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.

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.

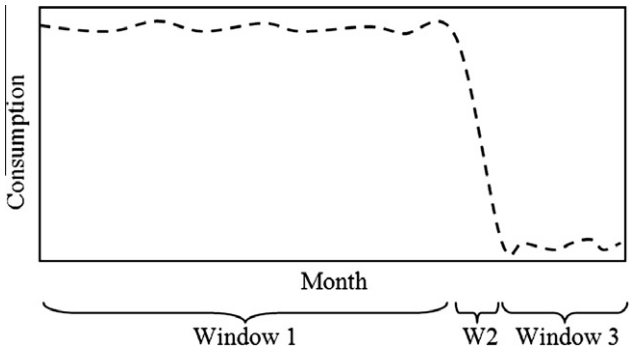

Fig. 1. Consumption patterns searched with first algorithm.

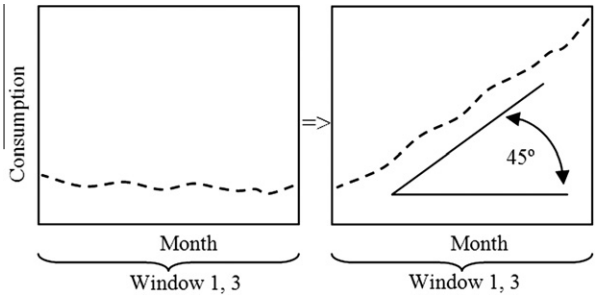

Fig. 2. Offset of the consumption of the windows 1 and 3.

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.

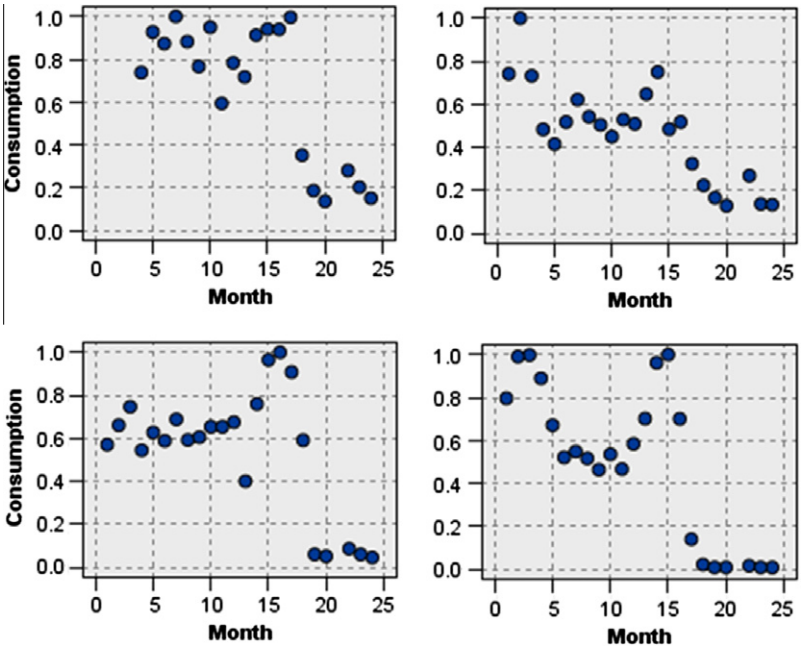

Fig. 3. Examples of customer detected with first algorithm.

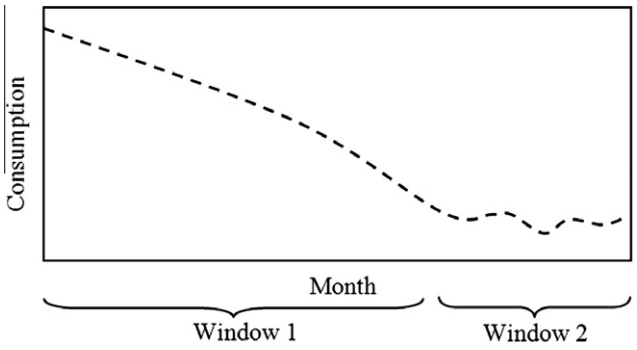

Fig. 4. Consumption patterns searched with second algorithm.

A value of 1 indicates that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and with Y increasing with X. A score of -1 shows that all data points lie on a single line but Y increases as X decreases. At last, a value of 0 shows that a linear model is inappropriate – there is no linear relationship between the variables.

Paper [25] presented two preliminary versions of algorithms for the detection of drops. In this work, two novel algorithms were described:

- An algorithm for the detection of a type of pattern of drastic drop very usual in the customers with NTLs.
- A more effective and robust algorithm for the detection of a type of pattern which was detected with worse adjustment by means of an algorithm of the work [25].

