrica são divididas em duas categorias: as técnicas e as não-técnicas. As perdas não-técnicas (também conhecidas como perdas comerciais) são causadas pela manipulação ilegal de medidores ou por falhas nas instalações de consumo, enquanto as técnicas são causadas através de efeitos físicos (o efeito *Joule*, por exemplo), devidos à distribuição de energia [Guerrero *et al*. 2014]. As perdas não-técnicas, dada sua natureza, geram diminuição no registro do consumo de energia, ocasionando assim modificações nas topologias das curvas de carga.

Os consumidores comerciais representam uma parcela considerável da energia consumida. Em 2014, a energia consumida pela classe comercial foi de 3.946.188

MW/h, significando 16.9% do consumo total (representando um aumento em relação à 2013, que teve um consumo de 3.604.418 MW/h) [FIESC 2015]. Considerando que podem haver aumentos anuais do consumo de energia da classe comercial, o desvio passou a tornar-se cada vez mais significativo. Os consumos de energia elétrica residencial e comercial subiram mais do que o PIB (Produto Interno Bruto) com uma taxa de aumento de 3,6% [EPE 2016]. Dados estes registros, a representatividade das perdas não-técnicas sobre a classe comercial torna-se significativa.

Muitos métodos existentes para lidar com estas perdas de energia impõe um alto custo operacional, e requerem uso extensivo de recursos humanos [Nizar *et al*. 2006], pois elas representam prejuízo econômico, colocam em risco a segurança pública e criam impactos sociais. As concessionárias de energia elétrica fazem uso de uma metodologia conhecida como Inteligência de Negócios (*Business Intelligence*) com o objetivo de reduzir os problemas gerados pelas perdas comerciais [Immon 2005].

A própria Superintendência de Pesquisa e Desenvolvimento e Eficiência Energética (SPE) da ANEEL (Agência Nacional de Energia Elétrica) propôs, em 2013, a criação de um sistema de informações envolvendo os agentes do setor, que possibilitasse (dentre outros objetivos) a criação e uma base de dados consistente para a aplicação de técnicas de inteligência analítica e de mineração de dados [SPE 2013]. Considerando o a implementação desta tecnologia para as concessionárias de energia, as técnicas de Inteligência Computacional também podem ser utilizadas sobre os dados dos clientes de uma distribuidora.

As etapas da Descoberta de Conhecimento em Base de Dados (ou KDD - *Knowledge Database Discovery*) permitem o que os dados sejam selecionados, pré-processados e transformados para a etapa de Mineração de Dados. Esta etapa consiste na extração de informações previamente desconhecidas e potencialmente úteis de bases de dados, com o objetivo de descobrir perfis de consumidores e outros comportamentos que não poderiam ser identificados somente com análises manuais de especialistas. Ela pode utilizar diversos algoritmos para fins específicos tais como classificação, regressão, agrupamento, dentre outros [Queiroz *et al*. 2016].

Com o objetivo de detectar consumidores que possuam anomalias (definidas por variações incomuns nas curvas de cargas, ou características que não pertençam a um determinado consumidor em relação às regras de associação, representando certos padrões definidos) nos perfis de consumo de energia elétrica, este estudo propõe um método de detecção de fraudes baseado no agrupamento das curvas de carga através de sua tipologia, utilizando o algoritmo K-Means e o reconhecimento de características desses *clusters* através de regras de associação *FP-Growth*.

Como existem diferenças nos perfis de consumo dentre um mesmo consumidor em diferentes estações do ano, optou-se pela análise de apenas uma estação (verão), representados pelos dados de janeiro, fevereiro e março, e com um banco de dados de uma distribuidora de energia, contendo os dados de curvas de carga mensais e as informações de consumidores comerciais durante todo o ano de 2011. Com isso, identificam-se quais os consumidores que apresentam um perfil discrepante do seu *cluster* de acordo com as regras de associação do mesmo, o que pode vir a ser um indício de fraude.

2. Padrões de Curvas de Carga

Os valores de consumo de energia elétrica mensal de cada usuário são armazenado pelas concessionárias após a leitura dos valores de consumo nos medidores. Através da análise dos valores de energia consumida é possível classificar os consumidores por padrões de consumo. Esses padrões de consumo variam de acordo com as tipologias das edificações, situações sociais e financeiras de cada consumidor, situação econômica do país, clima e eventos em períodos do ano (férias escolares, viagem, entre outros) [Maria and Sattler 2000].

