Optimal and Linear F-Measure Classifiers Applied to Non-technical Losses Detection

Fernanda Rodriguez, Matías Di Martino, Juan Pablo Kosut, Fernando Santomauro, Federico Lecumberry, and Alicia Fernández(✉)

Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Julio Herrera Y Reissig 565, Montevideo, Uruguay
alicia@fing.edu.uy

Abstract. Non-technical loss detection represents a very high cost to power supply companies. Finding classifiers that can deal with this problem is not easy as they have to face a high imbalance scenario with noisy data. In this paper we propose to use Optimal F-measure Classifier (OFC) and Linear F-measure Classifier (LFC), two novel algorithms that are designed to work in problems with unbalanced classes. We compare both algorithm performances with other previously used methods to solve automatic fraud detection problem.

Keywords: Class imbalance · One class SVM · F-measure · Recall · Precision · Fraud detection · Level set method

1 Introduction

Improving non-technical loss detection is a huge challenge for electrical companies. In Uruguay the national electric power utility (henceforth call UTE) addresses the problem by manually monitoring a group of customers. A group of experts inspects the monthly consumption curve of each customer and indicates those with some kind of suspicious behavior. This set of customers, initially classified as suspects are then analyzed taking into account other factors (such as fraud history, electrical energy meter type, etc.). Finally a subset of customers is selected to be inspected by an UTE's employee, who confirms (or not) the irregularity (illustrated in Figure 1). The procedure described before has major drawbacks, mainly, the number of customers that can be manually controlled is small compared with the total number of customers (around 500.000 only in Montevideo).

Several pattern recognition approaches have addressed the detection of non-technical losses, both supervised, unsupervised or recently semi-supervised as shown in [36]. Leon et al. review the main research works found in the area between 1990 and 2008 [23]. Here we present a brief review that builds on this work and wide it with new contributions published between 2008 and 2014. Several of these approaches consider unsupervised classification using different techniques such as fuzzy clustering [2], neural networks [25,35], among others. Monedero et al. use regression based on the correlation between time and

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 83–91, 2015.
DOI: 10.1007/978-3-319-25751-8_11

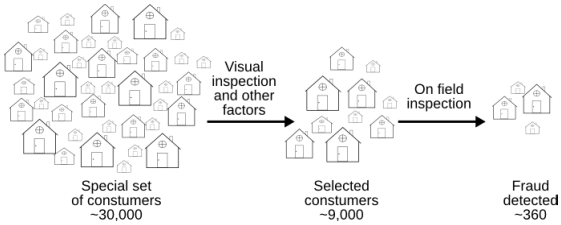

Fig. 1. Scheme of the manual procedure to detect fraudulent customers.

monthly consumption, looking for significant drops in consumption [26]. Then they go through a second stage where suspicious customers are discarded if their consumptions vary according to the moment or the year's season. Only major customers were inspected and 38% were detected as fraudulent. Similar results (40%) were obtained in [14] using a decision tree classifier and customers who had been inspected in the past year. In [9] and [37] SVM is used. In the latter, Modified Genetic Algorithm is employed to find the best parameters of SVM. In [38], is compared the methods Back-Propagation Neural Network (BPNN), Online-sequential Extreme Learning Machine (OS-ELM) and SVM. Biscarri et al. [5] seek for outliers, Leon et al. [23] use Generalized Rule Induction and Di Martino et al. [10] combine CS-SVM , One class SVM, C4.5, and OPF classifiers using various features derived from the consumption. In [34] it is compared the feature sets selected when using different classifiers with two different labelling strategies. Different kinds of features are used among this works, for examples, consumption [5,37], contracted power and consumed ratio [15], Wavelet transformation of the monthly consumption [20], amount of inspections made to each client in one period and average power of the area where the customer resides [2], among others.

This application has to deal with the class imbalance problem, where it is costly to misclassify samples from the minority class and there is a high overlapping between classes.