3.1. Algorithm for the detection of drastic drop with subsequent stabilization

An algorithm to identify the customers with abrupt drops and subsequent stabilization with low values was implemented. It detected them with a margin of only 6 months from the drop; this was very important to avoid that the NTL was prolonged in time. It was interesting because this type of pattern was typical when it was carried out with a gross manipulation of the measurement equipment.

In order to identify these patterns an analysis based on three windows was used:

- *Window 1*: First 16 consumption values of the customer (values 1–16 from the 24 values).
- *Window 2*: Intermediate 2 consumption values of the customer (values 17 and 18 from the 24 values).
- *Window 3*: Final 6 consumption values of the customer (values 19–24 from the 24 values).

We chose 3 windows in order to distinguish the three different types of behavior in the consumption of the customer. The searched pattern as well as the taken windows can be seen in Fig. 1. First of all, the algorithm detected a high steady consumption in window 1, with a subsequent sharp drop in window 2, and finally a low steady consumption in window 3. The length taken for the first window was quite long (16 months) to ensure that that was the typical consumption range of the customer in one year. On the other hand, the length of the third window (6 months) was a sufficient period of time to detect the important change of consumption range of the customer, but not too long so that the NTL was widespread long before being detected.

For the first and third window the Pearson coefficient was calculated for the values of consumption by moving the consumption 45° in the time axis of the months of both windows. This procedure is graphically described in Fig. 2.

Afterwards, the algorithm searched for values near 1 to the Pearson coefficient in windows 1 and 3; thus, steady consumption values in both sections were obtained.

Thus, we applied Eq. (2) to windows 1 and 3 and we extracted the following rule in order to detect those customers with this type of patterns:

$$R\_w1 > 0.8 \text{ and } R\_w3 > 0.8 \text{ and } Average\_w1 > 4 * Average\_w3 \quad (2)$$

where R_w1 is the Pearson coefficient for window 1 (with an offset of 45°); R_w3 is the Pearson coefficient for the window 3 (with an offset of 45°); $Average\_w1$ and $Average\_w3$ are the averages of the

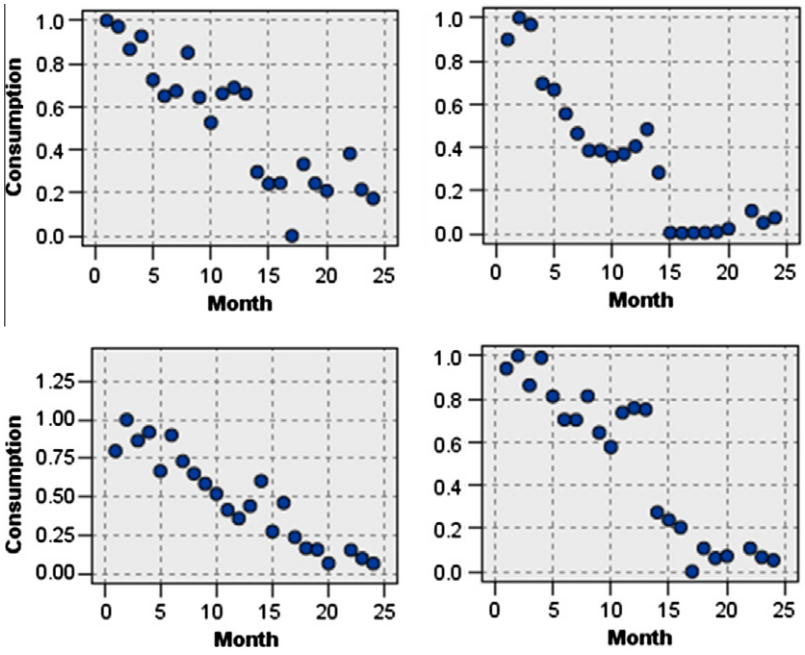

Fig. 5. Real examples of customers detected with second algorith.

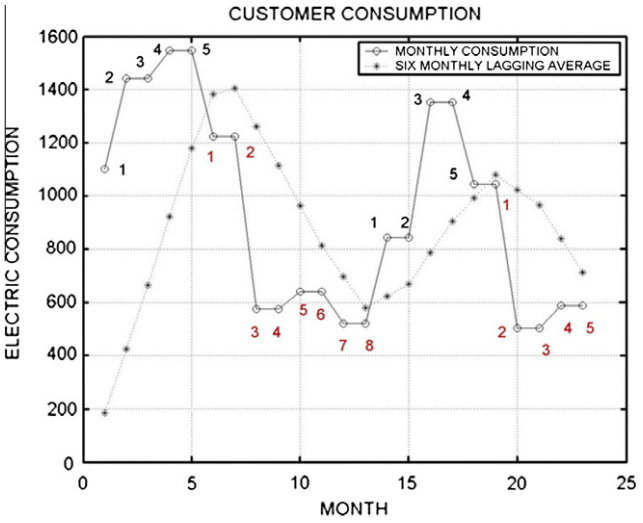

Fig. 6. Example of a customer with 3 streaks.

consumption for the first and second windows, respectively (without offset).