Independentemente do tipo de consumidor, o consumo de energia elétrica possui um comportamento sazonal e cíclico que pode ser demonstrado através das tipologias de curvas de carga quando é feita a análise de consumo dos usuários. O comportamento regular desta curva é chamado de padrão ou perfil de consumo. A curva de carga típica descreve os valores horários do consumo energético em uma base diária, e é associada a uma certa categoria de consumidor, em condições específicas de operação [Gavrilas *et al*. 2010].

As curvas de carga podem ser definidas como residenciais, industriais, comerciais ou de serviços, para estações quentes e frias, em dias de semana ou fins de semana, e são obtidas agrupando os perfis de carga de acordo com sua similaridade [Azad *et al*. 2014]. Nos dados utilizados para esta pesquisa, os perfis de consumo diário dos consumidores são registrados por hora, com dados coletados em um intervalo de cinco minutos.

Estes dados também foram divididos sazonalmente, pois há mudança na tipologia das curvas de um mesmo consumidor em estações diferentes. O perfil consumo do usuário mantém um padrão similar durante todos os dias, pois as atividades comerciais dos consumidores utilizados neste estudo são regulares na sua execução (ou seja, possuem intervalos definidos de início e fim, com interrupções sempre nos mesmos horários). Através da identificação de características comuns entre os *clusters* pelas regras de associação e através da análise das curvas de consumo de energia elétrica, é possível detectar padrões de consumo que podem ser usados para classificar consumidores e detectar anomalias nas redes de distribuição de energia elétrica [Queiroz *et al*. 2016].

Além dos dados de curvas de carga, foi utilizado um banco de dados com informações de consumidores, envolvendo os seguintes aspectos:

- Divisão e classe CNAE
- Fator de carga e porte da empresa
- Horários e turnos de funcionamento
- Demanda continua e Consumo médio mensal

Para a análise do comportamento de consumo utiliza-se o fator de carga (FC). Segundo a resolução normativa nº 414 de 9 de setembro de 2010 da ANEEL, o fator de carga é definido como sendo a razão entre a demanda média e a demanda máxima da unidade consumidora ocorridas no mesmo intervalo de tempo especificado [ANEEL 2010]. O fator de carga é um índice adimensional que varia de 0 a 1, e quanto mais

próximo de 1, maior a eficiência energética da instalação. Para isso, a diferença entre o consumo medido (numerador) e a demanda máxima registrada deve ser a menor possível. Este resultado próximo a 1 indica que as demandas instantâneas ao longo do dia são próximas à demanda máxima [Barros *et al*. 2010].

3. Pré-Processamento e Discretização dos Dados

Para identificar os comportamentos das curvas de carga, os dados da distribuidora de energia foram pré-processados de acordo com a sequência de etapas do KDD (*Knowledge Database Discovery*), que consiste, segundo Han e Kamber (2012), em: limpeza, integração, seleção, transformação, mineração e avaliação dos dados.

Inicialmente os dados (que em seu estado bruto estavam separados em arquivos de texto de curvas de carga e informações de consumidores) foram importados para uma base de dados PostgreSQL única e interrelacionados através do identificador do consumidor. Posteriormente, os dados de curvas de carga e informações de consumidores foram integrados através do seu identificador (que é comum entre eles, podendo ser utilizado como chave primária) e determinados grupos selecionados para serem submetidos aos algoritmos *K-Means* e *FP*-Growth.

Durante a fase de transformação, os dados das curvas de carga foram normalizados em valores entre 0 e 1, para evitar a discrepância causada pela diferença de valores de consumo entre consumidores do mesmo perfil, dado que o interesse está no padrão de consumo e não no montante consumido. Foi calculado também o desvio padrão para as tipologias de todas as curvas de cargas das amostras, pré-definindo então quais os possíveis valores máximo e mínimo para uma curva de carga daquele *cluster.* Os dados de fator de carga e número de funcionários são então transformados, sendo discretizados de acordo com a Tabela 1.

Tabela 1. Dados de números de funcionários e curvas de cargas discretizados para utilização com as regras de associação.

