Abnormal consumption analysis for fraud detection: UTE-UdelaR joint efforts

Juan Pablo Kosut
Fernando Santomauro
Andrés Jorysz
UTE, STC Montevideo, Unidad de Recuperación de Energía
Montevideo, Uruguay
jkosut@ute.com.uy

Alicia Fernández
Federico Lecumberry
Fernanda Rodríguez
UdelaR, Facultad de Ingeniería
Montevideo, Uruguay
{alicia,fefo,frodriguez}@fing.edu.uy

***Abstract*—Within the framework of the Energy Recovery Unit of the Technical-Commercial Service of Montevideo, UTE, for the reduction of Non Technical Losses, a research project was carried out jointly with the Institute of Electrical Engineering of UDELAR. The project had the aim of designing different strategies of automatic classification that separate normal consumption measurements from abnormal ones which represent clues of possible sources of Non Technical Losses. Different classifiers were implemented and several field tests were conducted, with promising results. Several criteria for the incorporation of new features are proposed in this work. These criteria are complementary to those derived from consumptions. An analysis of the performance of said features was conducted, showing that improving classifier performance is possible with this method.**

Index Terms– Unbalance Class Problem, Combining Classifier, Feature Selection, Performance Measurement

I. Introduction

Irregular or fraudulent use of electrical power represents a problem of great proportions, causing substantial economic losses to distribution companies in several countries. In the area of electricity, Nontechnical Losses (NTL) are defined as non-billed consumption, whether it be due to flawed equipment, billing errors, fraudulent manipulation (fraud) or direct connections to the grid (theft). In short, NTL can be defined as total losses other than technical losses (resulting from dissipation of components in the grid). In Uruguay, there are two areas that differ from the rest of the country, where high values of NTL are observed. These are Montevideo and the Center. In the case of Montevideo, total losses constitute a 19.1% according to December 2014 balances, while those of the center of the country are of an 18.6%. Country-wise, total distribution losses amount to a 15.8%. Due to the fact that UTE is in possession of information that allows for the estimation of losses linked to deprived neighborhoods, as well as calculating technical losses, it is possible to identify, within NTL, losses hereinafter referred to as fraud. The aim of the tools developed in this project is to reduce fraud-related losses, which, according to what has been stated above, are estimated in 3% in Montevideo and 6% in the central area.

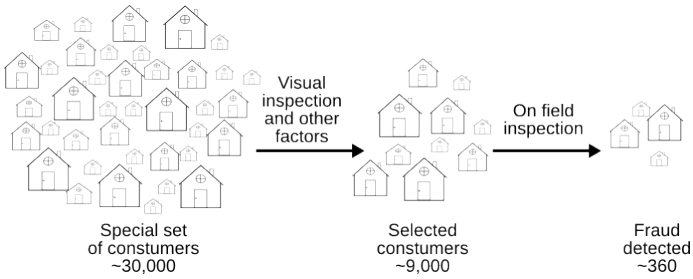

Fig. 1. Manual fraud detection scheme

II. Related works

Different machine learning aproaches have addressed the detection of non-technical losses, both supervised or unsupervised. Leon et al. review the main research works found in the area between 1990 and 2008 [1]. Here we present a brief review that builds on this work and wide it with new contributions published between 2008 and 2013. Several of these approaches consider unsupervised classification using different techniques such as fuzzy clustering [2], neural networks [8], [3], among others. Monedero et al. use regression based on the correlation between time and monthly consumption, looking for significant drops in consumption [9]. Then they make a second stage where suspicious customers are eliminated if the consumption of these depend on the economy of the moment or the year's season. Only major customers were inspected and 38% were detected as fraudulent. Similar results (40%) were obtained in [10] using a tree classifier and customers who had been inspected in the past year. In [11] and [12] SVM is used. In the latter, Modified Genetic Algorithm is employed to find the best parameters of SVM. In [4], is compared the methods Back-Propagation Neural Network (BPNN), Online-sequential Extreme Learning Machine (OS-ELM) and SVM. Biscarri et al. [5] seek for

978-1-4673-6605-2/15/$31.00 ©2015 IEEE

outliers, Leon et al. [1] use Generalized Rule Induction and Di Martino et al. [13] combine CS-SVM classifiers, One class SVM, and C4.5 OPF using various features derived from the consumption.

