NEW METHODS TO DETECT NON-TECHNICAL LOSSES ON POWER UTILITIES

Iñigo Monedero[1], Félix Biscarri[1], Carlos León[1], Juan I. Guerrero[1], Jesús Biscarri[2], Rocío Millán[2]
[1]Department of Electronic Technology, University of Seville, C/ Virgen de Africa, 7, 41011 Sevilla, Spain
[2]Endesa, Avda. Borbolla S/N, 41092 Seville (Spain)
imonedero@us.es, fbiscarri@us.es, cleon@us.es, juguealo@us.es, jesus.biscarri@endesa.es, rmillan@endesa.es

ABSTRACT
A non-technical loss is defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment. This paper describes new approaches to the detection of these for electrical companies. Concretely two different methods are proposed: one based on the detection by means of a clustering process with decision trees and another one with a simple but effective algorithm that allows detecting the customers with drastic drops of consumption. The analysis and the obtained results correspond to real customers of Endesa Company (one of the most important electrical companies of Spain).

KEY WORDS
Data mining, non-technical loss, fraud, decision tree, electrical company, power utility

1. Introduction

A non-technical loss (NTL) is defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment. Therefore, detection of NTLs in a particular field includes detection of fraudulent users. For the electrical distribution business, minimizing NTLs is a very important task. In Spain, it is estimated that the percent of fraud in terms of energy with respect to the total NTLs round about 35%-45%. Not only electrical companies have NTLs in its customers but also telecommunication companies, credit cards, automobile companies or even in medical insurances. Thus, although about these ambits in the literature there are many works and researches [1-13], however there is not too much research about NTL detection in electrical companies [14-20] although as we have said and it is verified the NTLs are very extended in this field.

Thus, nowadays the main methodologies carried out by the electrical companies in the detection of NTLs are 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 the need for a large number of inspectors and, therefore, a high cost. The problem with the second option is the impossibility of detecting users with non-null consumption (these are only the clearest cases of non-technical losses).

Nowadays, data mining techniques [21-22] are being applied to multiple fields and detection of NTLs is one field in which it has met with success recently [23-26]. Thus, it is possible to apply data mining techniques in order to improve the inspection success and the profitability rate, highly dependent of the cluster of customers researched, i.e, of the set of features that made the cluster and the class of customers studied (domestic customers, medium or high consumption customers.).

This paper describes a set of data mining techniques included on a prototype for NTL detections from the databases of Endesa Company. The work is within the framework of MIDAS project which we are developing at the Electronic Technology Department of the University of Seville with the funding of the electrical company. Concretely in this paper we describe two methods: one based on decision trees and another one based on detection of drops of consumption.

2. Selection of the samples

In first place in order to carry out the tests for the development of the prototype we selected a data set to work. For it we covered the two most important regions of Endesa Company: Catalonia and Andalusia. We extracted a sample set for each region made by customers with rate 3.0.2 and 4.0 and besides with a high contracted power (>40 KW). Our objective were to include in our analysis those customers with highest consumption (this was interesting because each detected NTL could mean a lot of money for the company). The number of users in each set, as well as the number of NTLs registered by the electrical company occurred in the analysis period, is shown in Table 1 (it was interesting to have identified these customers with NTL in order to detect similar patterns in the rest of the customers).

Thus, four tables were used from the main database of Endesa Company for each one of the regions: one of contracts data (with the information relative to the type of rate, address, contracted power, etc of the customer), another one with the reading values of the measurements equipment of the customers, another one for bills (which included for the study period each one of the bills of the

customers) and a last fourth one including all the cases with NTLs registered previously by the company by means its strategies non-based on data mining through its inspections. We configured an analysis period of 2 years which were a time enough to see a sufficiently detailed evolution of the consumption of the customer and, on the other hand, not too long to register along the contract the possible changes of type of business or the changes in the consumption habits of the client. We generated a new table from the linking of these four tables which included condensed all the information of consumption and type of contract for each costumer: reading values of the measurements equipment, bills from the last 2 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. With this information in our study we could identify the type of customer, its expected consumption and the evolution of its consumption in the last two years (as well as the bill paid by the client in each period).