Parâmetro	Nomenclatura
Fator de carga <= 0.3	FC_BAIXO
Fator de carga > 0.3 e <= 0.55	FC_MEDIO_BAIXO
Fator de carga > 0.55 e <= 0.8	FC_MEDIO_ALTO
Fator de carga acima > 0.8	FC_ALTO
Número de funcionários <= 625	FUNC_BAIXO
Número de funcionários > 625 e <= 1250	FUNC_MEDIO_BAIXO
Número de funcionários > 1250 e <= 1875	FUNC_MEDIO_ALTO
Número de funcionários > 1875	FUNC_ALTO

Os dados foram categorizados para que pudessem ser utilizados pelas regras de associação. As categorias foram selecionadas de acordo com o número médio de funcionários em cada empresa, e pela definição dos valores de fator de carga, que podem variar de 0 a 1. Uma vez que os dados tenham sido submetidos ao pré-processamento, eles estão aptos a serem submetidos aos algoritmos de clusterização e regras de associação.

4. Número Ideal de Clusters

Técnicas baseadas em agrupamentos de elementos criam níveis de particionamento dos objetos de dados. O k-means define o elemento como se fosse um centróide (elemento central), o que normalmente é a média de um grupo de pontos e é tipicamente aplicado a objetos em um espaço contínuo n-dimensional [Pang-Ning *et al*. 2005]. Ele toma aleatoriamente *k* pontos de dados (dados numéricos) como sendo os centróides dos *clusters*. Em seguida cada ponto (ou registro da base de dados) é atribuído ao *cluster* cuja distância deste ponto em relação ao centróide de cada *cluster* é a menor de todas as distâncias calculadas. Um novo centróide é computado pela média dos pontos, caracterizando a configuração do *cluster* para a interação seguinte. O processo termina quando os centróides param de se modificar, ou após um número limitado de iterações especificado pelo usuário [Goldschmidt and Passos 2005].

Nos dados das curvas de carga, o K-Means é aplicado na etapa de Mineração de Dados. Para permitir o agrupamento das curvas de cargas por características, utilizou-se como parâmetro do algoritmo a distância euclidiana e a distância média entre os centróides. Uma vez que há *clusters* com diversos valores de *k*, é possível determinar um bom número de *clusters* usando uma medida que é independente de *k*, tal como o coeficiente de silhueta média [Frahling and Sohler 2005]. Este coeficiente é utilizado para estudar a distância de separação entre os *clusters* resultantes. O gráfico apresentado na Figura 1 descreve uma medida de quão perto cada ponto em um *clusters* está próximo dos pontos na sua vizinhança, e provê um meio de medir o número de *clusters* visualmente. Essa medida possui um alcance que varia de -1 a 1.

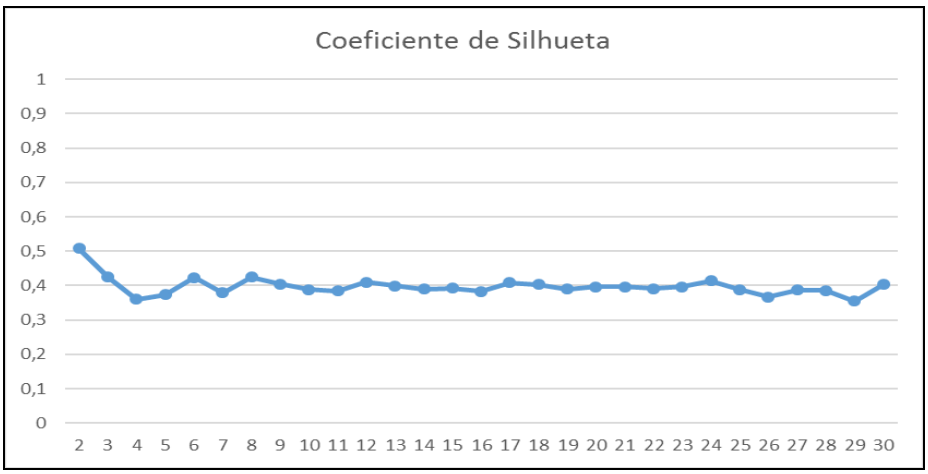

Figura 1. Coeficiente de silhueta de dois a trinta clusters para os dados de curvas de carga.

