

Large-Scale Detection of Non-Technical Losses in Imbalanced Data Sets

Patrick Glauner*, Andre Boechat*, Lautaro Dolberg*, Radu State*, Franck Bettinger†, Yves Rangoni†
and Diogo Duarte†

*Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
2721 Luxembourg, Luxembourg

Email: {first.last}@uni.lu

†CHOICE Technologies Holding Sàrl

2-4, rue Eugène Ruppert

2453 Luxembourg, Luxembourg

Email: {first.last}@choiceholding.com

Abstract—Non-technical losses (NTL) such as electricity theft cause significant harm to our economies, as in some countries they may range up to 40% of the total electricity distributed. Detecting NTLs requires costly on-site inspections. Accurate prediction of NTLs for customers using machine learning is therefore crucial. To date, related research largely ignore that the two classes of regular and non-regular customers are highly imbalanced, that NTL proportions may change and mostly consider small data sets, often not allowing to deploy the results in production. In this paper, we present a comprehensive approach to assess three NTL detection models for different NTL proportions in large real world data sets of 100Ks of customers: Boolean rules, fuzzy logic and Support Vector Machine. This work has resulted in appreciable results that are about to be deployed in a leading industry solution. We believe that the considerations and observations made in this contribution are necessary for future smart meter research in order to report their effectiveness on imbalanced and large real world data sets.

Index Terms—Electricity Theft Detection, Fuzzy Logic, Imbalanced Classification, Non-Technical Losses, Support Vector Machine

I. INTRODUCTION

Electrical power grids are the backbone of today’s society. Losses during generation and distribution cause major problems, including financial losses to electricity providers and a decrease of stability and reliability. They can be classified into technical losses and non-technical losses. Technical losses are naturally occurring and mainly include losses to power dissipation in electrical components, such as in generators, transformers and transmission lines due to internal electrical resistance. They are possible to detect and control given a knowledge of the quantities of loads.

Non-technical losses (NTL) faced by electricity providers include, but are not limited to, electricity theft by rewiring or manipulating meters. Other types include faulty meters and errors in meter readings and billing. There are different estimates of the financial losses caused by NTLs and they can range up to 40% of the total electricity distributed in countries such as Brazil, India, Malaysia or Lebanon [6], [14]. They are also of relevance in developed countries, for example estimates of NTLs in the US range from USD 1-6 billion [6].

In order to detect NTLs, inspections of customers are carried out, based on predictions whether there may be a NTL at a customer. The inspection results are then used in the learning of algorithms in order to improve predictions. However, carrying out inspections is expensive, as it requires physical presence of technicians. It is therefore important to make accurate predictions in order to reduce the number of false positives.

Detecting NTLs is challenging because of the wide range of possible causes of NTLs, such as different fraudulent types of customers. From a machine learning perspective, a key problem is the imbalance of the data, meaning that there are significantly more regular customers than customers with NTLs. We believe that this property has not adequately been addressed and reported in the literature. We therefore assess various prediction models for different proportions of NTLs in the data and discuss representative performance measures for a reliable assessment of them. We believe that an accurate discussion of this topic is necessary for future work on NTL detection in a smart meter environment.

The rest of this paper is organized as follows. Section II provides a literature review of NTL detection and its challenges. Section III describes different proposed NTL detection models and the respective data set. Section IV presents experimental results and comparison of the models on the data for different NTL proportions in the data. Section V summarizes this work and provides an outreach on future work.

II. RELATED WORK

A. Literature review

NTL detection can be treated as a special case of fraud detection, for which a general survey is provided in [10]. It highlights two approaches as key methods to detect fraudulent behavior in credit card fraud, computer intrusion and telecommunications fraud: (i) expert systems that represent domain knowledge in order to make decisions typically using hand-crafted rules and (ii) data mining or machine learning techniques that employ statistics to learn patterns from sample data in order to make decisions for future unseen data. Both

approaches have their justification and neither is generally better or worse than the other one in artificial intelligence [8].

One method to detect NTLs is to calculate the energy balance [17], which requires topological information of the network. This does not work accurately for those reasons: (i) in developing countries, network topology undergoes continuous changes in order to satisfy the rapidly growing demand of electricity, (ii) infrastructure may break and lead to wrong energy balance calculations and (iii) it requires transformers, feeders and connected meters to be read at the same time.