The use of those requirements for the averages of the windows 1 and 3 implied a decrease in consumption for the second window. Thus, applying this rule on the sample set it obtained a group of 23 customers to be inspected by Endesa. We can observe the pattern of four of these customers in Fig. 3.

3.2. Algorithm for the detection of progressive drop with subsequent stabilization

The objective of this second algorithm was to detect customers whose consumption became steady with low values after a progressive drop and therefore with a consumption pattern as shown in Fig. 4.

In order to get this objective, an algorithm by means of the analysis of the consumption of the customer in two windows was designed; the first windows with the consumptions of the first 18 months (which was a very long period of drop, more than a year), and the second one with the last 6 months (which was sufficient time to detect the new range of consumption of the customer). We configured 6 months for the second window because it was enough to see a possible anomaly in the consumption of the customer and, at the same time, to detect it quickly (with a shorter duration). The consumption pattern which was searched with this algorithm was the one shown in Fig. 4.

Thus, for this algorithm, the Pearson coefficient for the first window was calculated applying Eq. (1) for the 18 first consumption values.

For the second window, again the Pearson coefficient was calculated for the values of consumption of this window with an offset of 45° (described in Fig. 2).

After applying the Pearson coefficient to this moved consumption, the algorithm searched values near 1 (this corresponds to a steady value in the original consumption which was the objective).

$$R\_w1 < -0.75 \text{ and } R\_w2 > 0.75 \text{ and } Average\_w1 > 2 * Average\_w2 \quad (3)$$

where R_w1 is the Pearson coefficient for window 1; R_w2 is the Pearson coefficient for window 2 (with the moved consumption); *Average*_w1 and Average_*w*2 are the average of the consumptions for the first and second windows, respectively (without offset in the second window).

Applying this rule on the sample set, the algorithm obtained 18 customers to be inspected. Through a display of their consumption, we could verify that these clients had a pattern as shown in Fig. 4. Fig. 5 shows the patterns for four of these 18 customers.

In addition to greater accuracy in the detection of the searched pattern, this algorithm compared to the one presented in paper [25], had the advantage of detecting more quickly (in only 6 months) the possible NTLs.

4. Models based on Bayesian networks and decision trees

The objective of the Pearson coefficient was to search drops in consumption (which is often a characteristic indicative of losses

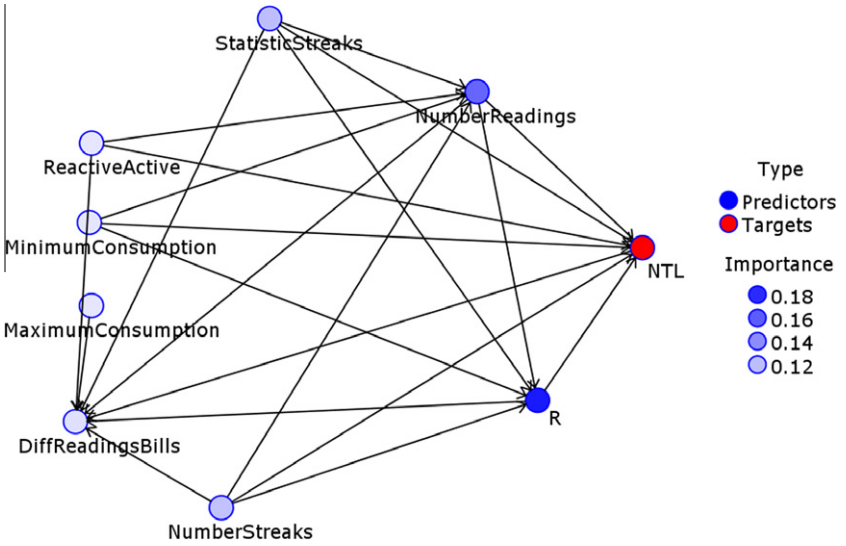

Fig. 7. Bayesian network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Results for output field NTL

Comparing $B-NTL with NTL

Correct	36.756	95,28%
Wrong	1.819	4,72%
Total	38.575	

Coincidence Matrix for $B-NTL (rows show actuals)

	0	1
0	36.612	1.357
1	462	144

Fig. 8. Detailed results of the Bayesian network.

techniques). With this approach the previously-described algorithms detected NTLs with drops in consumption but not with other patterns (not only drops in consumption is the consequence of NTLs). Therefore, we developed a complementary technique to search patterns with different shapes.