Table 1.- Data sets selected for analysis

Sample set	**Number of customers**	**Cases previously detected with NTLs**	**% NTLs vs. number of customers**
Catalonia	27695	598	2.15%
Andalusia	14706	396	2.69%

For these data sets, we carried out a pre-processing of data and so these were prepared for the mining process in this phase. An interesting point of the pre-processing was the concerning one to the reading values of the measurement equipment. Normally, the consumption billed is the result of consumption read, but it is not always true. If the company has no access to read the data, and there is no doubt of consumption has been made, company experts estimate the actual consumption, based on the recent historical consumption. Several and continuous differences between read data and billed data show abnormal behaviour. In this sense, a filling up of missing values has been performed.

So, with the result of the pre-processed samples we generate two new tables (one for each region) with one entry for each customer condensing all their information relating to the two last years': including bills (for the active and reactive energy consumed by the client), processed reading values of the measurements equipment of each customer, the power supplied, the type of rate and the type of client or business.

3. Method based on the use of decision trees

In order to detect possible NTLs inside the samples, we have developed a first method based on the recognition of customers with the same consumption patterns that those customers with NTL previously detected by the company in its inspections. This method was held on a process of generation of clusters (this technique has already been used previously in another works although with different approach [11-12], [27]). Thus, firstly we designed a feature vector that could identify the consumption pattern of each one of the customers. This vector included the following parameters:

- Number of hours of maximum power consumption: This parameter calculated the range of consumption of the customer in relation to its contracted power and, therefore, to the consumption expected for that customer in the two years of analysis. It is calculated as the division between the total consumption of the client and its contracted power.
- Standard deviation of the monthly or bimonthly consumption: This parameter identified the possible irregularities or changes of consumption of the customer along the two years of analysis. It is calculated as the standard deviation of the consumption values referring to the whole set of the bills (24 o 48 bills, depending the type of billing: monthly o bimonthly) of the client for the two analysis years. In this way, we intended to identify those customers with sudden changes of consumption due to possible frauds or some type of anomaly.
- Maximum and minimum value of the monthly or bimonthly consumptions: With this parameter our aim was to detect the different peaks and landings in the consumption of the customer during the analysis period. Thus, it is calculated as the maximum or minimum value of the different values of the bills in the analysis period.
- Reactive / Active energy coefficient: This coefficient measures for the two years of analysis the proportion of consumed reactive energy by the customer in relation to its reactive one. Thus, in this way, we intended to measure important imbalances which characterized anomalies in these consumptions due to possible fraud by the customer.

In our research we deduced that an important indicator of some types of NTLs in the customers could be the irregularity in their consumption. Thus, we developed two additional parameters to the previously described ones. These parameters were based on the concept of streak. A streak in a time series is the number of times that the values of this series cross the mean of their values (calculated from the initial value of the sample to that value). Streaks of past outcomes (or measurements), for example of gains or losses in the stocks market, are one source of information for a decision maker trying to predict the next outcome (or measurement) in the series.

In the case of gambling, each gamble is an independent event, so there is no casual mechanism linking outcomes (hence the fallacy). The customer consumption trend has, presumably, an underlying casual model. It depends on the seasonality, economical activity and other hidden features. The discovery of the theoretical consumption model is not the target of this paper. This model is strongly dependent of the cluster of customers considered and highly changeable amongst different clusters. But it is an interesting customer feature that their consumption trend depends of the consumption trend of the other customers in the same cluster.

In order to integrate in our work the concept of streak we used two additional parameters to the previously described ones:

- Number of streaks of the customer: This parameter is obtained calculating the six-month simple moving average for the consumption of each customer. Afterwards we counted how many times the consumption line is over the mean line (positive streaks) and how many times the consumption line was below the mean line (negative streaks). An example of counting of streaks for a costumer is shown in figure 1.
- Estimator from the streaks of the customer: The number of streaks for each customer offers interesting information about their consumption behaviour but it is also interesting to know the weight of each streak. Thus, we generate the following estimator for each customer in order to integrate this information:

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

Where *l* is the customer identifier, *Ns* is the number of streaks of this customer and *Nt* is the number of measurements of the streak.

