# FRAUD DETECTION IN ELECTRIC POWER DISTRIBUTION NETWORKS USING AN ANN-BASED KNOWLEDGE-DISCOVERY PROCESS

Breno C. Costa, Bruno. L. A. Alberto, André M. Portela, W. Maduro, Esdras O. Eler

PDITec, Belo Horizonte, Brazil – www.pditec.com.br

## *ABSTRACT*

*Nowadays the electric utilities have to handle problems with the non-technical losses caused by frauds and thefts committed by some of their consumers. In order to minimize this, some methodologies have been created to perform the detection of consumers that might be fraudsters. In this context, the use of classification techniques can improve the hit rate of the fraud detection and increase the financial income. This paper proposes the use of the knowledge-discovery in databases process based on artificial neural networks applied to the classifying process of consumers to be inspected. An experiment performed in a Brazilian electric power distribution company indicated an improvement of over 50% of the proposed approach if compared to the previous methods used by that company.*

## *KEYWORDS*

*Fraud Detection, Electric Power Distribution, Low-Voltage, KDD, Artificial Neural Networks*

## 1. INTRODUCTION

The losses in electric power distribution networks are a common reality for the electric utilities and can be classified as technical or non-technical losses, where the technical loss corresponds to the electrical energy dissipated between the energy supply and the delivery points of the consumer. On the other hand, a non-technical loss is the difference between the total losses and the technical losses, i.e. all other losses associated to the electric power distribution, such as energy theft, metering errors, errors in the billing process, etc. This loss type is directly related to the distributor's commercial management [1].

One of the usual ways to decrease non-technical losses rates is to perform local inspections to check if there are any thefts or hoaxes being committed by the consumers. For that purpose, the field staff is responsible for creating a methodology to generate an inspection schedule with suspected consumers. However, this task can be a too complex on due to the large amount of existing consumers in an electric power distribution network. Whenever a field staff performs an inspection, it returns only two possible outcomes: consumer is fraudster (there is irregularity) or non-fraudster (there is no irregularity).

Therefore, whenever a fraud is found, appropriate measures are applied and the situation returns to normal. Nevertheless, if a fraud is not found some costs such the inspector's hours and vehicles costs are unnecessarily spent, instead of being applied to inspect others consumers that are committing energy frauds. Increasing the hit rate of identifying irregularities is needed to reduce

DOI : 10.5121/ijaia.2013.4602

costs in the fraud detection, saving operational costs, even more importantly, identifying the fraudsters of the distribution network.

Some studies address this kind of problem applying well known knowledge-discovery techniques. Cabral et al. [2] [4] proposed a fraud detector for low and high voltage consumers. At that study, a knowledge-discovery methodology with artificial intelligence for data preprocessing and mining was successfully applied in different scenarios. Monedero et al. [3] developed an approach to identify non-technical losses in electric power distribution networks using artificial neural networks applied to the city of Seville (Spain), outperforming by over 50% precision compared to the previous methodologies.

Muniz et al. [5] presented an approach combining artificial neural networks and neuron-fuzzy systems to identify irregularities of the low voltage consumers. J. Nagi et al. [1] proposed a framework to detect non-technical losses in power distribution companies using pattern recognition by means of Support Vector Machines (SVM). E. W. S. dos Angelos [8] presented a cluster-based classification strategy with an unsupervised algorithm of two steps to identify suspected profiles of power consumption, providing a good assertiveness in real life systems. Over the past years, other studies in this field have been addressed applying different computational techniques to improve the detection of non-technical losses [9-13].

Although some studies have been conducted in the recent years in this field, the fraud rate in Brazilian low-voltage consumers is still very high, especially in metropolitan areas. There are still many electric power companies using only human expert knowledge to treat this problem and the results has been unsatisfactory. This paper addresses the fraud detection problem in electric power distribution networks (low-voltage consumers). Our methodology is based on the knowledge-discovery process using artificial neural networks to classify consumers as fraudsters or non-fraudsters. Thereafter a list of consumers classified as fraudsters is generated to help perform the inspections with the main objective being the improvement of the hit rate of inspections to reduce unnecessary operational costs.