One of the important advantages which can find in Bayesian networks is that the generated diagrams can readily be interpreted by humans; therefore, it could provide results understandable for Endesa Company.

Endesa Company has a history of NTLs which they have detected over time. The aim was to use these NTLs to detect other customers with similar NTLs through Bayesian networks (which are a supervised method).

Concretely, the sample (described in Section 2) constituted the following:

- Total number of customers: 38,575.
- NTLs registered for the two years of analysis: 606.
- Percentage of NTLs registered with respect to the total: 2.44%.

In order to characterize the pattern of consumption of each customer, a set of indicators to use as inputs of the decision trees was generated:

- Maximum and minimum value of the monthly or bimonthly consumptions (*MaximumConsumption*/*MinimumConsumption*): With this parameter the aim was to detect the different peaks and landings in the consumption of the customer during the analysis period. Thus, the maximum or minimum value of the different values of the bills in the analysis period was calculated and was divided by the contract power of the customer (this division was done in order to normalize these maximum and minimum values with respect to the expected consumption values for the customer).
- Number of readings (*NumberReadings*): This parameter stores the number of readings taken by the company at the counters of the customer during the two corresponding years of the analysis.
- Reactive/Active energy coefficient (*ReactiveActive*): This coefficient measures the proportion of consumed reactive energy by the customer for the two years of analysis in relation to its reactive one. In this way, the system would intend to measure important imbalances which characterized anomalies in these consumptions due to possible NTLs by the customer.
- Number of hours of maximum power consumption (*HoursConsumption*): This parameter calculated the range of consumption of the customers in relation to their contracted power and, therefore, to the consumption expected for the customers in the two years of analysis. It is derived from the total consumption of the client divided by its contracted power.
- Number of streaks of the customer (*NumberStreaks*): This parameter is obtained calculating the 6-month simple moving average for the consumption of each customer. Afterwards it counted the number of times the consumption line went above (positive streaks) and below (negative streaks) the mean line. An example of the counting of streaks for a customer is shown in Fig. 6.
- Estimator from the streaks of the customer (*StatisticStreaks*): The number of streaks and the weight of each streak for each customer offer interesting information about their consumption behavior. Thus, the following estimator for each customer was generated in order to integrate this information:

$$\text{Estimator\_Streak} = \sqrt{\frac{\sum_{t=1}^{Ns}(Nt)^2}{Ns}}$$

where *Ns* is the number of streaks of this customer and *Nt* is the number of measurements of the streak *t*.
- Coefficient relative to the difference between the energy billed to the customer (according to the information relative to the bills for that customer registered by the company) and the consumed energy (the energy consumed by the customer according

