MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques

Íñigo Monedero[1], Félix Biscarri[1], Carlos León[1],
Jesús Biscarri[2], and Rocío Millán[2]

[1] Escuela Técnica Superior de Ingeniería Informática,
Departamento de Tecnología Electrónica, Avda,
Reina Mercedes s/n, 41012 Seville (Spain)
imonedero@us.es
[2] Endesa, Avda. Borbolla S/N, 41092 Seville (Spain)

Abstract. Datamining has become increasingly common in both the public and private sectors. 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. The detection of non-technical losses (which includes fraud detection) is a field where datamining has been applied successfully in recent times. However, the research in electrical companies is still limited, making it quite a new research topic. This paper describes a prototype for the detection of non-technical losses by means of two datamining techniques: neural networks and statistical studies. The methodologies developed were applied to two customer sets in Seville (Spain): a little town in the south (pop: 47,000) and hostelry sector. The results obtained were promising since new non-technical losses (verified by means of in-situ inspections) were detected through both methodologies with a high success rate.

1 Introduction

Datamining [1][2] is a computing tool which involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Nowadays, datamining is being applied to multiple fields and detection of non-technical losses is one field in which it has met with success recently [3]. 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. Therefore, detection of non-technical losses includes detection of fraudulent users.

This datamining field involves identifying non-technical losses as quickly as possible once it has been happened. Normally, cases of non-technical loss have to be detected from huge data sets such as the logged data and user behaviour. The workforce is thus not sufficient to analyze these huge data sets and datamining techniques are the only tools which make it possible to study all the data in an acceptable time. The main research and applications in the non-technical losses and fraud detection field have been carried out on credit cards, telecommunications, and computer intrusion

[4][5][6][7]. Thus, for instance, telecommunication non-technical loss can be found in subscriptions where access to a service is obtained, often with false identity details, with no intention of paying. Other kinds of non-technical loss in this field occur from using a service without the necessary authority (for example, using mobile phone cloning) detected by the appearance of unknown calls on a bill. The tools and techniques in these cases involve detecting the users with non-technical loss quickly in order to report them and to recover the lost money.

Not only telecommunication companies and banks have non-technical losses in its users but also electrical companies. However, there is still very little non-technical detection research in electrical companies. Thus, once we have carried out a study of the state of art in this field and we have only found a very few papers [8], being therefore a research topic quite new.

Current methodology work by the electrical companies in the detection of non-techincal losses 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 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).

This paper describes a prototype for the detection of non-technical losses which has been developed at the Electronic Technology Department of the University of Seville.

2 Midas Project

MIDAS is the name of a project which developed two methodologies for the detection of non-technical losses, one by means of neural networks and the other by means of statistical techniques. The project was financed by Endesa, the most important electrical company of Spain and one of the most important ones in the world, and by Sadiel which is the most important consulting company of Andalusia and one of the most important ones of Spain. A representation of the stages carried out in the development of the project is shown in Figure 1. Each of the phases shown in the figure are described below:

1 Data Selection: The aim of this first phase was the selection from the database of the electrical company of a data set with which to work for the development of the prototype. Specifically, two sets were selected: Hostelry business in the province of Seville and private users in a little town in the south of Seville with a population of about 47,000. The first set was selected for being a traditionally sector of many non-technical losses. On the other hand, the chosen little town was the Sevillian postal code in which the electrical company had registered, through its inspections, the largest number of non-technical loss cases. The non-technical loss files of these users were also collected.

2 Query and formatting data: In this second phase the SQL queries were carried out and the data was formatted. Thus, three tables were obtained from the database: one for contracts, one for bills and a third one for files of cases with non-technical