This paper is structured as follows: the Section 2 presents the methodology used in the study. In the Section 3, the experiment and results are shown, and Section 4 presents the final discussions and conclusions.

## 2. METHODOLOGY

Our methodology is based on knowledge-discovery in databases (KDD) process using data mining, commonly used to extract previously unknown, non-trivial, and useful patterns from different data sources. The steps are presented in the Figure 1.

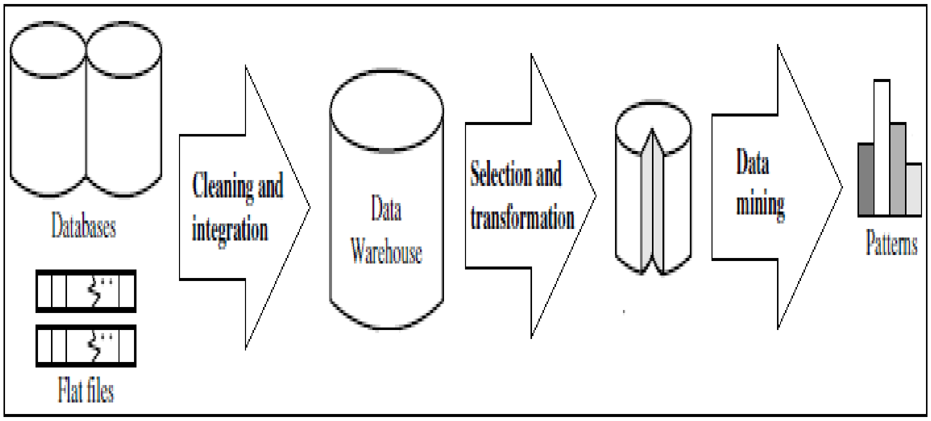

Figure 1: KDD process [6]

This process contains three main steps: cleaning and integration, selection and transformation, and data mining. In this case, they were applied to select consumers for inspection against consumption irregularities in electric power distribution networks. The development is presented in the next subsections.

## 2.1. Cleaning and Integration

In this step, different data sources obtained from databases and text files were processed to generate a single and consistent database. Initially, seven data sources were considered: (1) database of all consumers and their socio-economic characteristics; (2) history of inspections; (3) historical consumption; (4) history of services requested by clients; (5) history of ownership exchanges; (6) history of queries debits; and (7) history of meter reading.

All data sources were integrated using a common key: consumer installation code. Thus, it was possible to remove incorrect or duplicate records, generating a single database from the seven sources, used as input for the data mining algorithm (training and classification). In this context, only the records with valid inspection results (fraudsters or non-fraudsters) were considered.

## 2.2. Selection and Transformation

In this step, statistical techniques were applied for data selection and transformation. The attribute selection was performed using a multivariate correlation analysis and Information Gain method, generating some key attributes of the model as shown in the Table 1.

Table 1: Key attributes of the model

| # | Name | Type | Description |
|---|---|---|---|
| 1 | Location | nominal | Consumer location (city + neighborhood) |
| 2 | Business class | nominal | Business class (e.g. residential, industrial, commercial, among others) |
| 3 | Activity type | nominal | Activity type (e.g. residence, drugstore, bakery, public administration, among others) |
| 4 | Voltage | nominal | Consumer voltage (110v, 220v) |
| 5 | Number of phases | nominal | Number of phases (1, 2, 3) |
| 6 | Situation | nominal | Is consumer connected? (yes, no) |
| 7 | Direct debit | nominal | Type of direct debit |
| 8 | Metering type | nominal | Type of electricity metering |
| 9 | Mean consumption | numeric | Mean consumption during the previous 12 months |
| 10 | Service notes | nominal | Are there service notes requested by consumers during the previous 12 months? (Yes or no) |
| 11 | Ownership exchange | nominal | Are there ownership exchanges during the previous 12 months? (Yes or no) |
| 12 | Query debits | nominal | Are there query debits during the previous 12 months? (Yes or no) |
| 13 | Meter reading | nominal | Are there meter reading notes during the previous 12 months? (Yes or no) |
| 14 | Inspection (output) | nominal | The inspection result (Fraudster or Non-fraudster) |

The data transformation was performed by normalization methods. The Min-Max method was used to transform the single numerical attribute Mean Consumption, wherein each value is converted to a range between 0 and 1. For the nominal attributes, two distinct methods were applied: if the number of distinct values was less than five, the Binary method was applied; otherwise the One-of-N method was used.