Another approach is to analyze the customer load profile using artificial intelligence methods, such as machine learning or expert systems. Support Vector Machines (SVM) are used in [15], working on daily average consumption features of the last 24 months for less than 400 highly imbalanced training examples, ignoring the class imbalance in the results reported. That work is combined with fuzzy logic [14] or genetic algorithms [13], focusing on an optimization of the SVM output. A rule-based expert system outperforms a SVM in [6] for an unknown amount of customers, focusing on high performance implementations. Fuzzy logic following C-means fuzzy clustering is applied to a data set of ~20K customers in [2]. Furthermore, neural networks using hand-crafted features calculated from the consumption time series plus customer-specific pre-computed attributes are used in [18] for ~1K balanced customers. Applying smart half-hour meter readings of three weeks of ~6K customers are fed into a neural network in [7]. Optimum-path forest are applied to NTL detection in [20] for ~10K customers outperforming different SVMs and a neural network. A different method is to estimate NTLs by subtracting an estimate of the technical losses from the overall losses [21]. In many electricity grids it may be challenging to accumulate the entire losses and furthermore, this method does not scale to large numbers of meters. The class imbalance problem of electricity theft detection has initially been addressed in [11]. It applies an ensemble of two SMVs, optimum-path forest and C4.5 decision tree learning to ~300 on-field inspection test data. However, the degree of imbalance of the ~1.5K training examples is not reported. Furthermore, in the optimization of the classifiers, the true negative rate is ignored, which results in too many costly inspections of non-fraudulent customers.

Many of these results are constrained to either just a few test examples or report accuracies on highly imbalanced classes. To date, working on large and long-term data sets and properly measuring the performance of the classifiers on imbalanced data sets has not adequately been studied in the literature. However, ignoring the class imbalance in reported results is also true for many other machine learning applications. In this paper, we focus on large data sets comprising each of ~100K inspection results spanning four years of consumption data and apply different NTL detection methods on it. We particularly address the class imbalance problem using accurate performance measures.

B. Challenge of supervised learning for anomaly detection

It must be noted that most NTL detection methods are supervised. Anomaly detection - a superclass of NTL - is generally challenging to learn in a supervised manner for the reasons stated in [16]: (i) anomaly data sets contain a very small number of positive examples and large number of negative examples, resulting in imbalanced classes, (ii) it is used for many different kinds of anomalies as it is hard for any algorithm to learn from just a few positive examples what the anomalies might look like and (iii) there may be also future anomalies which may look completely different to any of the anomalous examples learned so far. In contrast, supervised learning works best for (i) large numbers of both positive and negative examples, (ii) when there are enough positive examples so that the algorithm can get a sense of what positive examples might look like and (iii) future positive examples are likely to be similar to the ones in the training set.

III. NTL DETECTION

A. Data

The data used in this paper is from an electricity provider in Brazil. It consists of three parts: (i) ~700K customer data, such as location, type, etc., (ii) ~31M monthly consumption data from January 2011 to January 2015 such as consumption in kWh, date of meter reading and number of days between meter readings and (iii) ~400K inspection data such as presence of fraud or irregularity, type of NTL and inspection notes.

Most inspections do not find NTLs, making the classes highly imbalanced. In order for the models to be applied to other regions or countries, they must be assessed on different NTL proportions. Therefore, the data was subsampled using 17 different NTL proportion levels: 0%, 0.1%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%. Each sample contains ~100K inspection results.

B. Models

In this Section, the different models for NTL detection of this paper are described. The first model is a CHOICE Technologies product based on Boolean logic and is used as a baseline. It is extended to fuzzy logic in the second model in order to smoothen the decision making process. The third model is a Support Vector Machine, a state-of-the-art machine learning algorithm.