Once we generated this set of parameters to identify the features of consumption of each customer, we developed the algorithm for the clustering process of the different customers. As it was said our final objective was the detection of consumption pattern similar to the previously detected by the electrical company in its inspections. Thus, afterwards, from the different obtained clusters we extracted rules which allowed us to separate or isolate those clusters with a high number of NTLs vs total number of customers (therefore, those ones without registered NTL being potentially suspicious to have it). In order to carry out this process we used a powerful software very extended for the use of data mining: SPSS Clementine (in its version 11). This software has got libraries for the automatic generation of decision trees that make it possible to carry out a clustering process. This type of algorithm (which is very extended in the data mining literature [1-2], [24-25]) build trees looking for a given objective (in our case to isolate NTLs). The final result is a set of leafs whose elements are covered by rules. Thus, in order to get validated results we decided to obtain the rules from the region with higher rate of NTLs (Andalusia) and to validate these rules with the other region (Catalonia).

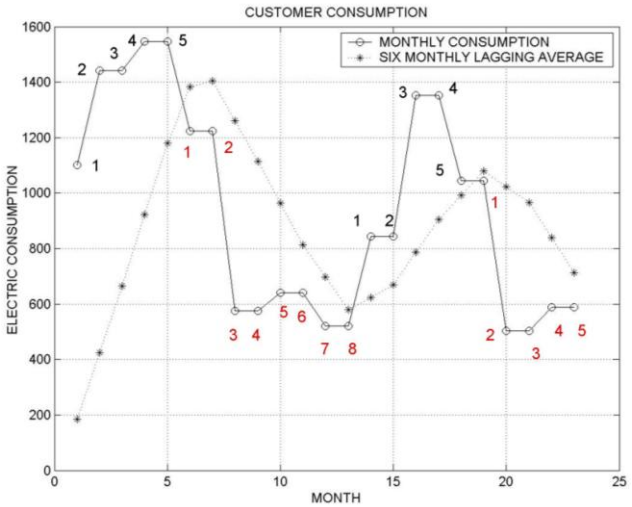

Figure 1.- Example of a customer with 3 streaks

Once we formatted the information for SPSS Clementine and built the corresponding algorithm we obtained the decision tree for Andalusia (it is shown in figure 2). Among the different nodes of the decision tree we selected two as more interesting nodes (nodes 6 and 10). These two nodes combined a high rate of NTLs with respect to the number total of customer and, on the other hand, a manageable number of customers for field inspections. This last feature was important by the fact that the final stage for the validation of the detection after our work was the checking in-situ of the possible NTL by the inspectors of the company. Thus, and as our system is still a prototype, we did not wanted to generate too many cases for inspection to validate completely our system.

Thus, the two selected sets had 50 and 36 customers respectively. The first set had 19 NTLs while the second one had 14. Thus, the percent of NTLs was around 40% for the first rule and around 25% for the second rule (improving importantly the original percent of 2.69%). The rules of both nodes were the following ones:

- Node 6: Minimum consumption < 0.05, Estimator from streaks> 2.272 and Number of streaks > 4.5 (as it could be deduced this rule reached those customers with an erratic consumption).
- Node 10: Minimum consumption < 0.05, Number of streaks < 4.5, Maximum consumption > 2226 and Estimator from streaks> 8.349 (this rule seemed to reach the customers with little streaks and those consumption has evolved from high values to lower ones).