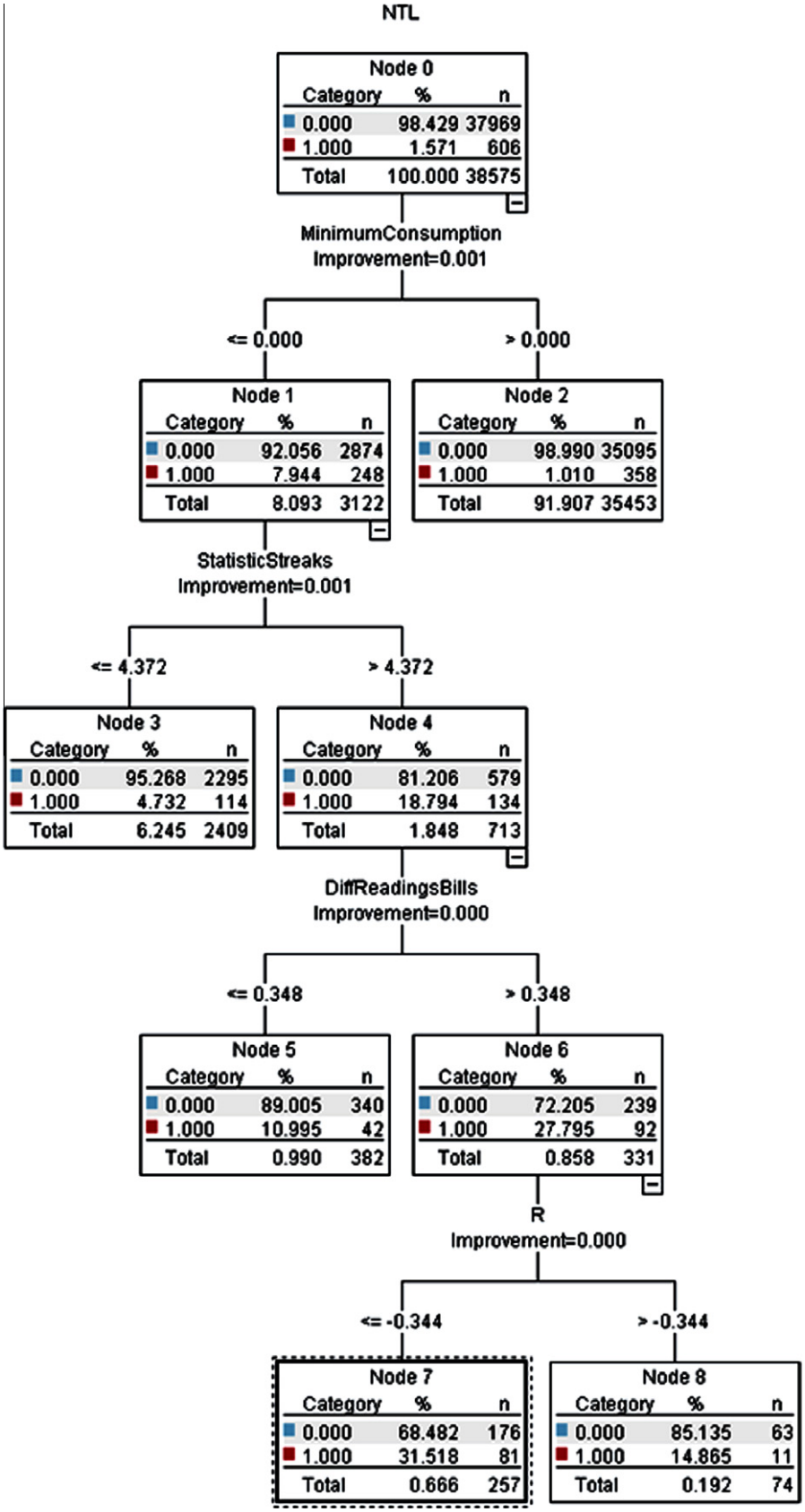

Fig. 9. Decision tree.

to the reading values at his counters) divided by the contracted power (to normalize the result) (*DiffReadingsBills*).

– The Pearson Coefficient (R) for the consumption in the time for the two years of analysis.

One of the problems that we could have to get a Bayesian network was the low number of NTLs compared to the total customers (38,575 customers and only 2.44% of NTLs registered by the company). Thus, we decided to balance the sample. The aim was to get a good ratio of customers without registered NTLs against customers with NTLs in order to obtain better results in the training of the Bayesian algorithm. This process is usual when the sample is very disproportionate and so the algorithms gets 'to take notice' better of the features of the minority cases (in this case the NTLs). This process is carried out copying n-times each one of the minority cases (in this case, the customers with NTL previously registered).

After some tests the best results was obtained balancing the NTLs with a coefficient of 20 (obtaining in the sample used to train the algorithm a final rate of 31% of NTL with respect to the total number of contracts).

After obtaining the balance mentioned earlier, different configurations of Bayesian networks were tested. We arrived at the best configuration with a structure of Markov blanket. Fig. 7 shows the network as well as the grade of importance of each one of the input parameters.

This network obtained good results since it achieved a 95% of rate success for the original sample (after the balance is undone). Fig. 8 shows the results of this network in detail. Besides, it shows the pattern of matches between each predicted field (NTL or no NTL) and its target field for the categorical target. Thus, the result is displayed with rows defined by actual values and columns defined by predicted values, with the number of records having that pattern in each cell.

In addition to concluding that the algorithm reached a 95% success rate in the detection, if we analyze the results in detail it is possible to observe that, specifically, the network obtained a 24% success in NTLs (144 correctly classified against the total 606). Thus, a possible conclusion from Fig. 8 is that if this Bayesian network was applied the system could detect customers with patterns similar to this 24% that we found inside the selected sample.

Besides the output prediction, Bayesian networks include an output with the probability that the prediction is correct. Thus, the 1357 customers predicted as NTLs were analyzed and those with a higher probability of correct prediction by the Bayesian network were selected. Concretely, a probability prediction higher than 0.8 would be an excellent warranty for the selection of these customers to be inspected by the company. There were 43 customers (from the said 1357) with a higher probability that this of having an NTL.