Coeficientes de silhueta próximos a 1 indicam que a amostra está longe da vizinhança. Um valor próximo a 0 indica que a amostra está dentro ou muito próxima do limite de decisão entre os *clusters* e valores negativos indicam que essas amostras foram associadas ao *cluster* errado. A Figura 1 indica os valores dos coeficientes de silhueta (eixo *y*) em relação ao número de clusters (eixo *x*, que variou de 2 a 30 para esta pesquisa). Quando os dados são divididos em apenas dois *clusters*, o coeficiente é de 0.5086, decrescendo com a divisão em quatro (0.3607) e estabilizando entre dez (0.3888) e vinte e três (0.3971) *clusters*.

Sendo assim, a partir da análise da Figura 1, é possível dizer que no ponto onde o gráfico se encontra estável há uma similaridade de divisão dos agrupamentos, tendo sido selecionado para esta pesquisa a divisão em quatorze *clusters*, um número que não granulariza muito a quantidade de características (e consequentemente a tipologia das curvas de carga), e é próximo da média da área estável (de 10 a 23). Divisões de *clusters* acima de 14 consideram variações de tipologia de forma muito detalhada, criando múltiplas categorias para pequenas modificações de valores, dificultando assim a sua classificação por similaridades. O número de *clusters* definidos para os dados de curvas de carga desta pesquisa fixaram-se em 14, sendo o parâmetro de entrada a média do valor horário para cada hora do dia (totalizando 24 pontos por dia), durante toda a estação do verão.

5. O Algoritmo FP-Growth na Tarefa de Associação das Curvas de Cargas e das Informações de Consumidores

Para a fase das regras de associação, foi necessário um algoritmo que calculasse o *Frequent Item Set* (Amostragem de Itens Frequentes) para identificar a incidência de combinações. Porém, dado o volume de itens, a utilização do algoritmo *Apriori* apresentou um desempenho inadequado (levando dias para processar todas as combinações existentes). Então para que as regras de associação fossem realizadas, optou-se pelo algoritmo FP-Growth (*Frequent Pattern-Growth*).

Na abordagem do algoritmo *Apriori*, se qualquer padrão de comprimento k não é frequente na base de dados, seu comprimento $(k + 1)$ não será frequente. Espera-se com isso que, através de um processo iterativo, gerar o conjunto de padrões de candidatos de comprimento $(k + 1)$, e verificar suas frequências de ocorrência na base de dados. Porém, a geração de um conjunto candidato é ainda dispendiosa, principalmente quando há um número muito grande ou longo de padrões. Também é dispendioso percorrer repetidas vezes a base de dados para verificar e testar todos os conjuntos de candidatos e seus padrões correspondentes [César *et al*. 2015].

O algoritmo *FP-Growth* utiliza uma abordagem que não faz uso da geração de candidatos e do paradigma do *Apriori* de gerar e testar cada um deles. Ao invés disso, codifica o conjunto de dados em uma estrutura denominada *Frequent Pattern Tree* (*FP-Tree*) e extrai os conjuntos de itens frequentes diretamente desta estrutura. Isso possibilita uma melhora eficiência na geração das regras de associação, pois evita constantes acessos na base de dados. Ele necessita de, no mínimo, dois parâmetros de entrada [Pang-Ning *et al*. 2005]:

a) **Suporte:** é a métrica utilizada pelo algoritmo para encontrar todos os *N itemsets*. O suporte de uma regra de associação $X, A \Rightarrow B$, é a porcentagem das transações que contém $A \cup B$ em relação ao número total de transações analisadas.

b) **Confiança:** calcula a força da regra. Assim, sendo C a confiança de uma regra de associação $A \Rightarrow B$, C é, na verdade, a porcentagem das transações que contém $A \cup B$ em relação à todas as transações que contém A.

O método consiste no desenvolvimento de uma estratégia baseada na técnica de dividir para conquistar, onde o problema é fracionado em subdivisões, considerando cada um dos itens da sua tabela de cabeçalhos. O primeiro item a ser processado é o

último, o segundo o penúltimo e assim sucessivamente até o processamento de todos os itens. Para cada *itemset* presente na tabela de cabeçalhos, bases de padrões condicionais são geradas, as quais são utilizadas para a construção das *FP-tree* condicionais, que relacionam os caminhos frequentes que se conectam aos nós correspondentes ao *itemset* em questão [César *et al*. 2015]. Uma vez criadas, as *FP-tree* condicionais são utilizadas para encontras os padrões frequentes que apresentam o *itemset* como sufixo [Han *et al*. 2000].