1) *Boolean logic*: This model is an expert system, it consists of hand-crafted rules created by the CHOICE Technologies expert team which are conjunctions of (in)equality terms, such as:

$$(N_1 > v_1) \wedge (N_1 < v_2) \wedge (N_2 < v_3) \wedge (N_3 = v_4) \dots \quad (1)$$

N_x is a so-called attribute. Possible attributes are change of consumption over the last 3 months, slope of consumption curves, etc. and v_x is a numeric value. In total, 42 attributes are used in 14 rules. If at least one rule outcome is true, that customer is considered to potentially cause a NTL.

2) *Fuzzy logic*: Fuzzy systems [3] have a long tradition in control applications allowing to implement expert knowledge in a softer decision making process. They allow to relate to classes of objects, breaking up boundaries, making membership a matter of degree. In this paper, the 14 Boolean rules were fuzzified and incorporated in a Mamdani fuzzy system using the centroid defuzzification method [3]. Fuzzy rules rely on membership functions. The number of membership functions for each attribute depends on the ranges of values found in the rules among which 1 attribute has 1 function, 32 attributes have 2 membership functions and 9 attributes have 4 functions. In most cases, trapezoid membership functions are used to keep the model simple. The exact parameters, such as membership function boundaries or the mean of sigmoid membership functions were determined from the distribution of attribute values.

However, these parameters could be optimized using: (i) gradient techniques [22], (ii) genetic algorithms [22] or (iii) neuro-fuzzy systems [1]. Techniques (i) and (ii) are highly constrained optimization problems due to dependence among parameter values to keep the fuzzy system valid. Technique (i) was studied further and its results are reported in Section IV.

3) *Support Vector Machine*: A Support Vector Machine (SVM) [24] is a maximum margin classifier, i.e. it creates a maximum separation between classes. Therefore, a SVM is less prone to overfitting than other classifiers, such as a neural network [4]. Support vectors hold up the separating hyperplane. In practice, they are just a small fraction of the training examples.

The training of a SVM can be defined as a Lagrangian dual problem having a convex cost function. In that form, the optimization formulation is written in terms of only the dot product $x^{(i)} \cdot x^{(j)}$ between points in the input space. By default, the separating hyperplane is linear. For complex problems, it is advantageous to map the data set to a higher dimension space, where it is possible to separate them using a linear hyperplane. A kernel is an efficient function that implicitly computes the dot product in the higher dimensional space. A popular kernel is the Gaussian radial basis function: $K(u, v) = \exp(-\gamma \|u - v\|^2)$.

Inspired by [15], for M customers $\{0, 1, \dots, M - 1\}$ over the last N months $\{0, 1, \dots, N - 1\}$, a feature matrix F is computed, in which element $F_{m,d}$ is a daily average kWh consumption feature during that month:

$$x_d^{(m)} = \frac{L_d^{(m)}}{R_d^{(m)} - R_{d-1}^{(m)}} \quad (2)$$

where for customer m , $L_d^{(m)}$ is the kWh consumption increase between the meter reading to date $R_d^{(m)}$ and the previous one $R_{d-1}^{(m)}$. $R_d^{(m)} - R_{d-1}^{(m)}$ is the number of days between both meter readings of customer m . Similarly, a binary target vector T is created in which element $T^{(m)}$ is the most recent inspection result for customer m in the respective period of time. NTLs are encoded by 1 if they are detected and 0 if not.

A. Metrics

In many classification problems, the classification rate, or accuracy is used as a performance measure. Given the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN): $ACC = \frac{TP+TN}{TP+TN+FP+FN}$. However, many publications ignore that it is only of minor expressiveness for imbalanced classes. For a NTL detection example, given a data set of 990 negative and 10 positive test examples, a classifier that always predicts negative has an accuracy of 0.99. This example clearly demonstrates that other performance measures must be used for NTL detection. The recall is a measure of the proportion of the true positives found. It is also named true positive rate (TPR) or sensitivity: $Recall = \frac{TP}{TP+FN}$. The specificity is a measure of the proportion of the true negatives classified as negative. It is also named true negative rate (TNR): $Specificity = \frac{TN}{TN+FP}$. The false positive rate (FPR) is $1 - TNR$. A receiver operating characteristic (ROC) curve plots the TPR against the FNR. The area under the curve (AUC) is a performance measure between 0 and 1, where any binary classifier with an AUC > 0.5 performs better than random guessing. While in many applications multiple thresholds are used to generate points plotted in a ROC curve, the AUC can also be computed for a single point, when connecting it with straight lines to (0, 0) and (1, 1) as shown in [9]: $AUC = \frac{Recall+Specificity}{2}$.