In almost all the approaches that deal with an imbalanced problem, the idea is to adapt the classifiers that have good accuracy in balanced domains. Many solutions have been proposed to deal with this problem [16,18]: changing class distributions [7,8,21], incorporating costs in decision making [3,4], and using alternative performance metrics instead of accuracy [17] in the learning process with standard algorithms. In [24] a comparative analysis of the two former methodologies is done, showing that both have similar performance and that they could be improved by hybrid procedures that combine the best of both methodologies. In [12] and [11] a different approach to this problem is proposed: designing a classifier based on an optimal decision rule that maximizes the F-measure [33] instead of the accuracy. In contrast with common approaches, this algorithm does not need to change original distributions or arbitrarily assign misclassification costs in the algorithm to find an appropriate decision rule.

In this work we propose to study and compare the following classifiers: Optimal F-measure (OFC) and Linear F-measure (LFC) with some classical approaches as presented in [10] applied to non-technical losses detection. In Section 2 theory is presented. In Section 3 we present non-technical loss problem and the data set used. In Section 4 experimental results are shown and in the last section we share conclusions and future work.

2 Classifiers for Unbalanced Problems

In this section we are going to introduce a brief descriptions of OFC and LFC classifiers. These classifiers were designed to face imbalance problems by looking for maximizing the F-measure value. Since high value of F-measure (F_β) ensures that both Recall and Precision are reasonably high, which is a desirable property since it indicates reasonable values of both true positive and false positive rates.This is relevant to non-technical loss detection problem since it has great imbalance between *normal* and *fraud/suspicious* classes and where, ideally, we want to detect all frauds with a minimum number of inspections to normal clients.

The goal of the OFC is to find class frontiers that guarantee maximum F-measure. The algorithm assumes that there are two classes, one called the **negative** class (ω_-), that represents the majority class, usually associated to the normal scenario (no suspicious, nor fraud), and the other called the **positive** class (ω_+) that represents the minority class (suspicious or fraud). Let us recall some related well known definitions:

$$\text{Accuracy: } \mathcal{A} = \tfrac{TP+TN}{TP+TN+FP+FN} \qquad \text{Recall: } \mathcal{R} = \frac{TP}{TP+FN}$$

$$\text{Precision: } \mathcal{P} = \tfrac{TP}{TP+FP} \qquad \text{F-measure: } F_\beta = \frac{(1+\beta^2)\mathcal{R}\,\mathcal{P}}{\beta^2\mathcal{P}+\mathcal{R}}$$

Where, TP (true positive) is 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.

As we stated before, Precision and Recall are two important measures to evaluate the performance of a given classifier in an imbalance scenario. The Recall indicates the True Positive Rate, while the Precision indicates the Positive Predictive Value. The F-measure combines them with a parameter $\beta \in [0, +\infty)$. With $\beta = 1$, F_β is the harmonic mean between Recall and Precision, meanwhile with $\beta \gg 1$ or $\beta \ll 1$, the F_β approaches the Recall or the Precision respectively.

It can be seen that maximizing F-measure is equivalent to minimizing the quantity:

$$\epsilon = \frac{\beta^2 FN + FP}{TP}. \tag{1}$$

The quantities FN, FP, and TP can be expressed as:

$$FN = P \int_{\Omega_-} f_+(x)dx,\ FP = N \int_{\Omega_+} f_-(x)dx \text{ and } TP = P \int_{\Omega_+} f_+(x)dx$$

where P and N are the number of positive and negative classes in the training database, and $f_+(x)$ and $f_-(x)$ are the probability distribution functions of the positive and negative class respectively.

Therefore, the task of training a classifier $u : \Omega \rightarrow \mathbb{R}$, that maximizes the F-measure (and minimizes ϵ) can be approached as finding the regions $\Omega_+(u) = \{x : u(x) \geq 0\}$ and $\Omega_-(u) = \{x : u(x) < 0\}$ that minimize:

$$\epsilon(u) = \frac{\beta^2 \int_{\Omega_-(u)} f_+(x)dx + \int_{\Omega_+(u)} f_-(x)dx}{\int_{\Omega_+(u)} f_+(x)dx}. \tag{2}$$

OFC looks for the classifier u that minimize $\epsilon(u)$ solving the optimization problem using a gradient descent flow, inspired by the level-set method [30]. A complete description of this classifier can be found in [12]. On the other hand, LFC proposes a way to get the regions Ω_+ and Ω_- that minimize energy ϵ thresholds for each dimension in an iterative way. A rectangular partition of the space is found by considering independently probability distributions in each dimension. Following this procedure in all the dimensions, one at a time, a set of hyper-rectangles are defined. The main difference between OFC and LFC, is that in the case of OFC, decision boundaries can have any arbitrary shape while in the case of LFC they are always parallel to input feature space coordinates axes. Although OFC is a more general approach and fewer hypothesis are assumed, LFC has the advantage of been very fast and its implementation is very simple and straightforward. For this reasons, in this work both strategies are considered and compared for the case of automatic fraud detection. A complete description of LFC algorithm can be found in [11].