## 2.3. Data Mining

The problem addressed here is a supervised classification one because there are real inspections' records to be used for the training step. An artificial neural network/multilayer perceptron (ANN-MLP) was used for the dataset training and classification. According to Haykin [7], the MLPs using the Backpropagation algorithm have been successfully applied to solve many similar complex problems, such as pattern recognition, classification, data preprocessing, etc. Therefore, we applied it to solve our problem with the configurations that are presented here.

The ANN-MLP's architecture used three layers: input, hidden and output layers. The input layer contains *n* neurons, wherein *n* is equal to number of bits of the normalized values. In the hidden layer, the number of neurons is equal to the geometrical mean between the numbers of neurons in the input and output layers [14]. The output layer contains a single neuron to classify the consumers as fraudsters or non-fraudsters.

After defining the architecture, the data mining algorithm was coded using the training/test/validation schema and the stratified *k*-Fold Cross Validation method. In this case, *k* is equal to 10, i.e., at the each iteration the dataset is randomly partitioned into 10 equal size subsets. Of the 10 subsets, 9 are used as training data and a single is used for testing the model. This process is then repeated *k* times, and the average error is calculated. This implementation requires parameter settings to perform the experiments successfully.

## 3. RESULTS AND DISCUSSION

The evaluation of the results was performed using a confusion matrix that compares real inspections to classified inspections by the ANN-MLP, indicating four result types: (1) TP is a fraudster consumer correctly classified as fraudster; (2) FN is a fraudster consumer incorrectly classified as non-fraudster; (3) FP is a consumer non-fraudster incorrectly classified as fraudster; (4) TN is a consumer non-fraudster correctly classified as non-fraudster.

Some measures are used to check the classifier efficiency: hit rate is the percentage of records correctly classified; recall: TP / (TP+FN); precision = TP / (TP+FP); True Positive rate = TP / (TP+FN); and False Positive rate = FP / (FP+TN).

This approach was evaluated using real data from a Brazilian electric power distribution company with more than seven million consumers. A lot of inspections conducted in the metropolitan region of Minas Gerais (Brazil) during the year 2011 were used to supply the KDD process.

The first step of our methodology was applied to clean and integrate databases and text files, generating a single data source with 22,848 records. After that, the second step was applied to select key attributes of the model and transform data to feed the ANN-MLP. Each record consisted of thirteen input attributes and one output attribute indicating fraud or not, as seen in the Section 2.2.

After that, the ANN-MLP was trained with Backpropagation learning algorithm using the Logistic activation function for all neurons. The training step was performed for 500 epochs, with learning rate equals to 0.01, momentum factor equals to 0.95, activation threshold equals to 0.5, and maximum error equals to 0.001. The parameter settings were empirically defined.

The algorithm was performed in 903 seconds on a personal computer and the Table 2 shows the classification results.

Table 2: Confusion matrix with the classification results

| | **Fraudster (a)** | **Non-fraudster (b)** |
|---|---|---|
| **Classified as (a)** | 945 | 508 |
| **Classified as (b)** | 2261 | 17869 |

The classification results were obtained with an error rate of 0.3562 after all epochs, and the performance measures were calculated from the confusion matrix. The accuracy is equal to 87.17%, precision is equal to 65.03% and recall is equal to 29.47%.

After obtaining the classification results, a list of inspections is generated to check if consumers are fraudsters. The precision indicates the percentage of consumers correctly inspected by the staff team. In this context, increasing of hit rate is essential to improve the number of fraudsters found, and reducing the number of non-fraudsters inspected. This directly implies two consequences: recovery of financial income and reducing of unnecessary operational costs.
According to the field measurement, currently the same electric power distribution company has obtained a precision around 40% using different methods. Our methodology improved this number to 65.03%, outperforming in 50% the current scenario. In other words, as previously seen, we obtained 22,848 records of inspections from the year 2011, i.e. only 9,139 consumers were correctly inspected by the company. At the same time, our methodology could obtain a precision of 14,858 consumers – a difference of 5719 per year.