Este algoritmo foi aplicado nas informações de consumidores cedidos pela distribuidora de energia, divididos de acordo com o *cluster* ao qual eles pertencem, visando identificar as regras mais proeminentes destes consumidores (que por sua vez definem as características comuns do *cluster*). Foram utilizados como parâmetros mínimos 35% de suporte e 70% de confiança e aproveitadas apenas regras que possuíam mais de uma incidência em relação à sua busca. Para este estudo foram selecionados os *clusters* 0 e 1 (que possuíam o maior número de consumidores), visando identificar características dos mesmos com a maior precisão possível. As selecionadas para a definição do perfil do *cluster* são apresentadas na Tabela 2.

Tabela 2. Regras de Associação selecionadas para os clusters 0 e 1 geradas pelo algoritmo FP-GROWTH

Cluster	Premissa	Conclusão	Suporte	Confiança
0	fc = FC_MEDIO_BAIXO tarifa = verde dem_continua = BAIXA divisao = div52	in_func = ini_manha fim_func = ini_noite	37.7%	100%
1	tp_tarifa = verde fc = FC_MEDIO_ALTO in_func = flat fim_func = flat de_turnos = T1111 dem_continua = MEDIA	tp_divisao = div85	39.3%	100%

As regras de associação selecionadas para os *clusters* 0 e 1, apresentadas na Tabela 2 indicam premissas, conclusões, suporte e confiança que sugerem uma similaridade dentre os consumidores daquele *cluster*. Algumas similaridades que podem-se identificar de acordo com as regras é que no *cluster* 0, os consumidores que iniciam seu horário de funcionamento pela manhã, possuem um fator de carga de médio para baixo, a tarifa verde, uma demanda contínua baixa e um consumo médio baixo possuem um número baixo de funcionários (regra 1), assim como no *cluster* 1, aqueles que possuem tarifa verde, fator de carga médio alto, fazem parte da divisão 85 e têm uma demanda contínua média trabalham todos os turnos, ininterruptamente (regra 5).

De acordo com estas regras, identifica-se então um padrão de consumo e características entre os usuários daquele *cluster*. Usuários que diferenciam-se desse padrão, mesmo fazendo parte do mesmo perfil de consumo, podem indicar algum tipo de anomalia no consumo.

6. Resultados

Para obterem-se os resultados, foram utilizados como ambiente de testes os softwares Matlab e Rapidminer, aonde todos os *clusters* apresentaram perfis de consumo próprios,

similares entre os consumidores que fazem parte dele. O eixo *x* representa a hora de registro do consumo, enquanto o eixo *y* o consumo em kw/h, assim como as linhas pontilhadas representam os desvios padrão superior e inferior. No *cluster* 0 as atividades iniciam-se às 06:00hs e encerram-se às 18:00hs, diferentemente do *cluster* 1, que já mantém uma atividade constante de consumo, que aumento às 06:00hs e sofre variações às 17:00hs, de acordo com o comportamento apresentado na Figura 2.

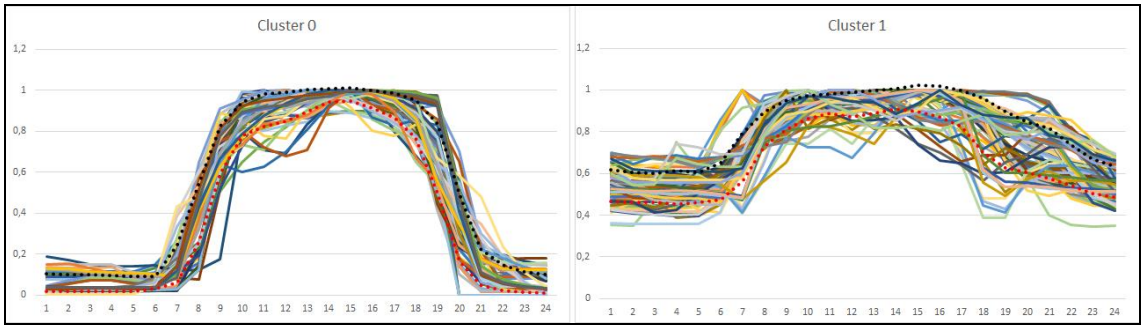

Figura 2. Tipologia de curvas de carga dos clusters 0 e 1 para todos os consumidores.