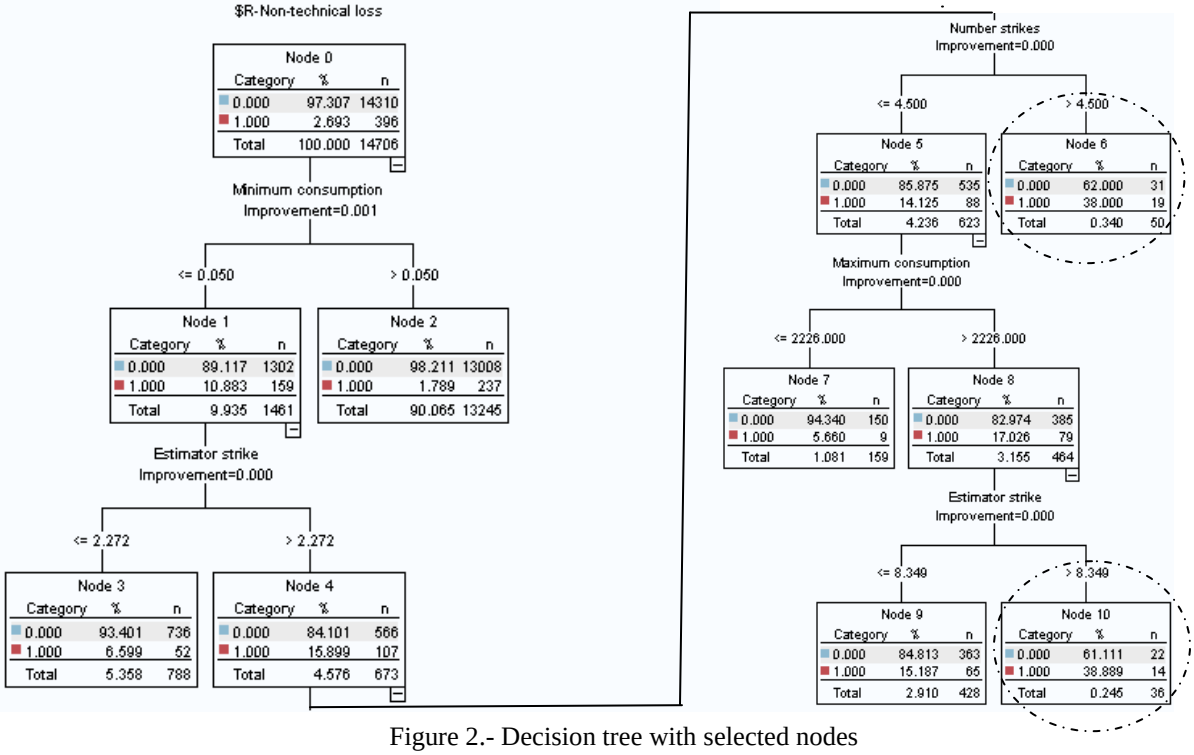

Figure 2.- Decision tree with selected nodes

Afterwards, we carried out an additional validation process of these rules generated for Andalusia testing them in Catalonia. Thus, applying these rules on the other sample set we obtained the whole results (for test and validation) of table 2. As it can be observed the results obtained with the generated rules for Catalonia were satisfactory getting a rate percent around 20%. Therefore it could be observed that the generated rules were valid for both sample sets and thus the validation process could be considered performed.

Table 2.- Results of the rules on both sample sets

Sample set	**Rule 1**		**Rule 2**	
	Custo-mers	**NTL**	**Custo-mers**	**NTL**
Andalusia (test set)	50	19	36	14
Catalonia (validation set)	57	10	45	9

4. Method based on the detection of drop in consumption

An evident symptom for the detection of NTL in the customers is a drastic drop of their consumptions. These drops can be due to a real slope of the consumptions of the customers (for example due to a change of type of contract or by a different use of the consumed energy). But turn these slopes can be due to failures in the measurement equipment or voluntary alterations of these equipment (both cases generates NTLs to the company and therefore loss of money for it). In order to detect these types of NTLs again with the help of SPSS Clementine we have developed an effective algorithm for detecting drops of consumption. This algorithm is based on the comparison of the consumption of the second year with the consumption of the first one (of the two years we used in order to carry out the detections). Thus, our algorithm of detection carries out the following steps (additional to the previous filling up of missing values of readings performed during the stage of pre-processing of data):