Apart from, and as complementary algorithm to the previous technique, we developed another line based on decision trees. For this, we used the tree generated from the C&RT algorithm [17]. This algorithm is a tree-based classification and prediction method. The C&R Tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. This algorithm has the advantage that the understandable rules that embed the NTLs can be extracted once the model is generated.

After a diverse test with some configurations, the best results were achieved (combining support – or the number of customers included in the leaves– and the confidence – or the rate between NTLs and customers without NTLs –) with a tree with depth 5. This decision tree is shown in Fig. 9. The selected leaf (Node 7) is those with highest of NTLs and with a support acceptable to be inspected in field. Thus, we counted 176 customers in this leaf without registered NTLs (it is supposed to be 31.5% on the total of customers of this leaf: 257). As they are too many to inspect in situ, we merge these customers with those from Bayesian network and with a probability higher than 0.5 to reduce the number of customers to be inspected. Thus, after this process, a total of 64 customers was obtained and selected to be inspected by the company.

Table 1
Customers selected to be inspected by the company.

Algorithm	No. customers selected
Progressive drop and stabilization	18
Drastic drop and stabilization	23
Bayesian network	43
Decision tree	64
Total	148
Once merged the results of the previous algorithms	140

5. Results and conclusions

Prior to sending the customers for the inspection, the results of all generated models were crosschecked in order to see how many customers were matching and to ensure that the different algorithms were not redundant.

Thus, after merging the customers detected with each one of the algorithms (the 2 algorithms to detect drops, the Bayesian network and the decision tree), the results shown in Table 1 were obtained.

As is evident from the table only 8 customers from the 148 selected customers were detected by more than one algorithm. Thus, we could deduce that each algorithm detected a type of different patterns of NTL.

Thus, a list of 140 customers with an evident and suspicious pattern of consumption with NTLs was obtained. These cases of NTL could be due to a drop of electrical demand for their business but never due to a low contract because in that case they would have reading information in their equipment.

Therefore, it was significant as additional information to study the type of business of these suspicious customers in order to know whether it was a business in which the demand is currently falling (e.g., currently, the construction business in Spain). Thus, we studied the business information for each customer in order to be able to control this fact and to avoid unnecessary inspections. It is known by the inspectors of the Company that the following types of business are more likely to have consumption drops innate to their use of the energy (and not due to possible NTLs): wells, lightings, irrigation pumps, water purification and construction (previously mentioned). So, from the 148 selected customers, we filtered those with these types of contracts (and therefore likely to have an anomaly pattern of consumption) and a definitive list of 101 was obtained.

In summary, a complete flow chart is shown in Fig. 10. In this diagram it is possible to observe the global scheme and the different steps for the detections of the NTLs.

Currently, the Endesa Company is carrying out inspections with a set of customers from the ones who were detected by the presented methods. Up to now, with the results obtained in the in situ inspections, the system has reached a success rate of 38%. These results are considered very satisfactory taking into account, first, the rate of success of the company in its routine inspections (less than 10%) and, second, the little input information used in the algorithms (only the evolution of the consumption of the customer and the type of contract).

As conclusions, it is necessary to remark that NTL is an important issue in power utilities because it has a high impact on company profits. Despite this, nowadays the methodology of detection of NTLs of the companies is very limited since these

