Electrical Power and Energy Systems 34 (2012) 90–98

Contents lists available at SciVerse ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees

Iñigo Monedero [a,*], Félix Biscarri [a], Carlos León [a], Juan I. Guerrero [a], Jesús Biscarri [b], Rocío Millán [b]

[a] Electronic Technology Department, University of Seville, Spain
[b] Endesa Distribution Company, Measure & Control Department, Spain

ARTICLE INFO

Article history:
Received 2 March 2011
Received in revised form 1 September 2011
Accepted 16 September 2011
Available online 1 November 2011

Keywords:
Non-technical loss
Data mining
Pearson correlation coefficient
Decision tree
Bayesian network

ABSTRACT

For the electrical sector, minimizing non-technical losses is a very important task because it has a high impact in the company profits. Thus, this paper describes some new advances for the detection of non-technical losses in the customers of one of the most important power utilities of Spain and Latin America: Endesa Company. The study is within the framework of the MIDAS project that is being developed at the Electronic Technology Department of the University of Seville with the funding of this company. The advances presented in this article have an objective of detecting customers with anomalous drops in their consumed energy (the most-frequent symptom of a non-technical loss in a customer) by means of a windowed analysis with the use of the Pearson coefficient. On the other hand, besides Bayesian networks, decision trees have been used for detecting other types of patterns of non-technical loss. The algorithms have been tested with real customers of the database of Endesa Company. Currently, the system is in operation.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

A non-technical loss (NTL) is defined as any consumed energy or service which is not billed because of a measurement equipment failure or an ill-intentioned and fraudulent manipulation of the said equipment. For the electrical distribution business, detecting NTLs is a very important task; since, for instance, in Spain it is estimated that the percentage of fraud in terms of energy with respect to the total NTLs about 35–45%. Although in the literature there are many works and researches on fraud and NTL detection in other fields [1–9], there is not much research about NTL detection in power utilities [10–15] in spite of the percentage of NTLs is high in this field. Besides, these works are basically theoretical and limited to the use of few types of detection techniques (rough sets, support vector machines and wavelet transform).

Thus, the current methodology adopted by the electrical companies in the detection of NTLs is basically of two kinds. The first one is based on making in situ inspections of some users (chosen after a consumption study) from a previously chosen zone. The second one is based on the study of the users which have null consumption during a certain period. The main problem of the first alternative is it requires a large number of inspectors and, therefore, involves a high cost. The problem with the second option is the possibility of detecting users only with null consumption (these are only the clearest cases of non-technical losses) and not those customers with non-null consumption but quite lower than the consumption that they might have. Nowadays, data mining techniques [16,17] are applied to multiple fields and power utility is an industry in which it has met with success recently [18–22].

The work is within the framework of MIDAS project which is being developed at the Electronic Technology Department of the University of Seville with the funding of the electrical company. We have presented the results of the MIDAS project using a detection process based on extraction rules and clustering techniques [23,24] as well as preliminary versions of the algorithms for the detection of drops [25].

This article describes new advances in the data mining process applied to detection of NTLs in power utilities. Besides, it includes a complete process of NTL detections from the databases of the Endesa Company. Thus, other additional lines have been developed in order to detect other types of NTLs. One of the ideas of these methods is to identify patterns of drastic drop of consumption. It is because it is known that the main symptom of an NTL is a drop in the billed energy of the customers. Thus, with this purpose, these methods are based on the use of the Pearson coefficient [26,27] on the evolution of the consumption of the customer. Besides, in order to carry out the detection of NTLs that include other type of consumption pattern, a model based on a Bayesian network [18] and a decision tree [18] has been developed.

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. Bayesian networks are applied in cases of uncertainty when we know certain probabil-

* Corresponding author. Address: Escuela Politécnica Superior, C/Virgen de África 7, 41011 Sevilla, Spain.
E-mail address: imonedero@us.es (I. Monedero).

0142-0615/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijepes.2011.09.009

ities and are looking for unknown probabilities given specific conditions. Some applications of Bayesian networks are: churn prevention [28], generation of diagnostic in medicine [29], pattern recognition in vision [30] and fault diagnosis [31] as well as forecasting [32] in power systems. Besides, these networks have also been used to detect anomaly and frauds in disciplines other than power utilities such as credit card or telecommunication networks [2,33,34]. On the other hand, it is possible to find some works that suggest the use of decision trees in power systems [35,36] and to detect some types of frauds [7,37]. However, besides our studies [23,24], as we said, not much research is done on detection of NTLs and frauds in power utilities [12–17] and nothing about the detection of consumption drops or development of models with the use of Bayesian networks.