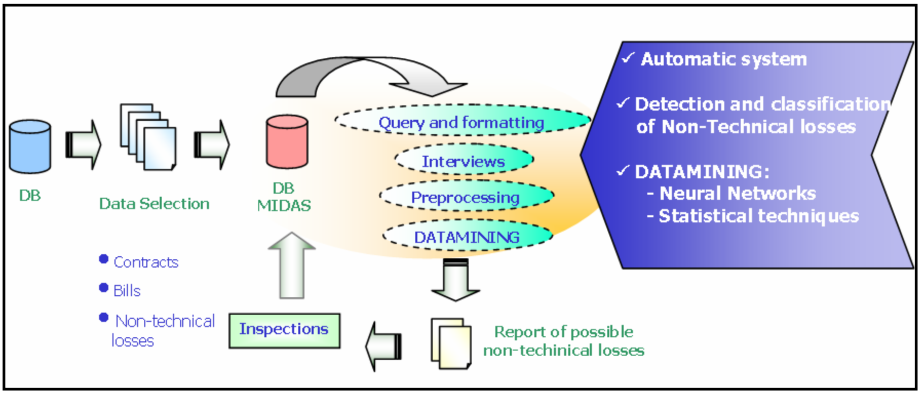

Fig. 1. Phases in the development of Midas

losses. The number of users in each set as well as the number of registered non-technical losses is shown in Table 1. As may be observed, the number of detected cases with non-technical losses was low in relation to the total number of contracts. The three database tables include, for each user, the following fields: bills from the last 4 years (one bill every two months), amount of power contracted and the type of customer (private or the what kind of business), address, type of rate, etc.

Table 1. Analyzed data

Type of customer	Number of Contracts	Cases with Non-Technical Losses
Little town	60048	17
Hostelry business	12879	5

3 Interviews: This task involved a number of interviews (exactly five were carried out) with specialized staff of the company for the detection of non-technical losses. In these interviews the different kinds of non-technical loss were studied as well as the various characteristics in the consumption evolution of each of them. This phase was carried out parallel to the two other previous phases.

4 Preprocessing: Data was prepared for the mining process in this phase. To this purpose, the entire fields from the tables shown above were studied and two new tables (one for each kind of customer) were generated with a set of fields selected for mining. Specifically, each register in these tables (one for each user) included: the six bills corresponding to the last year’s consumption, the power supplied and a non-technical loss sign (the value of this field was 1 if the user had registered some non-technical loss and otherwise the value was 0).

5 Datamining (with Neural Networks and Statistical Techniques): These two stages, which were carried out in parallel, involved the data-mining process. The techniques applied were neural networks and a statistical study. Our aim was to try to

obtain results through two different methods in order to obtain two result sources and compare them. The data mining process using these two techniques is described in detail in Sections 3 and 4.

6 Reporting possible non-technical losses and inspections: Once the data mining process was carried out, the electrical company would receive reports on customers detected with possible non-technical losses. Finally, the company would study these customers individually to decide on which ones to carry out the inspections. Thus, at the same time, by means of the inspections we could test the validity of the datamining process.

3 Data Mining with Neural Networks

Artificial neural networks are abstract simulations of a real nervous system that contains a collection of neuron units communicating with each other via axon connections. Algorithms based on neural networks are among the most popular data mining techniques used today. In general, neural networks are used when the exact nature of the relationship between inputs and outputs is not known (if the relationship was known, the system could be modeled directly).

There are two types of neural networks depending on the training used: supervised and unsupervised neural networks. Thus, supervised training involves the use of the inputs to come up with an output that can be compared to the given output. On the other hand, unsupervised learning by way of neural network training is unique in that the network is given a set of inputs but no indication of what the output should be. The goal, then, is to have the network itself begin to organize and use those inputs to modify the weights of its own neurons.

We used unsupervised neural networks in our datamining process due to the conditions of our problem. Thus, we had a couple of tables about customers in which we wanted to distinguish users with non-technical losses from users with a normal consumption. It was thus not possible carry out a supervised training since we could not be certain that customers were in the first case (we only had one set of files registered in some inspections). Besides, the total number of non-technical losses registered by the company for this data was very low (only 22 compared with the total number of 72,927 contracts –see Table 1–).

Specifically, we used Kohonen networks (whose structure and working are represented in Figure 2) which provide an objective way of clustering data by utilizing a self-organizing network of artificial neurons. The Kohonen network resembles statistical clustering algorithms as it is capable of finding intrinsic clustering in the input parameter space.

In order to carry out the clustering process, it was necessary to select an adequate set of inputs which would make it possible to characterize the patterns. Thus, after studying the different alternatives, we selected the following inputs as identifiers of customer consumption pattern:

- Maximum: the maximum value of the bills of the previous year.
- Minimum: the minimum value of the bills of the previous year.
- Average: the consumption average of the bills of previous year.