For NTL detection, the goal is to reduce the FPR to decrease the number of costly inspections, while increasing the TPR to find as many NTL occurrences as possible. In order to assess a NTL prediction model using a single performance measure, the AUC is the most suitable.

B. Methodology

Throughout the experiments, consumption readings and inspection result data are used. Further data, such as location of customers are not used. In the comparison of the three classifiers, the AUC performance measure is used for the different levels of NTL proportion mentioned in Section III-A. We assessed different values for the number of the most recent meter readings N . Only customers with complete time series of the last N months before the respective inspection are considered. The larger N , the less data is available. At least 12 months should be considered in order to represent seasonality effects. Experiments for the last 12, 18 and 24 months were carried out, for which 12 months have proven to lead to the best results as the other experiments lead to more overfitting. Due to lack of space, those results are omitted.

The SVM is the only classifier that requires training in our experiments. However, since it is a binary classifier, it could not be trained on NTL proportions of 0% and 100%. For the NTL proportions used for training, 10-folded cross validation is performed for every NTL proportion, splitting the data into a 60%/20%/20% training/validation/test ratio. The AUC score is used as the validation measure to pick the best classifier fold. Throughout the experiments, a linear SVM is used. The same

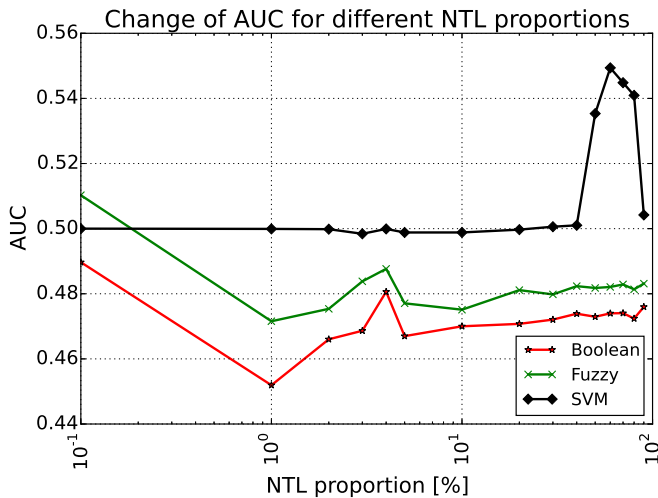


Fig. 1. Comparison of classifiers tested on different NTL proportions.

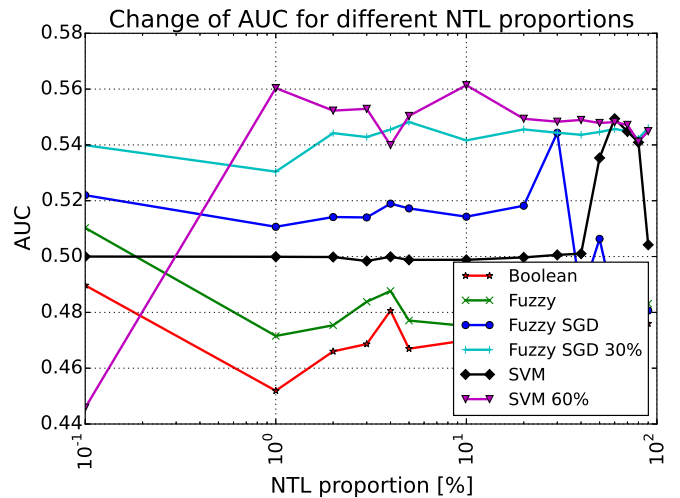


Fig. 2. Comparison of optimized classifiers tested on different NTL proportions.

experiments were repeated using a Gaussian Kernel, which proved to overfit for all NTL proportions.