Different kinds of features are used among this works, for examples, consumption [5], [12], contracted power and consumed ratio [14], Wavelet transformation of the monthly consumption [15], amount of inspections made to each client in one period and average power of the area where the customer resides [2], among others.

On the other hand, Romero proposes [16] a method to estimate and reduce non-technical losses, such as advanced metering infrastructure, fraud deterrence prepayment systems, system remote connection and disconnection, etc. Lo et al. based on real-time measurements, design [17] an algorithm for distributed state estimation in order to detect irregularities in consumption.

III. Managing Technical Losses

A. Conventional Procedures

In general, distribution companies face the problem of fraud when conducting inspections where energy meters are checked. The procedure requires planning, due to the great expenses it implies, which include transportation, training and expert hiring fees. This is why conducting a previous selection of abnormal consumption is of such importance; it helps reducing field inspections, avoiding the expenses these inspections imply. In Uruguay, both offices, Montevideo and Centro (downtown) have specific procedures for conducting a selection of suspicious clients. The heuristics selected are strongly conditioned by the technical resources allocated to the task, and by the specific characteristics of each region. In Montevideo, this process requires several stages, filtering, data pre-processing and processing using a decision rule for further inspection. In each stage, all the information available to the company is used. Particularly, some decision rules require manual analysis of data, such as visual inspection of consumption series. This procedure consists in a visual inspection of a series of three years or more, over which events such as titleholder changes, meter changes, cut offs due to failure to pay and other field activities, prior irregularities and trendlines are graphed. Other information such as area, fees and geographic location is used. When conducting a visual analysis, several elements that may indicate irregularities are looked for, e.g., consumption reductions, non-consistent seasonal variation, non-consistent variability and non-consistent levels of consumption. In Figure 2 it can be noted, for instance, that consumption levels increase after conducting adaptation works of measuring sites (activity code 639) and subsequently decrease abruptly, and no longer suffer seasonal variations.

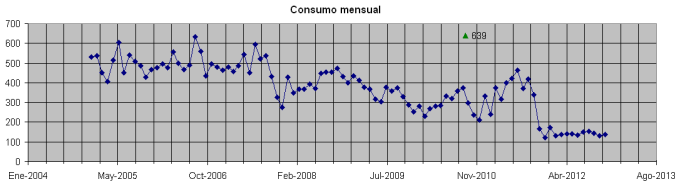

Fig. 2. Example of consumption with irregularities.

B. Problem Complexity

An example of irregular consumption is shown in Figure 2. Said irregular consumption, when inspected on site, allowed for the detection of a concrete fraudulent maneuver regarding the energy meter. Anyhow, the extraordinary number of clients, the variability in consumption modes, the existence of a great variety of frauds and forms of measurement alterations make consumption classification for abnormalities identification a very challenging issue. The manual procedure that is described, however showing good results, particularly when conducted by experienced personnel, requires too many technical hours, and its efficiency diminishes as the volume of cases to be processed rises excessively.

C. Automatic Classification

The problem described herein gives rise to the search for automatic classification tools that serve the purpose of detecting abnormal cases containing sings of possible fraud. Within the area of pattern recognition the case may be considered a classification problem with unbalanced classes, having a class of normal cases and one of abnormal cases, the latter being much less frequent than the former. The problem of unbalance must be approached in a particular way in order to stop classifiers from favoring the majority class, with the corresponding subsequent sub-optimal performance.