Finally, as was done in previous analysis, One-Class Support Vector Machine (O-SVM) [19], Cost-Sensitive Support Vector Machine (CS-SVM) [19], Optimum Path Forest (OPF) [32], and a decision tree proposed by Roos Quinlan, C4.5 [31] are also considered and compared with OFC and LFC approaches.

It should be noted briefly that Optimum Path Forest (OPF) was proposed by [32] to be applied to the problem of fraud detection in electricity consumption, showing good results. It consists in creating a graph with the training dataset, associating a cost to each path between two elements, based on the similarity between the elements of the path. This method assumes that the cost between elements of the same class is lower than those belonging to different classes. Next, a representative is chosen for each class, called prototypes. A new element is classified as the class that has lower cost with the corresponding prototype. Since OPF is very sensitive to class imbalance, we change class distribution of the training dataset by under-sampling the majority class.

The decision tree proposed by Ross Quinlan: C4.5 it is widely utilized since it is a very simple method that obtains good results. However, it is very unstable

Table 1. Fraud detection.

Description	*Recall* (%)	*Precision* (%)	$F_{measure}$ (%)[$\beta = 1$]
OPF	36	34	35
Tree (C4.5)	33	**37**	35
O-SVM	71	31	44
CS-SVM	74	33	46
LFC	75	32	45
OFC	**77**	34	**47**

and highly dependent on the training set. Thus, a later stage of AdaBoost was implemented, accomplishing more robust results. Just as with the OPF it was needed a resamplig stage to manage the dependency of the C4.5 with the class distribution.

Related to cost-sensitive learning (CS-SVM) and one-class classifier (O-SVM), in the former different costs were assigned to the misclassification of the elements of each class, in order to tackle the unbalanced problem while the second one considers the minority class as the outliers.

3 Experimental Results

In this work we used a data set of 456 industrial profiles obtained from the UTE's database. Each profile is represented by the customers monthly consumption in the last 36 months, with inspection results labels: fraud or not fraud. A pre-processing and normalization step is performed in order to normalizes the data and to avoid peaks from billing errors. A feature set was proposed taking into account UTE's technician expertize in fraud detection by manual inspection and recent papers on non-technical loss detection [1], [28], [29]. Some of them are:

- Consumption ratio for the 3, 6 and 12 months and the average consumption.
- Difference in wavelet coefficients 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 consumption for all the customers.
- Module of the Fourier coefficients of the total consumption.
- Difference in Fourier coefficients from the last and previous years.
- Variance of the consumption curve.
- Slope of the straight line that fits the consumption curve.

It is well known that finding a small set of relevant features can improve the final classification performance [6]; this is the reason why we implemented a feature selection stage. We used two types of evaluation methods: filter and wrapper. Filters methods look for subsets of features with low correlation between them and high correlation with the labels, while wrapper methods evaluate the

performance of a given classifier for the given subset of features. In the wrapper methods, we used as performance measure the F-measure. The evaluations were performed using 10 fold cross validation over the training set. As searching method, we used *Bestfirst* [27], since we obtained a good balance between performance and computational costs.

In order to confront the class imbalance problem in O-SVM, CS-SVM, C4.5 and OPF, the strategies of changing class distribution by re-sampling [22] were used.

Table 1 shows the results obtained by the different classifiers using 10-fold cross validation.