C. Implementation details

The Boolean and fuzzy classifiers were implemented in MATLAB, the latter using the Fuzzy Logic Toolbox [12]. The SVM classifier was implemented in Python using `scikit-learn` [19], which builds on top of `LIBSVM` [5]. The regularization parameter and the inverse variance parameter γ of the Gaussian kernel were not optimized explicitly, as `scikit-learn` optimizes them automatically. Using 10-fold cross-validation to train 10 SVMs and to select the best one takes about 2 minutes per NTL proportion on a state-of-the-art i5 notebook. Using the Boolean or fuzzy systems to classify the same amount of data takes about 1 second. However, both classifiers use pre-computed customer-specific attributes. Computing them takes a couple of hours in a cloud infrastructure.

D. Comparison of classifier performance

For different NTL proportions, the change of test AUC for the Boolean and fuzzy systems and the SVM can be observed in Fig. 1. The Boolean classifier has an AUC < 0.5 for all NTL proportions and therefore performs worse than random guessing. The same applies for the fuzzy system, except for a NTL proportion of 0.1%. The SVM performs only (noticeably) better than random guessing for NTL proportions between 50% and 80%.

Given the theory of fuzzy systems and their potential, the parameters of the fuzzy system were optimized using stochastic gradient descent (SGD) for each of the 15 binary NTL proportions: 0.1% to 90%. Out of the 15 optimized fuzzy systems, the one with the greatest AUC test score is picked and tested on all NTL proportions. The fuzzy system trained on 30% and tested on all NTL proportions - Fuzzy SGD 30% - significantly outperforms both, the Boolean and fuzzy systems, as shown in Fig. 2.

TABLE I
NORMALIZED CONF. MATRICES FOR TEST ON 5% NTL PROPORTION.

Classifier	Actual	Predicted	
		TNR	FPR
Boolean	Actual	0.53	0.47
		0.60	0.40
Fuzzy SGD 30%	Actual	0.87	0.13
		0.77	0.23
SVM 60%	Actual	0.36	0.64
		0.26	0.74

The same methodology as for the optimized fuzzy system is applied to the SVM. SVMs are trained on all 15 binary NTL proportions and SVM 60%, the SVM trained on 60% NTL proportion, is selected because of the greatest AUC test score. Its performance compared to the Boolean and Fuzzy SGD 30% are shown in Fig. 2. In summary, SVM 60% performs in a similar range as Fuzzy SGD 30% compared to the Boolean system, except for very small NTL proportions $< 1\%$.

However, comparing the confusion matrices of both classifiers, they perform very differently as shown in Tables I and II for selected NTL levels of 5% and 20%, respectively. The optimized fuzzy system has a higher TNR, but lower TPR compared to the optimized SVM. In return, the SVM has a higher TPR, but a lower FNR.

E. Discussion

The initial Boolean and fuzzy models perform worse than random guessing and are therefore not suitable for real data, as they trigger too many inspections while not many of them will lead to NTL detection. Optimized fuzzy and SVM models trained on 30% and 60% NTL proportion, respectively, result in significantly greater AUC scores. However, both perform very differently, as the optimized fuzzy system is more con-

TABLE II
NORMALIZED CONF. MATRICES FOR TEST ON 20% NTL PROPORTION.

Classifier	Actual	Predicted	
		TNR	FPR
Boolean	Actual	0.53	0.47
		0.58	0.42
Fuzzy SGD 30%	Actual	0.87	0.13
		0.78	0.22
SVM 60%	Actual	0.35	0.65
		0.25	0.75

servative in NTL production. In contrast, the optimized SVM is more optimistic, leading also to a higher FPR. In general, neither can be named better than the other one, as picking the appropriate model from these two is subject to business decisions.

However, this work also demonstrates that for real data, NTL classifiers using only the consumption profile are limited. Therefore, it is desirable to use more features like location, inspection notes, etc. Another issue with the real data is the potential bias of inspections so that this sample of customers does not represent the overall population of customers. We expect a correction of the bias to lead to better predictions, too.

V. CONCLUSION AND FUTURE WORK

In this work, we have proposed three models for NTL detection for large data sets of 100K customers: Boolean, fuzzy and Support Vector Machine. In contrast to other results reported in the literature, the optimized fuzzy and SVM models were assessed for varying NTL proportions on imbalanced real world consumption data. Both have an AUC > 0.5 for all NTL proportions $> 0.1\%$ and significantly outperform simple Boolean or unoptimized fuzzy models. The improved models are about to be deployed in a CHOICE Technologies product. The contribution methodology is necessary for future smart meter research, in order to report their effectiveness in imbalanced and large real world data sets.