That difference of fraudsters consumers found represents the recovery of financial income in three senses: (1) the recovery of value defrauded applying fines according to the fraud period, (2) normalization of the consumers' status and the regularization of their consumption, and (3) unnecessary expenses in incorrect inspections were avoided by reducing expenses with employees, transportation costs and fuel consumption.

Another important point is the use of the k-Fold Cross Validation method, whose goal is to prevent the ANN-MLP of over-fitting the training data. This allows the learning to be used for the classification of consumers outside the training set (the real world case).

Although the algorithm used the stratified validation method, the unbalanced class problem is a weakness that must be better handled, since the learning systems can find difficulties in the classification of records belonging to minority class when the class distribution is unequal. This is a need for the next steps of development of our methodology.

## 4. CONCLUSION

This paper proposed the use of the KDD process for fraud detection in the power distribution network (low-voltage consumers). For this purpose, a methodology with data preprocessing and mining using artificial neural networks was applied. The results of the experiment indicated that is possible to improve the hit rate obtained by the methods used currently for the classification of fraudsters.

The experiment was performed using only historical data and a deeper evaluation in field is needed to ensure the efficiency of our approach. Although this has not been done, we used the k-Fold Cross Validation method for assessing how the results will generalize to an independent data set. Furthermore, we intend on testing different supervised classification techniques and compare theirs results.

## REFERENCES

[1] J. Nagi et al., (2010) “Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines”, IEEE Transactions on Power Delivery, Vol. 25, N. 2, Apr 2010.

[2] Cabral et al., (2004) “Fraud Detection in Electrical Energy Consumers Using Rough Sets”, 2004 IEEE International Conference on Systems, Man and Cybernetics: The Hague, Netherlands, Oct 10-13.

[3] Monedero et al., (2006) “MIDAS Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques”, In proceeding of: Computational Science and Its Applications - ICCSA 2006, International Conference, Glasgow, UK, May 8-11.

[4] Cabral et al., (2009) “Fraud Detection System for High and Low Voltage Electricity Consumers Based on Data Mining”, Power and Energy Society General Meeting, Calgary, Jul 26-30.

[5] Muniz et al., (2009) “A Neuron-fuzzy System for Fraud Detection in Electricity Distribution”, Proceedings of the IFSA-EUSFLAT 2009, Lisbon, Portugal, Jul 20-24.

[6] J. Han, M. Kamber, J. Pei, (2011) “Data mining: concepts and techniques”, Morgan Kaufmann, 3rd ed.

[7] S. Haykin, (1999) “Neural Networks: A Comprehensive Foundation”, Prentice Hall International, 2nd ed.

[8] E. W. S. dos Angelos et al., (2011) “Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems”, IEEE Transactions on Power Delivery, Vol. 26, N. 4, Oct 2011.

[9] C. C. O. Ramos et al., (2012) “New Insights on Nontechnical Losses Characterization Through Evolutionary-Based Feature Selection”, IEEE Transactions on Power Delivery, Vol. 27, N. 1, Jan 2012.

[10] A. H. Nizar, Z. Y. Dong, Y. Wang, (2008) "Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method", IEEE Transactions on Power Systems, Vol. 23, N. 3, pp.946-955, Aug 2008.

[11] J. Nagi et al., (2008) "Detection of abnormalities and electricity theft using genetic Support Vector Machines" TENCON 2008 - 2008 IEEE Region 10 Conference, pp.1-6, Nov 2008.

[12] J. Nagi et al., (2008) "Non-Technical Loss analysis for detection of electricity theft using support vector machines", IEEE 2nd International Power and Energy Conference, pp.907-912, 1-3 Dec 2008.

[13] A. C. Ramos et al., (2011) "A New Approach for Nontechnical Losses Detection Based on Optimum-Path Forest", IEEE Transactions on Power Systems, Vol. 26, N. 1, pp.181-189, Feb 2011.

[14] M. Negnevitsky, “Artificial Intelligence: A Guide to Intelligent Systems”, Pearson Education Canada, 3rd ed.