Os consumidores C1 e C2 possuem todas as características inerentes às premissas e conclusões das regras de associação encontradas (similaridade em fator de carga, demanda média, consumo médio mensal, CNAE e etc) indicando uma possível pertinência aos *clusters* 0 e 1, respectivamente. Porém, quando observam-se suas curvas de carga, verificam-se tipologias diferentes das previstas para os *clusters* 0 e 1, conforme é mostrado na Figura 3.

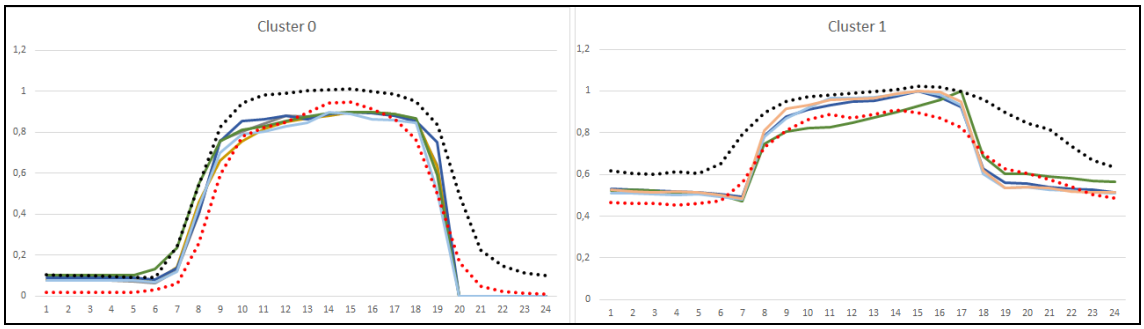

Figura 3. Tipologia de curvas de carga dos clusters 0 e 1 especificamente para os consumidores C1 e C2.

Apesar de possuírem as mesmas características que os categorizam dentro dos seus respectivos *clusters*, o consumidor C1 apresenta uma queda abrupta de consumo das 20:00h em diante, enquanto o consumidor C2 apresenta um consumo praticamente constante das 01:00h ás 07:00h, e novamente após ás 18:00h, em ambos os casos fora do desvio padrão. Este comportamento pode indicar, para o consumidor C1, que há um desligamento completo do sistema de medição das 21:00h ás 24:00h (com seu religamento após esse horário), enquanto que para o consumidor C2 apenas determinados equipamentos podem estar desligados do medidor quando fora do horário comercial (das 19:00h ás 07:00h), já que o C2 é classificado como divisão 85 do CNAE (Atividades de Atendimento Hospitalar) [IBGE 2016], e não possui horário para início e fim de funcionamento. Como ficou evidenciado um possível comportamento

fraudulento, caberia então às concessionárias fazer uma inspeção local para confirmar ou não a identificação do sistema inteligente.

7. Conclusões

Considerando-se a quantidade de dados armazenados diariamente com os registros das leituras mensais das distribuidoras de energia elétrica, e também nas informações dos consumidores, é possível afirmar que a aplicação de inteligência computacional para transformar tais dados em conhecimento pode auxiliar na identificação de anomalias nos perfis de consumo dos usuários desta distribuidora.

Dentre os resultados obtidos, permitiram-se então a divisão de consumidores pelo seu perfil de consumo (representado pela tipologia da curva de carga) e o reconhecimento das características de cada *cluster* através das regras de associação, possibilitando assim identificar quais elementos definem e diferenciam um consumidor que pertence àquele *cluster*. Sendo o principal objetivo deste estudo a identificação de padrões de possíveis fraudes através da análise do seu agrupamento por perfis, e das comparações de suas características com as do cluster ao qual pertencem, pode-se afirmar que é possível identificar padrões discrepantes que podem ser ocasionados pela perda não-técnica de energia elétrica.