- Difference average: the difference between the average parameter of the customer and the mean of the average parameters of the analyzed customers.
- Consumption coefficient: (Maximum–Minimum) / amount of contract power
- Difference average for month N: difference between the consumption in month N and the consumption average for month N of the analyzed customers.
- Difference maximum for month N: difference between the consumption in month N and the consumption maximum for month N of the analyzed customers.

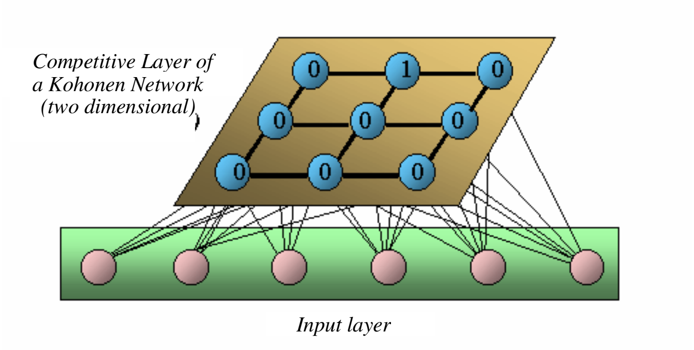

Fig. 2. Kohonen network structure

Once the input parameters had been selected, we designed a process for the detection of non-technical losses based on the search of similarities between the consumption pattern of the registered files and the consumption pattern of database customers. In short, first we carried out a clustering process of the all customers (including the customers with registered non-technical losses). Afterwards, we studied the different clusters in order to identify where the customers with non-technical losses were concentrated. So, in the clusters which were localized, the registered files were identified as possible customers with non-technical losses and we recommended the electrical company to carry out an inspection on them.

In case of hostelry business, we extracted for our analysis the 1,145 customers belonging to the interval of contracted power between 12 and 14 KW. We selected this interval in order to reduce the number of samples because 4 of the 5 files registered by the company belonged to that group. In the case of the little town, we extracted the 46,081 customers of the interval between 12 and 14 KW because 14 of the 17 files of this set were located within that group.

Two Kohonen neural networks were designed for each of the groups studied (hostelry business in the province of Seville and private users in the little town, making for a total of 4 neural networks) in order to carry out the clustering process: The first one was for clustering using parameters which involved, for each customer, an annual calculus of his consumption: Maximum, Minimum, Average, Difference average and Consumption coefficient and the second one was for clustering using the parameters related to monthly consumption: Difference average for month *N* and Difference maximum for month *N*.

First, the data was clustered using the first network. Afterwards, in order to reduce the large groups and the number of customer to be inspected, the second network was applied to the data belonging to clusters where the files registered by the electrical company were concentrated. Finally, the resulting groups were studied, that is groups with registered files and, therefore, with customers having a similar consumption pattern.

The two neural networks had 12 neurons and a 4*3 structure and therefore generated 12 clusters for each data set. After training the first network in the case of hostelry business, the registered files were localized into two clusters. The first cluster (cluster a1) only had 40 customers whereas the second one (cluster *a2*) had 352. In the case of the little town, the registered files were also localized into two sets (clusters *b1* and *b2*) of 2,520 and 852 customers, respectively. The second neural network was trained on clusters *a2*, *b1* and *b2* (cluster *a1* only had 40 customers) and the registered files were concentrated into an single cluster in each case (clusters *c1*, *c2* and *c3*) of 48, 2,200 and 244 customers, respectively.

Finally, the users of clusters *a1*, *c1*, *c2* and *c3* were identified as customers with possible non-technical losses and we recommended the electrical company to do a specific study for each case and if necessary, to inspect them.

Table 2. Selected study cases using neural networks

Type of customer	**Contracts**	**Selected study cases**	**Selected rate**
Little town	46,081	2,444 (*2,200+244*)	5.03 %
Hostelry business	1,145	88 (*40+48*)	7.68 %

4 Data Mining with Statistical Techniques

Outliers are elements in a data set which are grossly different or inconsistent with the remaining data. The statistical method developed for the non-technical losses is based on the detection of outliers, and it provides a general methodology for obtaining a list of abnormal users using only the general customer databases as input. In electrical consumption, outliers can be caused by measurement error or by fraud in customer consumption. But, alternatively, outliers may be the result of inherent data variability. Thus, the detection of outliers and its analysis is an interesting datamining task.