IV. UTE-UDELAR Cooperation

Kosut and Alcetegaray [6], UTE technicians in the area of nontechnical losses management, projected the possibility of incorporating pattern recognition tools to the analysis of consumption data. They took the course at the Instituto de Ingeniería Eléctrica (Institute of Electrical Engineering or IIE) in UDELAR and, within its framework, they proposed using One Class SVM technique in order to detect abnormal consumptions based on historical energy consumption series. Positive results were obtained when applying the classifier to a sample of commercial services in the area of grocery stores and supermarkets. Labels defined by UTE technicians through visual inspection of consumption curves were used for the training and validation stages. The team managed to demonstrate that the technique could reproduce manual classification correctly.

Decia, Di Martino and Molinelli [13] within the framework of their undergraduate thesis, added Cost Sensitive SVM, Optimum

Path Forest and C4.5 decision tree classification techniques, broadening databases used for training and validation, and proposing the use of other features, also based in historical energy consumption series. They analyzed the problem of class unbalance and proposed the use of F_{value} as an indicator for algorithm optimization.

$$F_{value} = \frac{(1+\beta^2)\text{Recall} \times \text{Precision}}{\beta^2 \text{Recall} + \text{Precision}}$$

Features used by classifiers are built upon historical energy consumption series. In [18] 28 features are proposed. Some examples are displayed below:

- Average consumption ratio compared to the average of the last 3, 6, and 12 months (features 1,2 y 3)
- Euclidean distance between each consumer and the average consumer (feature 20).
- Differences between ratios of curve polynomial approximation coefficients.
- Consumption variance.
- Straight line slope approximating consumptions (feature 28).

In Figure 3 some consumption-series-based features are displayed graphically.

Between June 2012 and June 2014 an UTE-UDELAR joint project was carried out, within the framework of the Programa de Vinculación Universidad Sociedad y Producción de CSIC (Comisión Sectorial de Investigación Científica). An exhaustive bibliographic revision was conducted on the subject. There was also extensive work conducted with relation to the identification of relevant qualities for the detection of suspicious records, together with the processing of different databases in order to validate algorithms with real data. Moreover, different approaches regarding data labeling were compared, new algorithms were proposed and analyzed and different field tests were carried out.

Intrioni and Lema [7], in their 2011 pattern recognition course, tried to identify clusters within the universe of consumers, analyzing whether performance improvements are accomplished when using said consumers in combination with defined classifiers.

Di Martino et al. [19], [20] proposed new classification techniques, using F_{value} as target measurement for the optimization of algorithms as a way of attacking the class unbalance problem.

Tacón et al. [21] proposed a semi-supervised approach for the detection of abnormal consumptions.

Rodriguez, Lecumberry and Fernandez [22] analyzed the impact different labeling strategies had in classifier performance.

V. Addition of New Features

So far, extensive investigation, validation and field test works have been conducted on different algorithms using features based in consumption curves.

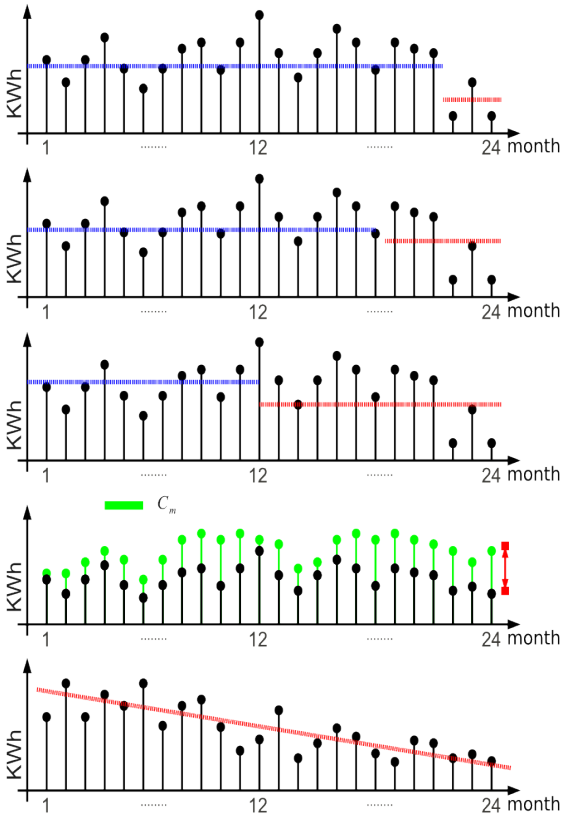

Fig. 3. Representation of features 1, 2, 3, 20 and 28

Moving forward towards finding new features that may be of use for the detection of electrical energy fraud is of utmost importance.

This goal is related with the difficulty regarding the capacity that algorithms shall have for managing new features, especially nominal features. The difficulty arising from the generation of features, based on the information registered in computer systems in UTE, is also significant.

A. Generating New Features

We began by working with a list of 30 new features that could potentially be added, which, according to the opinion of UTE experts, could affect the classification. Said features are briefly described in table I

Those features that count days since an event (8, 18, 21, 24 and 27) have the same information as those that indicate the date of the event itself (7, 17, 20, 23 and 26). Hence, the latter were discarded. In addition, some features (2, 5, 6, 8, 16 and 18) work as a base for the preprocessing of information, and are used in order to generate different previously identified clusters or client groups, which are to be analyzed separately by the classifiers,

TABLE I
PROPOSED NEW FEATURES

Cod	Feature	Description
01	Power Paid	Maximum power usable by the client
02	Strategic	Indicates if the client can be considered strategic
03	Grouping	Indicates if the client's meter is located in a centralization
04	Area	Indicates the geographical location of the client
05	Contract Status	Contract status: active, inactive
06	Period Recontract	Indicates whether new contracts exist or not in the analyzed period
07	Modification Date	Contract modification date
08	Days Since Recontract	Number of days passed since the execution of the new contract
09	Re 1570	Client deemed as of relatively low income
10	Three-phase	Three-phase energy supply
11	Fee	Code for the fee assigned to the client
12	Type of Client	Indicates the type of client
13	Type of Device	Indicates the type of device installed
14	Sgalum	General lighting service
15	Area	Area occupied by the estate/business
16	With Irregularities	Indicates whether there is a background of irregularities
17	Last Irregularity Date	Indicates the date of the last irregularity
18	Days Since Last Irregularity	Days passed since the last irregularity
19	Number of Irregularities	Number of irregularities
20	Prior Irregularity Date	Date of prior irregularity
21	Days Since Prior Irregularity Date	Days passed since prior irregularity
22	Existing Inspections in the Period	Indicates if the site was inspected during the considered period
23	Last Inspection Date	Indicates the date of the last inspection
24	Days Since Last Inspection	Days passed since the last inspection
25	Update	Indicates whether the meter site has been updated and placed in the property line of the estate
26	Update Date	Update date
27	Days Since Update	Days since update
28	Device ID	Identification number of the meter installed
29	Actual Reading Proportion	Frequency rate of actual readings
30	Default	Default of payment frequency rate

TABLE II
NEW FEATURES SELECTED FOR USE

Cod	Feature	Description
01	Actual Readings Proportion	Ratio between readings carried out by UTE employees over total readings. (Other kinds of readings can be submitted by the client or calculated by the system)
02	Power Paid For	Maximum power usable by the client, limited by a thermal switch
03	Number of Irregularities	Number of irregularities recorded for the period studied herein
04	Days Since Last Inspection	Days passed since the last field activity comprised in the PNT management framework
05	Days Since Update	Days passed since the realization of meter improvement works, after which the meter is located outside the estate, where it can be accessed by UTE controls at all times.
06	Default	Ratio between the number of months paid in arrears regarding the expiry date of the bill over the total number of months of the period.

TABLE III
ADDITION OF NEW FEATURES

Classif	Feat	New Feat	$Accu$	Rec	$Prec$	F_{value}
CS-SVM	1 2 3 12 20 28	-	82%	52%	24%	32%
Tree	1 3 4 10 12 13 23 14	-	73%	55%	17%	26%
Tree	3 4 12 19 24	2 3 5 6	88%	43%	35%	39%

using a similar strategy to that used in [7]. A lack of data needed for some features was detected. In consequence, such features had to be discarded. Pursuant to the above mentioned information, the team decided to take into account 6 features for analysis: (1, 19, 24, 27, 29 and 30)

For the sake of clarity, those features selected for the evaluation of a possible improvement in classifier performance are displayed again in table II.

B. Selecting Features

Different selection techniques were applied to the total group of features, given that, as it is widely known, finding a smaller subgroup with relevant features can improve classifier performance. Besides, in this case it is of utmost important to assess whether the selection methods used select some of the proposed new features as relevant.

The selection methods used were Filter and Wrapper. The Filter method works by searching the feature subgroup that presents the highest correlation levels between classes, as well as the lowest correlation levels possible between features. By conducting an exhaustive search through all of the size subgroups, the one with the best performance is selected. The Wrapper method is a supervised method, which takes into account the specific classifier that will later on be used during the classification stage. In this case, F_{value} was used as performance measurement, and the Best First method was used as search criteria. This resulted in a good balance between performance and computer costs.

C. Obtained Results

The C4.5 decision tree was the strategy selected for the evaluation of new feature performance, due to the fact that decision trees are especially suitable for the use of nominal and continuous features. The decision tree is implemented jointly with the Adaboost technique in order to alleviate the instability problem which is inherent to trees. In order to take into account the problem of class unbalance, different methods of random sub-sampling of the majority class, classifier cost attribution and definition of F_{value} as performance measurement are used.

The results obtained are shown in table III. CS-SVM classifier performance is also added, serving as a reference, given that it has been found to be one of the classifiers with the best performance in previous projects.

In the table, a significant improvement can be observed with regard to the performance of the C4.5 decision tree, as a result of the addition of new features.

VI. Conclusions

A summary of the results obtained in the research project carried out by UTE and the Instituto de Ingeniería Eléctrica of UDELAR has been presented herein. The joint efforts of the university team and its UTE counterpart were very productive, and enabled the generation and transfer of knowledge between both teams. The approach taken regarding the subject matter gave way to the generation of publishable articles, undergraduate thesis and other works in post-graduate courses. The project accomplished the creation of a Framework including the possibility of training a group of classification algorithms, and then using them in order to select the inspection plan for the management of Nontechnical Losses. Besides, the proposal of generation and addition of new features was further developed. It was demonstrated improving classifier performance is possible with this method.

VII. Acknowledgments

This work was partially supported by “Programa VUSP de CSIC UdelaR-UTE”.

Periodicals

[1] C. Leon, F. X. E. L. Biscarri, I. X. F. I. Monedero, J. I. Guerrero, J. X. F. S. Biscarri, and R. X. E. O. Millan, “Variability and trend-based generalized rule induction model to ntl detection in power companies,” *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 1798–1807, 2011.

[2] E. dos Angelos, O. Saavedra, O. Cortes, and A. De Souza, “Detection and identification of abnormalities in customer consumptions in power distribution systems,” *IEEE Transactions on Power Delivery*, vol. 26, no. 4, pp. 2436–2442, 2011.

[3] M. Sforna, “Data mining in power company customer database,” *Electrical Power System Research. London, U.K*, vol. 55, pp. 201–209, 2000.

[4] K. S. Yap, S. K. Tiong, J. Nagi, J. S. P. Koh, and F. Nagi, “Comparison of supervised learning techniques for non-technical loss detection in power utility,” *International Review on Computers and Software (I.RE.CO.S.)*, vol. 7, no. 2, pp. 1828–6003, 2012.

[5] F. Biscarri, I. Monedero, C. Leon, J. I. Guerrero, J. Biscarri, and R. Millan, *A data mining method based on the variability of the customer consumption - A special application on electric utility companies*. Inst. for Syst. and Technol. of Inf. Control and Commun., 2008, vol. AIDSS, pp. 370–374.

Unpublished Papers

[6] D. Alcetegaray and J. Kosut, “One class SVM para la detección de fraudes en el uso de energía eléctrica.” *Trabajo Final Curso de Reconocimiento de Patrones, Dictado por el IIE- Facultad de Ingeniería- UdelaR*, 2008.

[7] Intrioni and Lema, “Detección de clusters.” *Trabajo Final Curso de Reconocimiento de Patrones, Dictado por el IIE- Facultad de Ingeniería-UdelaR*, 2011.

Papers from Conference Proceedings (Published)

[8] Z. Markoc, N. Hlupic, and D. Basch, “Detection of suspicious patterns of energy consumption using neural network trained by generated samples,” *Proceedings of the ITI 2011 33rd International Conference on Information Technology Interfaces*, pp. 551–556, 2011.

[9] I. Monedero, F. Biscarri, C. Leon, J. Guerrero, J. Biscarri, and R. Millan, “Using regression analysis to identify patterns of non-technical losses on power utilities,” in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, R. Setchi, I. Jordanov, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2010, vol. 6276, pp. 410–419.

[10] J. Filho, E. Gontijo, A. Delaiba, E. Mazina, J. Cabral, and J. Pinto, “Fraud identification in electricity company customers using decision tree,” in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 4, Oct 2004, pp. 3730–3734 vol.4.

[11] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, “Support vector machine based data classification for detection of electricity theft,” *2011 IEEE/PES Power Systems Conference and Exposition*, pp. 1–8, 2011.

[12] K. S. Yap, Z. Hussien, and A. Mohamad, “Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm,” *3rd IASTED Int. Conf. Advances in Computer Science and Technology, Phuket, Thailand*, vol. 4, 2007.

[13] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, “Improving electric fraud detection using class imbalance strategies.” in *International Conference on Pattern Recognition and Methods, 1st. ICPRAM.*, 2012, pp. 135–141.

[14] J. Galván, E. Elices, A. M. Noz, T. Czernichow, and M. Sanz-Bobi, “System for detection of abnormalities and fraud in customer consumption,” *Proc. 12th IEEE/PES conf. Electric Power Supply Industry*, 1998.

[15] R. J. R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey, “Wavelet based feature extraction and multiple classifiers for electricity fraud detection,” *IEEE/PES Transmission and Distribution Conference and Exhibition*, vol. 3, 2002.

[16] J. Romero, “Improving the efficiency of power distribution system through technical and non-technical losses reduction,” *IEEE*, 2012.

[17] Y.-L. Lo, S.-C. Huang, and C.-N. Lu, “Non-technical loss detection using smart distribution network measurement data,” in *Innovative Smart Grid Technologies - Asia (ISGT Asia), 2012 IEEE*, 2012, pp. 1–5.

[18] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "A novel framework for nontechnical losses detection in electricity companies," in *Pattern Recognition - Applications and Methods*, ser. Advances in Intelligent Systems and Computing, P. Latorre Carmona, J. S. Sanchez, and A. L. Fred, Eds. Springer Berlin Heidelberg, 2013, vol. 204, pp. 109–120.

[19] M. Di Martino, A. Fernández, P. Iturralde, and F. Lecumberry, "Novel classifier scheme for imbalanced problems," *Pattern Recognition Letters*, vol. 34, no. 10, pp. 1146–1151, 2013.

[20] M. Di Martino, G. Hernández, M. Fiori, and A. Fernández, "A new framework for optimal classifier design," *Pattern Recognition*, vol. 46, no. 8, pp. 2249–2255, 2013.

[21] J. Tacón, D. Melgarejo, F. Rodríguez, F. Lecumberry, and A. Fernández, "Semisupervised approach to non technical losses detection," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2014, pp. 698–705.

[22] F. Rodríguez, F. Lecumberry, and A. Fernández, "Non technical loses detection - experts labels vs. inspection labels in the learning stage," in *ICPRAM 2014 - International Conference on Pattern Recognition Applications and Methods*, 2014, pp. 624–628. [Online]. Available: http://iie.fing.edu.uy/publicaciones/2014/RLF14