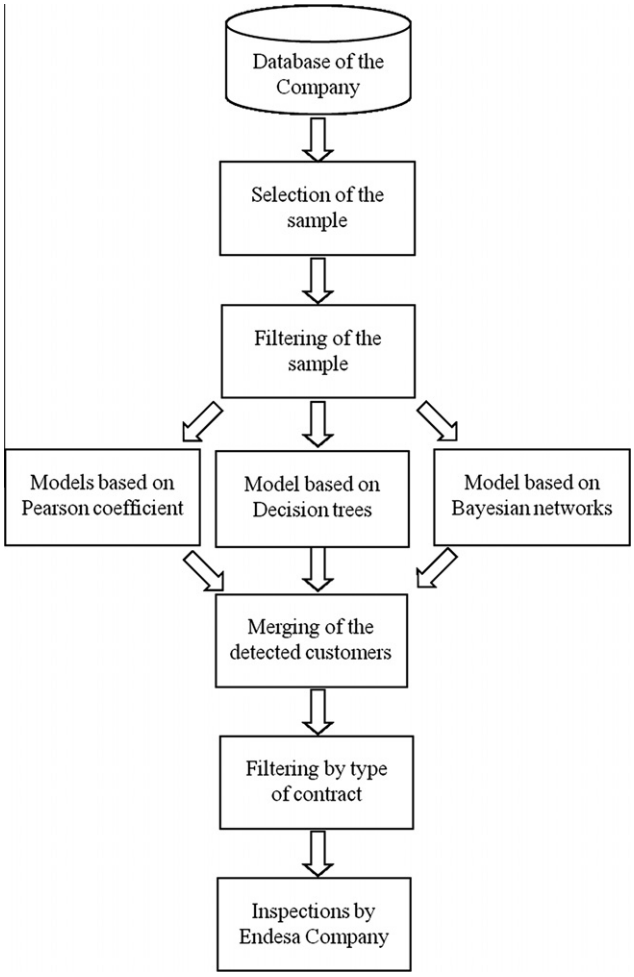

Fig. 10. Flow chart of the detection process.

companies use detection methods that do not exploit the use of data mining techniques. Different methods to detect NTLs have been developed and tested on a real database supplied by the Endesa Company. Concretely, in this paper, a line of work based on 5 different algorithms has been presented for the detection of NTLs using of the Pearson coefficient, Bayesian networks and decision trees. Before sending the customers for inspection, a table analysis involving a filtering task by the type of contract was carried out in order to enhance the accuracy percentage of the detections. The system obtained a success rate of 38% in the inspection of real customers aided by the presented algorithms. At present, in terms of energy, these detections are equivalent to a total energy recovery of about 2 millions of kWh, which implies a large amount of money is recovered for the Endesa Company.

Therefore, the contributions of this work with respect to the existing ones are as follows:

- The development of a system based on various complementary models with the application of techniques not implemented until now in the literature about the topic [11–16,23–25].
- A system tested in field and which is currently in operation by one of the most important power utilities of the world (Endesa Company).
- Good results (and better than those obtained with the lines of work existing in the literature) obtained both verification (with the different tests in the algorithms) and validation (the inspections carried out in situ).

Also, to avoid this filtering task carried out by the human component, we are currently developing an expert system that takes automatically the last filtering process (depending on the type of business of the customers) and the final selection of customers (from whose selected by the algorithms described in this paper) to be inspected in situ by Endesa Company.

Acknowledgments

The authors would like to thank the Endesa Company and Sadiel Company for providing the funds for this project (since 2005). The authors are also indebted to the following colleagues for their valuable assistance in the project: Gema Tejedor, Francisco Godoy and Joaquín Mejías. Special thanks to Jesús Macías, Eduardo Ruizbérriz, Juan Ignacio Cuesta, Tomás Blazquez and Jesús Ochoa for their help and cooperation.