The statistical approach to outlier detection implies the use of a distribution or probability model for the given data set and then identifies outliers with respect to the model using a divergence test. The application of this test requires knowledge of the data set parameters (such as the assumed data distribution), distribution parameters (such as the mean and variance) and, mainly, knowledge of the inherent data variability.

In short, the datamining process involves the following tasks: First, from the two work sets: in the province of Seville and private users in the little town. We normalized these samples erasing the temporary and the local components of the individual consumption. Thus, we considered the probability distribution of the transformed sample as Gaussian (for the normal operating condition). Afterwards, we calculated and adjusted the threshold of the sample variance. Finally, we used outliers to guide the inspections.

We developed this method working on the set of particular users in the little town, in which we extracted 105 customers with the same contracted power (4 KW) and the same yearly electric consumption (between 0 and 5000 KW).

On the set of selected customers, we carried out a set of calculus for the detection of non-technical losses. The method involved the following steps:

1. Given:
   - A data at a set of spatial location (different customers).
   - Several data acquisitions at each location but spaced in time. It is assumed that all the locations are sampled at the same time but are sampled several times.
2. The operating equation is defined as
3. Follows: *Data acquired* = *Dlt* where
   D is the current data point measurement.
   l is the location of the measurement (number of customers).
   t is the time of the measurement (this is the time at which all the data is recorded at all locations).
4. The next step is to obtain the average at each time across all locations. This is defined by equation:

$$At = \sum_{l=1}^{N} \frac{Dlt}{N}$$

   where
   At is the average of all data at time *t*, across all locations *l*
   N is the number of locations *1*
5. It can now be observed, considering the averages and their times, whether there is or not an effect on a change in time. This is something that cannot be seen during an analysis of variance, but which we may see here.
6. The following step was to obtain the differences comparing the data at each location to the average at that time in the following way:

$$Lt = Dlt - At$$

   where .*lt* is the difference between the data at each location *l* and the time *t*, average.
7. Now the average of the differences *lt* at each location across time needs to be obtained, that is:

$$\bar{\delta} l = \sum_{t=1}^{M} \delta lt \, / \, M$$

   where
   $\bar{\delta} l$ is the average of all. *lt* at location *l* across time *t*.
   M is the number of times averaged.

8. In the following step, it is necessary to obtain the differences, comparing each time difference *lt*, to its average at location *l*, as shown in equation:

$$\Delta lt = \delta lt - \overline{\delta l}$$

Thus, the values obtained are the residual electrical consumptions after the linear variations in time and space are averaged out.

9. The next step was to calculate the standard deviation associated to each customer which is used as a distribution parameter:

$$STD_{\Delta l} = \pm\sqrt{\frac{\sum_{t=1}^{6}(\Delta lt - \overline{\Delta l})^{2}}{6-1}}$$

where

$$\overline{\Delta l} = \sum_{t=1}^{6}\frac{\overline{\Delta lt}}{6}$$

10. Finally, we carried out an outlier analysis (inherent data variability). To do this, we estimated a threshold for *STD* calculated as the mean of $STD_{1..N}$ multiplied by a constant (1.96 corresponding to a level of significance of 0.05).

$$STD_{\Delta l} = \pm\sqrt{\frac{\sum_{t=1}^{6}(\Delta lt - \overline{\Delta l})^{2}}{6-1}}$$

Thus, plotting $\mathrm{STD}_{1..N}$ and the threshold (Figure 3), we found that 9 customers were outliers.

As mentioned, these outliers can be caused by measurement error or by fraud in customer consumption. But, alternatively, outliers may be the result of inherent data variability. Thus, the following task involved a careful study of these outliers by company staff specialized in non-technical loss detection.

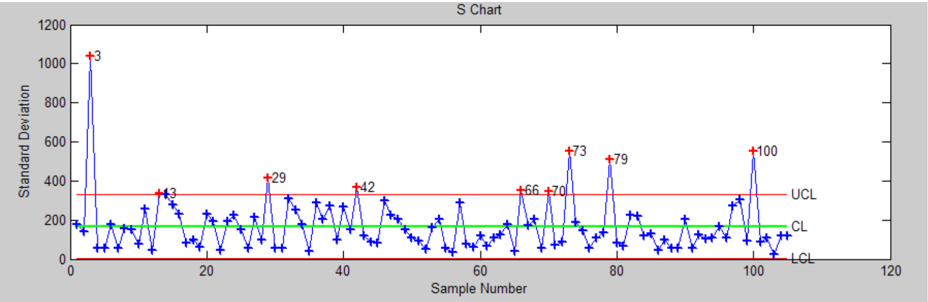

Fig. 3. An *STD* and threshold representation for the 105 selected customers

Table 3. Selected study cases using statistical techniques

Type of customer	Studied contracts	Selected study cases	Study rate
Little town	105	9	8.57%
Hostelry business	4,047	440	10.87%