In order to carry out the data mining process (including the algorithms as well as the models of Bayesian network and decision tree), we used a powerful software called IBM SPSS Modeler 14 used extensively in data mining. This software provide a quick access to the databases and many libraries for the generation of models such as: clustering processes, decision trees, neural networks and Bayesian networks.

The article is structured as follows: Section 2 describes the sample set which has been used to develop the algorithms and select the customers to be inspected by the company. Sections 3 and 4 describe the developed models. Finally, Section 5 contains the results as well as the conclusions from the study.

2. Data preparation

First of all, we selected a sample set made by customers with the, called by the company, ratings 3.0.2 and 4.0. These types of rate or contract are used by the company to design for customers with a high contracted power (which, in great majority, are belonging to contracts with companies). This sample set covered for the most important region of the Endesa Company: Catalonia (Spain). On the other hand we included those customers with highest consumption because this was interesting because each detected NTL could mean a lot of lost energy for the company. An analysis period of two years was adjusted. This is a time enough to see a sufficiently detailed evolution of the consumption of the customer and, also, not too long a period to register along the contract the possible changes of type of business or the changes in the consumption habits of the client. From these customers, all the information of consumption and type of contract for each customer: reading values of the measurements equipment, bills from the last two years, amount of power contracted and the type of customer (private client or the kind of business of the contract), address, type of rate, etc. was collected. These data were condensed and tabulated. Thus, the system would have details on the types of customer as well as the evolution of their consumption in the last two years.

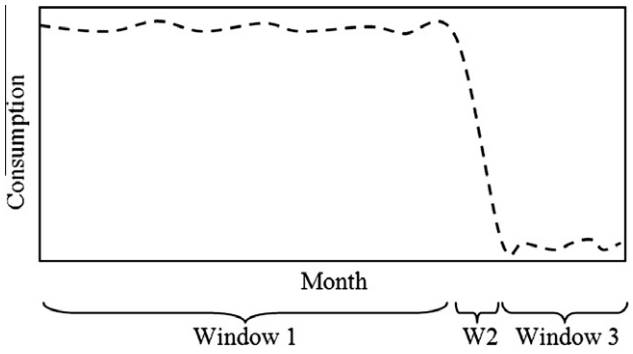

Fig. 1. Consumption patterns searched with first algorithm.

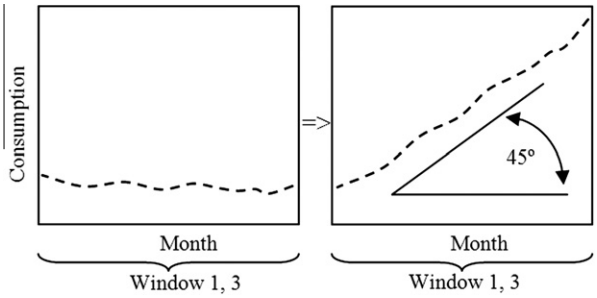

Fig. 2. Offset of the consumption of the windows 1 and 3.

As in our work described in the paper [25], a filling up of missing values of the consumption read was performed and a filtering of the customers:

- With less than 1000 KWs consumed in the two years.
- With less number of reading values from the measurements equipment (under 10 from the 24 months of the analysis.
- With no reading values in the last four months.

After the selection and filtering of the sample, a set of 38,575 customers was obtained for the analysis. The time interval that the data covered was from July 2008 to June 2010.

3. Models based on Pearson coefficient

The drastic drops of the consumption can be due to a real slope of the consumptions of the customers (e.g. due to a change of type of contract or by a different use of the consumed energy). But, in turn, these slopes can be due to failures in the measurement equipment or voluntary alterations of this equipment (both cases generate NTLs to the company and therefore a loss of money for it). We could verify this fact with a set of customers with NTLs previously registered by the company in its inspections, where this type of drop was clearly visible.

There were two problems in detecting this type of customers in Endesa Company:

- They detect the drops only when the drops reach to null consumption (and on a drastic manner).
- The inspections of the company detect these drops when the drop has been prolonged over a long time.

Thus, with the models described in this section these problems were solved by detecting other type of drops and doing it in a short interval of months.

In statistics, the Pearson correlation coefficient (r) is a measure of how well a linear equation describes the relation between two variables X and Y measured on the same object or organism. The result of the calculus of this coefficient is a numeric value that runs from -1 to 1. This coefficient (r) is calculated by means of the following equation:

$$-1 \leq r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{\sum_{t=1}^{n}(X_t - \bar{X}) * (Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2} * \sqrt{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}} \leq +1 \qquad (1)$$

where $Cov(X, Y)$ is the covariance between X and Y. $S_X S_Y$ is the product of the standard deviations for X and Y.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات