Semisupervised Approach to Non Technical Losses Detection

Juan Tacón, Damián Melgarejo, Fernanda Rodríguez,
Federico Lecumberry, and Alicia Fernández

Instituto de Ingeniería Eléctrica, Facultad de Ingeniería,
Universidad de la República, Uruguay
http://iie.fing.edu.uy

Abstract. Non-technical electrical losses detection is a complex task, with high economic impact. Due to the diversity and large number of consumption records, it is very important to find an efficient automatic method to detect the largest number of frauds with the least amount of experts' hours involved in preprocessing and inspections. This article analyzes the performance of a strategy based on a semisupervised method, that starting from a set of labeled data, extends this labels to unlabeled data, and then allows to detect new frauds at consumptions. Results show that the proposed framework, improves performance in terms of the $F_{measure}$ against manual methods performed by experts and previous supervised methods, avoiding hours of experts/inspection labeling.

Keywords: Electricity Fraud, Support Vector Machine, Semisupervised Approach, SVMlight, TSVM, Unbalance Class Problem.

1 Introduction

In the power market, electrical losses are increasingly being taken more into consideration due to the high economic impact generated. These can be separated in technical and non technical losses (NTL). The former ones are related to dissipation losses in the grid, either during transmission, voltage transformation, or energy measurement. On the other hand NTL involves energy that is transmitted, but is not billed, and essentially they are generated by faults or illegal manipulation on the side of the client. Inspecting customers on site implies a great economic cost, and this cost is not refunded through these inspections. In [1] is shown the procedure carried out in UTE[1], Uruguay. It is a manual procedure, that involves many hours of manual preprocessing, and results shows the posibility to improve the performance regarding true positive detections. In [2] Rodríguez et al. a comparative analysis of learning from experts labels and inspection labels is done using a supervised approach. In a more general review, there are several works with a Pattern Recognition approach that have addressed the detection of non technical losses (supervised or unsupervised). Leon et al.

[1] The national company (utility).

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 698–705, 2014.

review the main research works found in the area between 1990 and 2008 [3]. Since 2008, there are diverse contributions published. A few of these approaches consider unsupervised classification using different techniques such as fuzzy clustering (dos Angelos et al., 2011) [4], neural networks (Markoc et al., 2011 [5]; Sforna, 2000 [6]), among others. In (Depuru et al., 2011) [7] and (Yap et al., 2007) [8] Support Vector Machines (SVM) [9] is used. In (Yap et al., 2012) [10] are compared the methods Back-Propagation Neural Network (BPNN), Online-sequential Extreme Learning Machine (OS-ELM) and SVM. Di Martino et al. (Di Martino et al., 2012) [1] combine CS-SVM classifiers, One class SVM, and C4.5 OPF.

While these methods cited above utilize supervised or unsupervised techniques. The objective of this work is to address the problem from a complementary semi supervised approach using a variation of SVM, as well as implement feature selection, and compare results with previous works. There are multiple works that implement semi supervised methods, for example Transductive SVM (Joachims, 1999) [11], Large Scale Semisupervised Linear SVMs (Sindhwani et al., 2005) [12] and Gaussian Fields and Harmonic Functions (Zhu et al., 2003) [13] for text classification. They show that is efficient to use a semi supervised method in problems where is significantly more complicated and expensive to get labeled data, than unlabeled data. To make inspections on site, involves cost of technicians and transport for inspections of suspect customers. Then, the objective in a semisupervised approach, is to create a tool, that starting from a set of labeled data, extends this labels to unlabeled data, and then allows to detect new frauds at consumptions. In this paper we set out to analyze the behavior of the proposed semi supervised framework to fraud classification and compare it against manual methods performed by experts and previous supervised methods. Thus, as far as we know is a new way to approach this problem. The paper is organized as follows. Section 2 describes the semi supervised approach, Section 3 presents the experiments conducted and the obtained results, and finally Section 4 presents the conclutions and the proposed future work.

2 Semi Supervised Approach to NTL Detection

In NTL detection problems, usually the amount of labeled data is limited, and getting new labels is difficult and involves high economic costs. On the other hand, there exist more readily available and easily obtainable unlabled data. As was said in the introduction, semi supervised approach has succeded others approaches in these conditions: large amount of unlabeled data (consumption registers) and few labeled data (previously inspected). In the transductive learning, the labeled data is used to extends these labels into the unlabeled data. Thus, theoretically using unlabeled data, enables to get a better performance on the results, since as shown in figure 1, information about the unlabeled data can significantly change the classification.

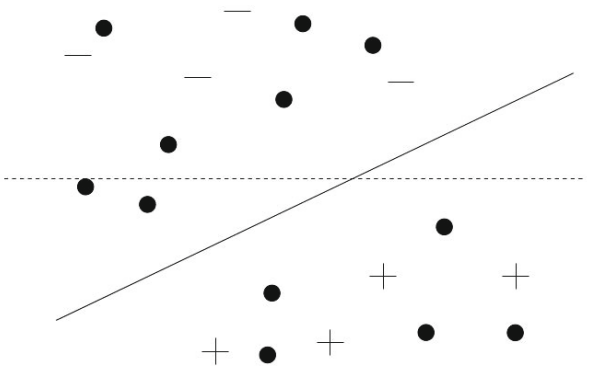

Fig. 1. Hyperplanes solution of SVM using positive/negative examples (marked as +/-, and SVM using also unlabeled data (marked as dots) [11]

2.1 Transductive SVM

Transductive SVM [11], is a variant of the SVM method, which takes into account a particular test set, and tries to minimize classification errors over that set. Thus, the required amount of labeled data is greatly reduced, allowing to implement a semi supervised method.

This method uses a labeled training set of size n, and an unlabeled test set of size k:

$$(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), ..., (\overrightarrow{x_n}, y_n) \tag{1}$$

$$\overrightarrow{x_1}^*, \overrightarrow{x_2}^*, ..., \overrightarrow{x_k}^* \tag{2}$$

The equivalent SVM primal problem (linearly separable case) is defined as:

$$\min \frac{1}{2} \| \overrightarrow{w} \|^2 \tag{3}$$

Subject to:

$$\forall i \in [1..n] : y_i [\overrightarrow{w} \cdot \overrightarrow{x_i} + b] \geqslant 1 \tag{4}$$

$$\forall j \in [1..k] : y_j^* [\overrightarrow{w} \cdot \overrightarrow{x_j^*} + b] \geqslant 1 \tag{5}$$

By varying w, b and the estimated labels y_k^*. Where w and b are the parameters that define the hyperplane of SVM. Solving this means finding a labelling of the test data and a hyperplane, so that hyperplane separates both training and test data with maximum margin, as shown in figure 1. To deal with non separable data, similar variables to those used in the conventional SVM are introduced:

$$\min \frac{1}{2} \| \overrightarrow{w} \|^2 + C \sum_{i=0}^{n} \xi_i + C^* \sum_{j=0}^{k} \xi_j^* \tag{6}$$

Subject to:

$$\forall i \in [1..n] : y_i[\overrightarrow{w} \cdot \overrightarrow{x_i} + b] \geqslant 1 - \xi_i \tag{7}$$

$$\forall j \in [1..k] : y_j^*[\overrightarrow{w} \cdot \overrightarrow{x_j^*} + b] \geqslant 1 - \xi_j^* \tag{8}$$

$$\forall i \in [1..n] : \xi_i > 0 \tag{9}$$

$$\forall j \in [1..k] : \xi_j^* > 0 \tag{10}$$

The problem is very similar to conventional SVM, the main difference is the handled of different variables ξ_i and C for labeled and unlabeled data.

To solve this problem, a first step is to label the test set according to a conventional SVM, then the iteration begins, where the test set labels are exchanged, so that the objective function decreases, and in each step the influence of the test set is increased. The iteration continues until there is no label exchange of the test set that reduces the objective function.

2.2 Performance Measure

As proposed in [14] we attempt to maximize $F_{meassure}$, and also monitor the values of Precision and Recall. We make the analysis for the default value beta equal to one, which translates in a commitment to equality between the Precision and Recall. We define $\Omega = \{\omega_+, \omega_-\}$ as the set of possible classes, being TP (true positive) the number of $x \in \omega_+$ correctly classified, TN (true negative) the number of $x \in \omega_-$ correctly classified, FP (false positive) and FN (false negative) the number of $x \in \omega_-$ and $x \in \omega_+$ misclassified respectively.

$$Recall : R = \frac{TP}{TP + FN} \tag{11}$$

$$Precision : P = \frac{TP}{TP + FP} \tag{12}$$

$$F_{meassure} : F = \frac{(1 + \beta^2)RP}{\beta^2 P + R} \tag{13}$$

2.3 Feature Selection

The initial database has 36 features per sample, which correspond to the consumption of the last 36 months. In [2] an attempt to find a small set of relevant features implementing a feature selection stage is done using a wrapper method to evaluate the performance of CSVM for the wanted features. The feature subset includes:

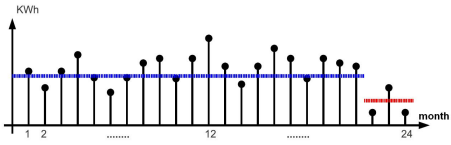

Fig. 2. Feature 1

– Consumption ratio for the last three months and the average consumption

$$car1 = \frac{mean(C[n-3:n])}{mean(C[1:n-4])} \tag{14}$$

– Consumption ratio for the last six months and the average consumption

$$car2 = \frac{mean(C[n-6:n])}{mean(C[1:n-7])} \tag{15}$$

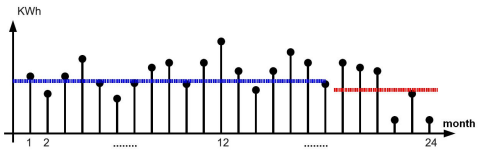

Fig. 3. Feature 2

– Difference between fifth Wavelet coefficient from the last and previous years
– Euclidean distance of each customer to the mean customer, where the mean customer is calculated by taking the mean for each month between all the customers

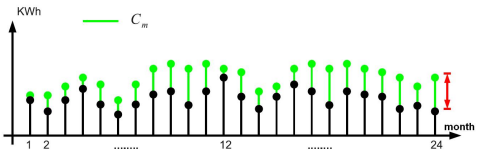

Fig. 4. Feature 4

– Slope of the straight line that fits the consumption curve

$$car28 = polyfit(C, 1) \tag{16}$$

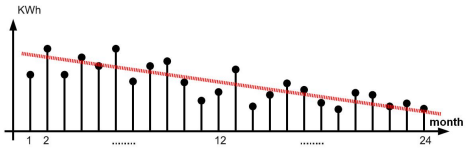

Fig. 5. Feature 5

3 Experiments and Results

3.1 Database

In this work we used a data set of 446 profiles obtained from the UTE's database. Each profile is represented by the customers monthly consumption in the last 36 months, and has two labels, one defined by technicians previous the inspection (normal or suspect) and another based on the inspection results (fraud or no fraud). While in CSVM training we only use the suspect set, in TSVM training, first we delete the fraud and no fraud labels from customers labeled as normal. Then we extend the fraud and no fraud labels from customers labeled as suspect to the customers labeled as normal (previously deleted). These final fraud and no fraud labels were utilized in the training stage. Performance evaluation was done given only the inspection labels (fraud and no fraud original labels).

The consumptions can be separated into 353 normals and 93 suspects, and 123 fraud, 323 no fraud, distributed as follows in table 1.

Table 1. Labels

Normal		Suspect	
Fraud	No fraud	Fraud	No Fraud
76	277	47	46

3.2 Algorithm Performance

SVMlight[2] [15] algorithm is used to try to improve efficiency, if it receives some unlabeled data, it automatically implement TSVM, these algorithms can be found in [16]. We iterate with the parameters C from equation [6] and gamma of the RBF Kernel used in SVMlight.

A cross validation is implemented (using 5 folds). The database, is partitioned randomly in 80% for training and 20% for test, keeping the proportions of the labels (suspect and fraud). This approach would maintain the same original ratio between fraud and no fraud labels, and allows to obtain training and test datasets for use with the algorithm.

[2] SVM library used to run TSVM.

An exhaustive search varying C from equation 6 and γ of the RBF Kernel used in SVMlight is performed to obtain optimal values, and this values are used to clasify the test set.

Results obtained after using the algorithms mentioned above can be seen in table 2

Table 2. Fraud detection results

•	P	R	Fm
Manual	51	38	44
CSVM	33.66	82.93	47.89
TSVM	34.72	81.30	48.66

We compare the performance of this semisupervised approach with supervised algorithms CSVM and with manual classification. It can be seen an improvement achieved with respect to the Fmeassure. Besides the former has the advantage over the manual classification that not require so many hours of manual preprocessing.

4 Conclusion and Future Work

In this work we analyze the performance of a strategy based on a semisupervised method, that starting from a set of labeled data (suspects) as fraud or no fraud, extends this labels to unlabeled data (no suspects), and then allows to detect new frauds at consumptions. Results show that the proposed framework, improves performance in terms of the $F_{measure}$ with inspection labels against manual methods performed by experts and previous supervised methods, avoiding hours of previous data inspect. As future work, we propose to analyze the performance of the proposed method, utilizing more amount of data, that reflects the typical imbalance between fraud and no fraud customers. Also, we propose to analize the performance of other semisupervised methods, as well as the extension of supervised methods (that have had good performance with unbalanced problems) to the semisupervised approach, as for example LFC [17] and OFC [14].

Acknowledgments. Authors would like to thank UTE, especially Juan Pablo Kosut and Fernando Santomauro, for providing datasets and share fraud detection expertise. Also thanks to Matías di Martino for fruitful discussions. This work was partially supported by the Comisión Sectorial de Investigación Científica (CSIC) - Universidad de la República and UTE.

References

1. Di Martino, M., Decia, F., Molinelli, J., Fernández, A.: A Novel Framework for Nontechnical Losses Detection in Electricity Companies. In: Latorre Carmona, P., Sánchez, J.S., Fred, A.L.N. (eds.) Pattern Recognition - Applications and Methods. AISC, vol. 204, pp. 109–120. Springer, Heidelberg (2013)
2. Rodríguez, F., Lecumberry, F., Fernández, A.: Non Technical Loses Detection - Experts Labels vs. Inspection Labels in the Learning Stage. In: International Conference on Pattern Recognition Applications and Methods, ICPRAM 2014, Angers, France, pp. 624–628 (2014)
3. Leon, C., Biscarri, F., Monedero, I., Guerrero, J.I., Biscarri, J., Millan, R.: Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies. IEEE Transactions on Power Systems (2011)
4. dos Angelos, E.W.S., Saavedra, O.R., Cortés, O.A.C., de Souza, A.N.: Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. IEEE Transactions on Power Delivery (2011)
5. Markoc, Z., Hlupic, N., Basch, D.: Detection of suspicious patterns of energy consumption using neural network trained by generated samples. In: Proceedings of the ITI 2011 33rd International Conference on Information Technology Interfaces (2011)
6. Sforna, M.: Data mining in power company customer database. Electrical Power System Research, London, U.K. (2000)
7. Depuru, S.S.S.R., Lingfeng, W., Devabhaktuni, V.: Support vector machine based data classification for detection of electricity theft. In: 2011 IEEE PES Power Systems Conference and Exposition (2011)
8. Yap, K.S., Hussien, Z.F., Mohamad, A.M.: Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm. In: 3rd IASTED Int. Conf. Advances in Computer Science and Technology, Phuket, Thailand (2007)
9. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
10. Yap, K.S., Tiong, S.K., Nagi, J., Koh, J.S.P., Nagi, F.: Comparison of Supervised Learning Techniques for Non-Technical Loss Detection in Power Utility. International Review on Computers and Software, I.RE.CO.S. (2012)
11. Joachims, T.: Transductive Inference for Text Classification using Support Vector Machines. In: International Conference on Machine Learning, ICML (1999)
12. Sindhwani, V., Keerthi, S.S.: Large Scale Semi-supervised Linear SVMs (2005)
13. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In: Proceedings of the Twentieth International Conference on Machine Learning, ICML (2003)
14. Di Martino, M., Hernández, G., Fiori, M., Fernández, A.: A new framework for optimal classifier design. Pattern Recognition 46(8), 2249–2255 (2013), doi:10.1016/j.patcog.2013.01.006, ISSN: 00313203
15. Joachims, T.: Making Large-Scale SVM Learning Practical. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
16. SVMlight - Support Vector Machine, `http://www.svmlight.joachims.org`
17. Di Martino, M., Fernández, A., Iturralde, P., Lecumberry, F.: Novel classifier scheme for imbalanced problems. Pattern Recognition Lett. 2013, 1146–1151 (2013), `http://dx.doi.org/10.1016/j.patrec.2013.03`