1.- It is carried out a selection of the customers which have readings on their equipment both the first and second years. Thus we discard the customers whose contract finished in the first year of analysis.
2.- It is generated a parameter calculated as the difference between the mean of the consumption of the first year of analysis and the mean of the consumption of the second one. The result of this subtraction is divided by the mean of the consumption for the two years of analysis (this division was carried out in order to normalize the entire sample set in function of the range of consumption in where was moved each customer).
3.- The customers are sorted by the previously calculated coefficient. The result of this process is a table which the first entries have highest difference of normalized consumption between the first and the second analyzed year. Thus, in figure 3 we can observe the graphical representation of the consumption (KWs) for the two years of the analysis of four among the set of customers of both samples (first two ones for Andalusia and second two ones for Catalonia).

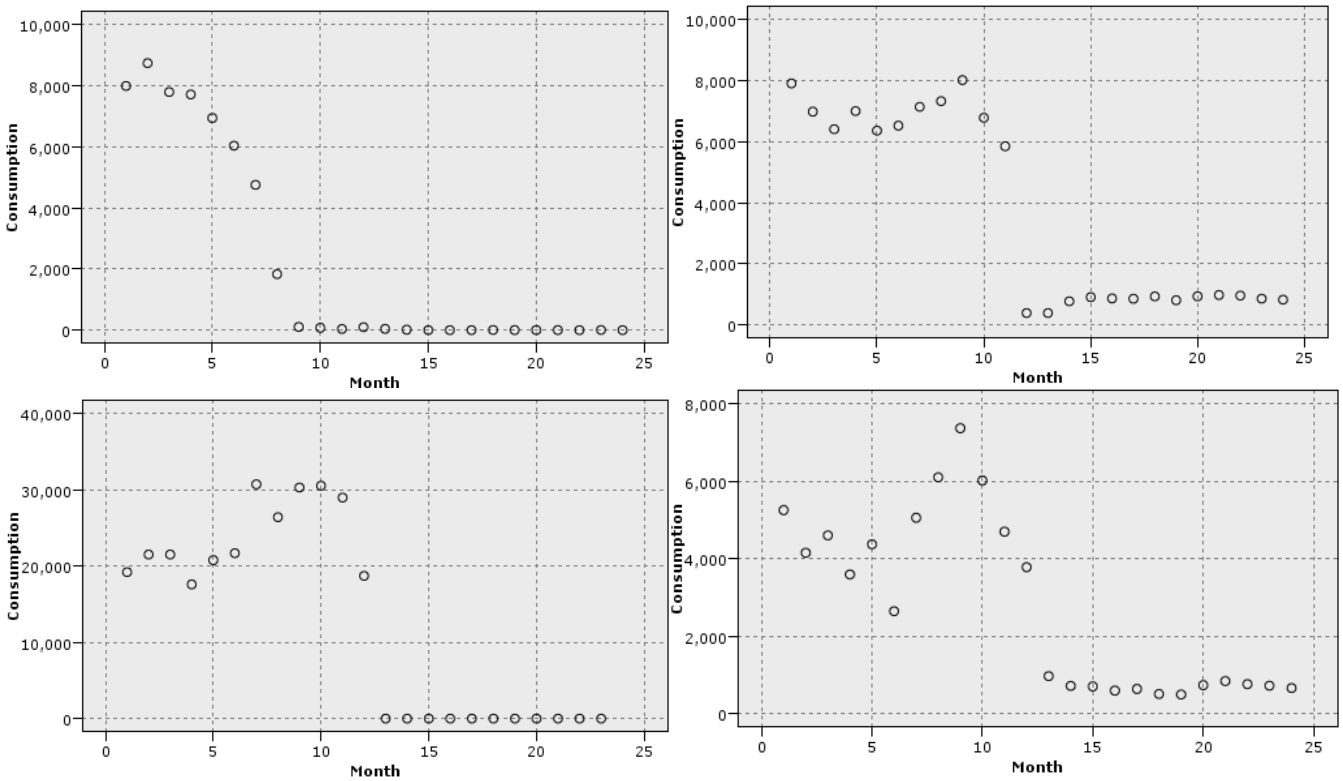

Figure 3.- Drastic drop in consumption of 4 customers

As it can be observed in these four customers the difference of consumption between the first and the second year is very significant. In the two customers on the left the consumption falls to zero while in the two ones on the right their consumption falls significantly but not to zero (these drops are especially difficult to detect by the electrical company in its current inspections). These consumptions are very suspicious to have some NTL (as some type of failure in their measurement equipment or fraudulent alterations of these). These cases could be due to a drop of electrical demand of their business but never due to a contract low because in that case they would have reading information in their equipment. Therefore, it was interesting as additional information to study the type of business of these suspicious customers in order to know if it was businesses in which currently is falling the demand (for examples currently building business in Spain). We added to our tables that business information for each customer in order to be able to control this fact and to avoid unnecessary inspections. Thus, in the customers of figure 3 their businesses were respectively: a bakery, a hostelry business (concretely a discotheque), a bank and a restaurant. A priori these types of business did not seem typical businesses with demand fall (even in the case of the hotel and the pub it could be very suspicious). Thus, in these cases the final testing of the motif of the drop would be task for the inspectors of Endesa Company. Using this method we selected the 20 more representative customers (10 in each one of the 2 samples) to be inspected.

5. Conclusion