5 Results and Conclusions

Once the datamining processes were carried out, the data on the customers detected by both methodologies (neural networks and statistical techniques) as possible customers with non-technical losses, was sent to the electrical company. Thus, some of these cases (selected in order of importance for the company since there was a large number of cases and their study required considerable time) were studied individually and in detail by specialized staff of the company which selected a set of these to be inspected in-situ. The rates of inspections as well as their results are shown in following tables:

Table 4. Datamining results (Neural Networks)

Type of customer	Selected study cases	Cases studied by the company	Cases inspected by the company	Non-technical losses detected
Little town	2,200	5	1	1
Hostelry business	89	27	6	3

Table 5. Datamining results (Statistical Techniques)

Type of customer	Selected study cases	Cases studied by the company	Cases inspected by the company	Non-technical losses detected
Little town	9 (of 105)	6	2	1
Hostelry business	440 (of 4047)	35	15	8

As may be observed in Tables 4 and 5, both methodologies detected cases of non-technical losses: 13 in total. The success rate from the inspections carried out by the company was around 50%. This represents an excellent rate, taking into account that up to that moment the company had carried out the study of a very large number of customers without any previous filtering.

In addition, both methodologies are general and not bound to a particular set or customer type. The whole input information needed is taken exclusively from the general customer database and is currently being integrated into a global system. Finally, we can enumerate three important conclusions for the work described in this paper:

1. We have developed two possible methodologies for the detection of non-technical losses (therefore, including cases of fraud): one by means of neural networks and the other by means of statistical techniques.
2. The company has tested the validity of both (by means of individual studies of the cases detected and by selective inspections in-situ). The resulting success rate of the inspections was around 50%.
3. The work described in this paper is a worldwide original work due to the fact that there is very little research on detection of non-technical losses and fraud detection in electrical consumption.

A possible line of work in the future might be the application of different and more-complex input parameters or other datamining techniques as well as the integration of human expert knowledge in these new techniques in order to improve the results. Therefore, this work is likely to be continued and, in fact, the Electronic Technology Department of the University of Seville and Endesa company are planning a continuation of the studies.

Acknowledgments

We would like to thank the initiative and collaboration of Endesa Company in the person of Mr. Juan Ignacio Cuesta.

References

1. M. Kantardzic, "Data Mining: Concepts, Models, Methods and Algorithms", Ed. IEEE Press,2003.
2. G. Piatetski-Shapiro, W.J. Frawley, "Knowledge discovery in databases", Ed. AAAI/MIT Press, 1991.
3. Yufeng Kou, Chang-Tien Lu, Sirirat Sinvongwattana, Yo-Ping Huang, "Survey of Fraud Detection Techniques", Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control, Taiwan, March 21, 2004.
4. J.R. Galván, A. Elices, A. Muñoz, T. Czernichow, M.A. Sanz-Bobi, "System for Detection of Abnormalities and Fraud in Customer Consumption", 12th Conference on the Electric Power Supply Industry, November,1998, Thailand.
5. R. Wheeler, S. Aitken, "Multiple Algorithms for Fraud Detection", Knowledge-Based Systems 13 (2000) 93–99
6. R. Richardson, "Neural Networks Compared to Statistical Techniques", Computational Intelligence for Financial Engineering (CIFEr), Proceedings IEEE/IAFE,1997.
7. S.Daskalaki, I.Kopanas ,M.Goudara, N.Avouris, "Data Mining For Decision Support on Customer Insolvency in the Telecommunications Business", European Journal of Operational Research 145 (2003), 239 –255.
8. José E. Cabral, Joäo Onofre P.Pinto, Edgar M.Gontijo, José Reis Filho, "Fraud Detection In Electrical Energy Consumers Using Rough Sets", 2004 IEEE Internacional Conference on Systems, Man and Cybernetics.