In spite of the fact that in this work we used a more complicated and challenging dataset than that analyzed in [13], results are consistent with the reported in [13] if we compare the performance between the classifiers. CS-SVM outperforms O-SVM, C4.5 and OPF, while the novel approaches included in the present work show one of the highest results with very promising performances. OFC approach outperforms LCF as expected but, LFC also seems to be a reasonably option to face automatic fraud detection problem for instance performing similar to the best state of the art algorithm, with computational efficiency. A deeper interpretation of the results, taking into consideration the specific problem of non-technical losses detection, shows that all algorithms obtained a similar value in the rate of fraud detected (TP) per number of inspections (TP + FP). However it can be seen that the OPF and C4.5 get that performance in an operating point which corresponds to a high threshold, where it is detected a low fraudulent registrations percentage, while C-SVM, OFC and LFC are working in an operating point where a high percentage is detected. Working in a more demanding operation point (detecting not only the obvious fraud but those which are more difficult, those similar to normal records) without deteriorating the precision, reaffirms the assessment that the new proposed algorithms have a very good performance and that the use the F-measure as the objective measure to be optimized is suitable for the problem of non-technical losses.

4 Conclusions and Future Work

We propose to use two novel algorithms specially design to deal with class imbalance problems to non-technical loss detection. We compare these algorithms performance with previous strategies used to solve this problem. Performance evaluation shows that OFC and LFC can achieve similar performance to the state of art such as SVM and outperforms C4.5 and OPF classifiers. In future work, we propose to extend OFC and LFC algorithms to semisupervised approach and study the impact of applying them to the non-technical losses detection.

Acknowledgment. This work was supported by the program *Sector Productivo CSIC UTE*.

References

1. Alcetegaray, D., Kosut, J.: 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 Ingeniera-UdelaR (2008)
2. dos Angelos, E., Saavedra, O., Corts, O., De Souza, A.: Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Transactions on Power Delivery **26**(4), 2436–2442 (2011)
3. Barandela, R., Garcia, V.: Strategies for learning in class imbalance problems. Pattern Recognition 849–851 (2003)
4. Batista, G., Pratti, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations **6**, 20–29 (2004)
5. Biscarri, F., Monedero, I., Leon, C., Guerrero, J.I., Biscarri, J., Millan, R.: A data mining method based on the variability of the customer consumption - A special application on electric utility companies, vol. AIDSS, pp. 370–374. Inst. for Syst. and Technol. of Inf. Control and Commun. (2008)
6. Bishop, C.: Neural Networks for Pattern Recognition. Oxford University Press, USA (1996)
7. Chawla, Nitesh V., Lazarevic, Aleksandar, Hall, Lawrence O., Bowyer, Kevin W.: SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, Nada, Gamberger, Dragan, Todorovski, Ljupčo, Blockeel, Hendrik (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
8. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
9. Depuru, S.S.S.R., Wang, L., Devabhaktuni, V.: Support vector machine based data classification for detection of electricity theft. In: 2011 IEEE/PES Power Systems Conference and Exposition pp. 1–8 (2011)
10. Di Martino, J., Decia, F., Molinelli, J., Fernández, A.: Improving electric fraud detection using class imbalance strategies. In: International Conference In Pattern Recognition Aplications and Methods pp. 135–141 (2012)
11. Di Martino, M., Fernández, A., Iturralde, P., Lecumberry, F.: Novel classifier scheme for imbalanced problems. Pattern Recognition Letters **34**(10), 1146–1151 (2013)
12. Di Martino, M., Hernández, G., Fiori, M., Fernández, A.: A new framework for optimal classifier design. Pattern Recognition **46**(8), 2249–2255 (2013)
13. Di Martino, Matías, Decia, Federico, Molinelli, Juan, Fernández, Alicia: A novel framework for nontechnical losses detection in electricity companies. In: Latorre Carmona, Pedro, Sánchez, JSalvador, Fred, Ana L.N. (eds.) Pattern Recognition - Applications and Methods. AISC, vol. 204, pp. 109–120. Springer, Heidelberg (2013)
14. Filho, J., Gontijo, E., Delaiba, A., Mazina, E., Cabral, J.E., Pinto, J.O.: Fraud identification in electricity company costumers using decision tree pp. 3730–3734 (2004)
15. Galván, J., Elices, E., Noz, A.M., Czernichow, T., Sanz-Bobi, M.: System for detection of abnormalities and fraud in customer consumption. In: Proc. 12th IEEE/PES conf. Electric Power Supply Industry (1998)
16. Garca, V., Mollineda, J.S.S.R.A., Sotoca, R.A.J.M.: The class imbalance problem in pattern classification and learning, pp. 283–291 (2007)
17. García, V., Sánchez, J., Mollineda, R.: On the suitability of numerical performance measures for class imbalance problems. In: International Conference In Pattern Recognition Aplications and Methods, pp. 310–313 (2012)

18. Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G.: On the class imbalance problem. In: International Conference on Natural Computation, pp. 192–201 (2008)
19. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification. National Taiwan University, Taipei (2010)
20. Jiang, R.J.R., Tagaris, H., Lachsz, A., Jeffrey, M.: Wavelet based feature extraction and multiple classifiers for electricity fraud detection. IEEE/PES Transmission and Distribution Conference and Exhibition vol. 3 (2002)
21. Kolez, A., Chowdhury, A., Alspector, J.: Data duplication: an imbalance problem?. Workshop on Learning with Imbalanced Data Sets, ICML (2003)
22. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
23. Leon, C., Biscarri, F.X.E.L., Monedero, I.X.F.I., Guerrero, J.I., Biscarri, J.X.F.S., Millan, R.X.E.O.: Variability and trend-based generalized rule induction model to ntl detection in power companies. IEEE Transactions on Power Systems **26**(4), 1798–1807 (2011)
24. López, V., Fernández, A., del Jesus, M.J., Herrera, F.: Cost sensitive and preprocessing for classification with imbalanced data-sets: Similar behaviour and potential hybridizations. In: International Conference In Pattern Recognition Aplications and Methods, pp. 98–107 (2012)
25. 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, pp. 551–556 (2011)
26. Monedero, Iñigo, Biscarri, Félix, León, Carlos, Guerrero, Juan I., Biscarri, Jesús, Millán, Rocío: Using regression analysis to identify patterns of non-technical losses on power utilities. In: Setchi, Rossitza, Jordanov, Ivan, Howlett, Robert J., Jain, Lakhmi C. (eds.) KES 2010, Part I. LNCS, vol. 6276, pp. 410–419. Springer, Heidelberg (2010)
27. Müller-Merbach, H.: Heuristics: Intelligent search strategies for computer problem solving. European Journal of Operational Research **21**(2), 278–279 (1985)
28. Muniz, C., Vellasco, M., Tanscheit, R., Figueiredo, K.: A neuro-fuzzy system for fraud detection in electricity distribution. In: IFSA-EUSFLAT, pp. 1096–1101 (2009)
29. Nagi, J., Mohamad, M.: Nontechnical loss detection for metered customers in power utility using support vector machines. In: IEEE Transactions on Power Delivery, pp. 1162–1171 (2010)
30. Osher, S., Sethian, J.A.: Fronts propagating with curvature- dependent speed: Algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics **79**, 12–49 (1988)
31. Quinlan, J.: C4.5: Programs for Machine Learning. C4.5 - programs for machine learning. J. Ross Quinlan. Morgan Kaufmann Publishers (1993). http://books.google.com.uy/books?id=HExncpjbYroC
32. Ramos, C.C.O., Sousa, A.N.D., Papa, J.P., Falcao, A.X.: A new approach for nontechnical losses detection based on optimum-path forest. IEEE Transactions on Power Systems **26**(1), 181–189 (2011)
33. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
34. Rodríguez, F., Lecumberry, F., Fernández, A.: Non technical loses detection - experts labels vs. inspection labels in the learning stage. In: ICPRAM 2014 - Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, ESEO, Angers, Loire Valley, France, 6–8 March, 2014, pp. 624–628 (2014). http://dx.doi.org/10.5220/0004823506240628

35. Sforna, M.: Data mining in power company customer database. Electrical Power System Research **55**, 201–209 (2000)
36. Tacón, Juan, Melgarejo, Damián, Rodríguez, Fernanda, Lecumberry, Federico, Fernández, Alicia: Semisupervised approach to non technical losses detection. In: Bayro-Corrochano, Eduardo, Hancock, Edwin (eds.) CIARP 2014. LNCS, vol. 8827, pp. 698–705. Springer, Heidelberg (2014)
37. Yap, K.S., Hussien, Z., Mohamad, A.: 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, vol. 4 (2007)
38. 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.) **7**(2), 1828–6003 (2012)