In electrical companies, NTL is an important issue because it has a high impact in the company profits. Despite this, nowadays the methodology of detection of NTL of the companies is not very advanced using detection methods that do not exploit the use of data mining techniques. In this paper, we have proposed two data mining methods that we have developed for Endesa Company in order to detect customers with NTL. The aim of the first method was the search by means of decision trees of customers with consumption patterns similar to customers with NTLs previously detected by the company. On the other hand, the second method is based on the detection of pronounced drops in the annual consumption of the customers with respect to their consumption of the following year.

The methods were tested on a real database supplied by Endesa Company. Thus, the first method obtained satisfactory results improving importantly to 20% the around 2% rate of NTLs vs number of customers from the original data sets of the company. Besides, as we had two different sample sets we used one of them for validating the rules extracted with the other. On the other hand, with the second method we selected a sample of 20 more representative customers which had clear drops of the consumption in the second year of analysis. Currently, Endesa Company is carrying out inspections with a set of customers from the ones who were detected by our methods.

One of the possible improvements that we could include in our approach is based on the fact that a hand analysis is currently needed after the application of our methods. It is because our methods use basically as input information the consumption data of the customers. But there is more information that can be determining to decide to study in field that customer (as for example the type of business, the stationary consumption in some types of business or even the location of the customer). Thus, we are currently working on an expert system which takes as input all this information from the database and so carries out the task of the hand analysis. On the other hand, we are working in the task of improving the detection results. For it, we are using new input parameters and testing other data mining algorithms.

Acknowledgements

The authors would like to thank the Endesa 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 and Francisco Godoy. Special thanks to Juan Ignacio Cuesta, Tomás Blazquez and Jesús Ochoa for their help and cooperation to extract the data from Endesa Company.

References

[1] R. Wheeler and S. Aitken, "Multiple algorithms for fraud detection," *Knowledge based systems*, no. 13, pp. 93–99, 2000.

[2] Y. Kou, C.-T. Lu, S. Sinvongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Proceeding of the 2004 IEEE InternationalConference on Networking, Sensing and Control. Taiwan, march 21 89–95*. IEEE press, 2004.

[3] T. Fawcett and F. Provost, "Adaptative fraud detection." in *Data mining and Knowledge Discovery 1 291–316*, 1997.

[4] M. Art´ıs, M. Ayuso, and M. Guillén, "Modeling different types of automobile insurance frauds behavior in the spanish market," in *Insurance Mathematics and Economics 24 67–81*. Elsevier Press, 1999.

[5] S. Daskalaki, I. Kopanas, M. Goudara, and N. Avouris, "Data mining for decision support on customer insolvency in the telecommunication business," in *European Journal of Operational Research 145 239–255*. Elsevier press, 2003.

[6] H. X. He, J. C. Wang, W. Graco, and S. Hawkins, "Application of neural networks to detection of medical frauds," in *Expert Systems with Applications 13 (4) 329–363*. Elsevier press, 1997.