Resultados obtidos através de algoritmos incentivam sua utilização na identificação de fraudes, o que pode reduzir e direcionar as fiscalizações manuais que são realizadas pelas concessionárias de energia a fim de encontrar possíveis fraudes nas redes de distribuição de energia elétrica.

Referências

ANEEL. "Resolução Normativa Nº 414, de 9 de Setembro de 2010". Disponível: http://www2.aneel.gov.br/cedoc/ren2010414.pdf, Dezembro/2016.

Azad, S. A., Ali, A. B. M. S. and Wolfs, P. "Identification of Typical Load Profiles using K-means Clustering Algorithm". In: Asia-Pacific World Congress on Computer Science and Engineering. IEEE, 2014.

Barros, B. F. De, Borelli, R. and Gedra, R. L.. Gerenciamento de Energia: Ações Administrativas e Técnicas de Uso Adequado da Energia Elétrica. São Paulo: Érica Ltda, 2010.

César, J., Nandi, B., Pereira, R. M. and Felippe, G.. "O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine". In: VII Congresso Sul Brasileiro de Computação. SULCOMP, 2015.

EPE. "Consumo de Energia Elétrica por Classe (Regiões e Subregiões)". Disponível: http://www.epe.gov.br/mercado/Paginas/Consumomensaldeenergiaelétricaporclasse(regiõeses, Maio/2016.

FIESC. "Santa Catarina em Dados 2015". Disponível: http://fiesc.com.br/sites/default/files/medias/sc_em_dados_site_correto.pdf, Dezembro/2016.

Frahling, G. and Sohler, C.. "A Fast K-Means Implementation Using Coresets ∗", International Journal of Computational Geometry & Applications, v. 18, n. 6, p. 605–625, 2005.

Gavrilas, M., Gavrilas, G. and Sfintes, C. V.. "Application of Honey Bee Mating Optimization Algorithm to Load Profile Clustering". In: 2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. IEEE, 2010.

Goldschmidt, R. and Passos, E.. Data Mining: Um Guia Prático. Rio de Janeiro, RJ: Elsevier, 2005.

Guerrero, J. I., León, C., Monedero, I., Biscarri, F. and Biscarri, J.. "Improving Knowledge-Based Systems with Statistical Techniques, Text Mining, and Neural Networks for Non-Technical Loss Detection". Knowledge-Based Systems, v. 71, p. 376–388, 2014.

Han, J. and Kamber, M.. Data Mining: Concepts and Techniques. 3. ed. Filadelfia: Elsevier, 2012.

Han, J., Pei, J. and Yin, Y.."Mining Frequent Patterns without Candidate Generation". In: 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD, 2000.

IBGE. "CONCLA - Comissão Nacional de Classificação". Disponível: http://cnae.ibge.gov.br/?option=com_cnae&view=atividades&Itemid=6160&tipo=cnae&chave=52&versao_classe=3.0.1&versao_subclasse=4.1.1, Dezembro/2016.

Immon, W. H.. Building the Data Warehouse: Getting Started. 4. ed. Indianapolis, Wiley Publishing, 2005.

Maria, A. H. D. and Sattler, M. A.. Padrões de Consumo de Energia Elétrica em Diferentes Tipologias de Edificações Residenciais, em Porto Alegre. 2000. 146 f. Dissertação (Mestrado) - Curso de Engenharia Civil, Programa de Pós-graduação em Engenharia Civil, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

Nizar, A. H., Dong, Z. Y., Jalaluddin, M. and Raffles, M. J.. "Load Profiling Method in Detecting non-Technical Loss Activities in a Power Utility". In: 2006 IEEE International Power and Energy Conference. IEEE, 2006.

Pang-Ning, T., Vipin, K. and Steinbach, M.. Introduction to Data Mining. Boston: Pearson Addison Wesley, 2005.

Queiroz, A. de S., Franco, E. M. C. and López, G. P.. "Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica Utilizando Técnicas de Inteligência Computacional". In: VI Simpósio Brasileiro de Sistemas Elétricos. SBSE, 2016

SPE. "SIASE – Sistema de Inteligência Analítica do Setor Elétrico". http://www2.aneel.gov.br/arquivos/PDF/PD Estratégico 018-2013_SIASE.pdf, Dezembro/2016.