References

[1] Wheeler R, Aitken S. Multiple algorithms for fraud detection. Knowl Based Syst 2000;13:93–9.
[2] Kou Y, Lu C-T, Sinvongwattana S, Huang Y-P. Survey of fraud detection techniques. In: IEEE int conf on netw sens and control. Taiwan; 2004. p. 89–95.
[3] Fawcett T, Provost F. Adaptative fraud detection. Data Min Knowl Discov 1997;1:291–316.
[4] Arťıs M, Ayuso M, Guillén M. Modeling different types of automobile insurance frauds behavior in the spanish market. Insur Math Econ 1999;24:67–81.
[5] Daskalaki S, Kopanas I, Goudara M, Avouris N. Data mining for decision support on customer insolvency in the telecommunication business. J Oper Res 2003;145:239–55.
[6] Brause R, Langsdorf T, Hepp M. Neural data mining for credit card fraud detection. In: Proc. 11th IEEE int conf on tools artif intell; 1999. p. 53–61.
[7] Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. Expert Syst Appl 2007;32:995–1003.
[8] Burge P, Shawe-Taylor J. Detecting cellular fraud using adaptative prototypes. In: Proc on AI approaches to fraud detect and risk manage, vol. 42, 1997. p. 9–13.
[9] Cabral J, Pinto J, Linares K, Pinto A. Methodology for fraud detection using rough sets. In: IEEE int conf on granul comput; 2006. p. 246–49.
[10] Jawad N, Keem SY, Sieh KT, Syed AK, Malik M. Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans Power Deliv 2010;25(2):1162–71.
[11] Yap KS, Hussien Z, Mohamad A. Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm. In: Proc third IASTED int conf adv comput science and tech Phuket. Thailand: Iasted press; 2007.
[12] Cabral J, Pinto J, Gontijo EM, Reis J. Fraud detection in electrical energy consumers using rough sets. In: IEEE int conf on systems, man and cybern; 2004. p. 3625–9.
[13] Cabral J, Pinto J, Martins E, Pinto A. Fraud detection in high voltage electricity consumers using data mining. In: IEEE transm and distrib conf and exposit. T&D. IEEE/PES; 2008. p. 1–5.
[14] Sforna M. Data mining in power company customer database. Electr Power Systs Res 2000;55:201–9.
[15] Jiang R, Tagiris H, Lachsz A, Jeffrey M. Wavelet based features extraction and multiple classifiers for electricity fraud detection. In: Trans and distrib conf and exhibit. 2002: Asia pacific. IEEE/PES; 2002. p. 2251–6.
[16] Kantardzic M. Data mining: concepts, models methods and algorithms. 1st ed. AAAI/MIT Press; 1991.
[17] Witthen I, Frank E. Data mining–practical machine learning tools and techniques with java implementations. New York, San Mateo, CA: Morgan Kaufmann, Acad Press; 2000.
[18] Editorial. Recent advances in data mining. Eng App Artif Intell; 2006. p. 19.
[19] McCarthy J. Phenomenal data mining. Commun ACM 2000;43(8):75–9.
[20] Ramos S, Vale Z. Data mining techniques application in power distribution utilities. In: IEEE transm and distrib conf and exposit. IEEE/PES; 2008. p. 1–8.
[21] Valero S, Ortiz M, Senabre C, Gabaldon A, Garcia F. Classification filtering and identification of electrical customer load pattern through the use of self-organizing maps. IEEE Trans Power Syst 2006;21(4):1672–82.
[22] Othman ML, Aris I, Othman MR, Osman H. Rough-set-and-genetic-algorithm based data mining and rule quality measure to hypothesize distance protective relay operation characteristics from relay event report. Int J Electr Power Energy Syst 2011;33:1437–56.
[23] Biscarri F, Monedero I, León C, Guerrero JI, Biscarri J, Millán R. A data mining method based on the variability of the customers consumption. In: 10th int conf on enterp inf syst, ICEIS2008, Barcelona, Spain, June 12–16.
[24] Biscarri F, Monedero I, León C, Guerrero JI, Biscarri J, Millán R. A mining framework to detect non-technical losses in power utilities. In: 11th Int conf on enterp inf syst, ICEIS2009, Milano, Italy, May 6–10.

[25] Monedero I, Biscarri F, León C, Guerrero JI, Biscarri J. Using regression analysis to identify patterns of non-technical losses on power utilities. KES 2010, LNAI 6276; 2010. p. 410–19.
[26] Pearson K. Mathematical contributions to the theory of evolution—III. Regression, heredity and panmixia. Philos Trans R Soc London Ser A 1896;187:253–318.
[27] Moore D. Basic practice of statistics. In: Freeman WH, editor. San Francisco, CA, USA; 2006.
[28] Lee KC, Jo NY. Bayesian network approach to predict mobile churn motivations: emphasis on general Bayesian network. Markov blanket, and what-if simulation, LNCS, 6284; 2010. p. 304–13.
[29] Przytul KW, Dash D, Thompson D. Evaluation of Bayesian networks used for diagnostics. In: Proc IEEE Aerospace Conf, vol. 60, 2003. p 1–12.
[30] Thornton J, Gustafsson T, Blumenstein M, Hine T. Robust character recognition using a hierarchical Bayesian network, LNCS, 4304; 2006. p. 1259–66.
[31] Aggarwal SK, Saini LM, Kumar A. Electricity price forecasting in deregulated markets: a review and evaluation. Int J Electr Power Energy Syst 2009;31:13–22.
[32] Riascos LAM, Simoes MG, Miyagi PE. A Bayesian network fault diagnosis system for proton membrane exchange fuel cells. J. Power Sources 2007;165:267–78.
[33] Buschkes R, Kesdogan D, Reichl P. How to increase security in mobile networks by anomaly detection. In: Proc 14th comput secur appl conf (ACSAC '98); 1998. p. 8.
[34] Maes S, Tuyls K, Vanschoenwinkel B, Manderick B. Credit card fraud detection using Bayesian and Neural networks. In: Proc. neuro fuzzy, Havana, Cuba; 2002.
[35] Samantaray SR, Dash PK. Decision tree based discrimination between inrush currents and internal faults in power transformer. Int J Electr Power Energy Syst 2011;33:1043–8.
[36] Ugedo A, Lobato E. Application of neural networks to the management of voltage constraints in the Spanish market. Int J Electr Power Energy Syst 2011;33:1261–71.
[37] Bonchi F, Giannotti F, Mainetto G, Pedreschi D. A classification-based methodology for planning audit strategies in fraud detection. In: conf on knowl discov data: proc. fifth Acm SigKdd; 15–18 August, 1999.