[7] H. He, W. Graco, and X. Yao, "Application of genetic algorithm and k-nearest neighbors in medical fraud detection," in *Lecture Notes in Computer Science 1585 74–81*. LNCS, 1999.

[8] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *Proceeding 11th IEEE International Conference on Tools with Artificial Intelligence*. IEEE press, 1999.

[9] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," in *Expert Systems with Applications 32 995–1003*, 2007.

[10] P. Burge and J. Shawe-Taylor, "Detecting cellular fraud using adaptative prototypes," in *Proceeding on AI Approaches to Fraud Detection and Risk Management. 9–13*. Menlo Park, CA: AAAI Press, 1997.

[11] P. L. Brokett, X. Xia, and R. A. Derrig, "Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud," in *The Journal of risk and Insurance 65(2) 245–274*, 1998.

[12] J. Cabral, J. Pinto, K. Linares, and A. Pinto, "Methodology for fraud detection using rough sets" in *2006 IEEE International Conference on Granular Computing*. IEEE press, 2006.

[13] D. Denning, "An intrusion-detection model," in *IEEE transactions on Software Engineering 13 222–232*. IEEE press, 1987.

[14] T. Lunt, "A survey of intrusion detection techniques," in *Computers & Security, 12 405–418*, 1993.

[15] K.S.Yap, Z. Hussien, and A. Mohamad, "Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm," in *Proceeding of the Third IASTED International Conference Advances in Computer Science and Technology*. IASTED PRESS, April 2-4, Phuket, Thailand, 2007.

[16] J. Filho and als, "Fraud identification in electricity company costumers using decision tree," in *IEEE International Conference on Systems, Man and Cibernetics*. IEEE/PES, The Hague, The Netherlands, 2004.

[17] J. Cabral, J. Pinto, E. M. Gontijo, and J. Reis, "Fraud detection in electrical energy consumers using rough sets," in *2004 IEEE International Conference on Systems, Man and Cybernetics*. IEEE press, 2004

[18] J. Cabral, J. Pinto, E. Martins, and A. Pinto, "Fraud detection in high voltage electricity consumers using data mining," in *IEEE Transmission and Distribution Conference and Exposition T&D*. IEEE/PES, April 21-24, 2008.

[19] M. Sforna, "Data mining in power company customer database," in *Electric Power Systems Reseach, 55, 201-209*. Elsevier Press, England, 2000.

[20] R. Jiang, H. Tagiris, A. Lachsz, and M. Jeffrey, "Wavelet based features extraction and multiple classifiers for electricity fraud detection," in *Transmission and Distribution Conference and Exhibition 2002: Asia pacific*. IEEE/PES, Oct. 6-10, 2002.

[21] M. Kantardzic, *Data mining: concepts, models methods and algorithms*,1st ed. Ed. AAAI/MIT Press, 1991.

[22] I. Witthen and E. Frank, *Data Mining–Practical Machine Learning Tools and Techniques with Java Implementations*. New York and San Mateo, CA: Morgan Kaufmann, Academic Press, 2000.

[23] Editorial, "Recent advances in data mining," in *Engineering applications of Artificial Intelligence 19 361–362*, 2006.
[24] J. McCarthy, "Phenomenal data mining," *Communications of the ACM*, vol. 43, no. 8, pp. 75–79, August 2000.
[25] S. Ramos and Z. Vale, "Data mining techniques application in power distribution utilities," in *IEEE Transmision and Distribution Conference and Exposition T&D*. IEEE/PES, April 21-24, 2008.
[26] A. Nizar, Z. Dong, and J. Zhao, "Load profiling and data mining techniques in electricity deregukated market," in *Power Engineering Society General Meeting*. IEEE/PES, June 18-22, 2006.
[27] S. Valero, M. Ortiz, C. Senabre, A. Gabaldón, and F. García, "Classification, filtering and identification of electrical customer load pattern through the use of self-organizing maps," *IEEE transactions on Power Systems*, vol. 21, no. 4, nov. 2006.