We are planning to evaluate unsupervised methods, in particular deep learning, in order to detect NTL more accurately by finding hidden correlations in the data. Furthermore, we are planning to use other features in our models, such as the location and latent features and to investigate cost-based optimization in order to maximize the total electricity recovered through inspections. Also, we are planning to make our implementations faster and more scalable using Apache Spark [25].

REFERENCES

- [1] A. Abraham, "Adaptation of Fuzzy Inference System Using Neural Learning", *Fuzzy Systems Engineering, Studies in Fuzziness and Soft Computing*, vol. 181, pp. 53-83, July 2005.
- [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] Plamen Angelov, "Autonomous Learning Systems: From Data Streams to Knowledge in Real-time", Wiley, ISBN: 978-1-119-95152-0, December 2012.
- [4] L. J. Cao and F. E. H. Tay, "Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting", *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1506-1518, November 2003.
- [5] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, vol. 2, issue 3, pp. 27:1-27:27, 2011.
- [6] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni and R. C. Green, "High Performance Computing for Detection of Electricity Theft", *International Journal of Electrical Power & Energy Systems*, vol. 47, issue 1, pp. 21-30, May 2013.
- [7] V. Ford, A. Siraj and W. Eberle, "Smart grid energy fraud detection using artificial neural networks", *IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, pp. 1-6, 9-12, December 2014.
- [8] D. Gorgevik, D. Cakmakov and V. Radevski, "Handwritten digit recognition using statistical and rule-based decision fusion", *11th Mediterranean Electrotechnical Conference (MELECON)*, pp.131-135, 2002.
- [9] W. B. van den Hout, "The area under an ROC curve with limited information", *Medical Decision Making*, vol. 23, issue 1, pp. 160-166, March-April 2003.
- [10] Y. Kou, C.-T. Lu, S. Sirwongwattana and Y.-P. Huang, "Survey of fraud detection techniques", *IEEE International Conference on Networking, Sensing and Control*, vol. 2, pp. 749-754, 2004.
- [11] M. Di Martino, F. Decia, J. Molinelli and Alicia Fernandez, "Improving electric fraud detection using class imbalance strategies", 2012.
- [12] MathWorks, "Matlab Fuzzy Logic Toolbox User's Guide", 2015.
- [13] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed and A. M. Mohammad, "Detection of abnormalities and electricity theft using genetic Support Vector Machines," *IEEE Region 10 Conference on TENCEN 2008*, pp. 1-6, November 2008.
- [14] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed and F. Nagi, "Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System", *IEEE Transactions on Power Delivery*, vol. 26, issue 2, pp. 1284-1285, April 2011.
- [15] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines", *IEEE Transactions on Power Delivery*, vol. 25, issue 2, pp. 1162-1171, Apr. 2010.
- [16] A. Ng, "Machine Learning", Coursera, 2014.
- [17] C. C. B. de Oliveira, N. Kagan, A. Meffe, S. L. Caparroz and J. L. Cavaretti, "A New Method for the Computation of Technical Losses in Electrical Power Distribution Systems", *Proceedings CIRED*, 2001.
- [18] C. Muniz, K. Figueiredo, M. M. B. R. Vellasco, G. Chavez and M. A. C. Pacheco, "Irregularity detection on low tension electric installations by neural network ensembles", *IEEE - INNS - ENNS International Joint Conference on Neural Networks*, June 2009.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, V., B. Thirion, et al., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [20] C. C. O. Ramos, A. N. Souza, J. P. Papa and A. X. Falcao, "Fast Non-Technical Losses Identification Through Optimum-Path Forest", *15th International Conference on Intelligent System Applications to Power Systems (ISAP)*, November 2009.
- [21] S. Sahoo, D. Nikovski, T. Muso and K. Tsuru, "Electricity theft detection using smart meter data," *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1-5, February 2015.
- [22] Lala Septem Riza, Christoph Bergmeir, Francisco Herrera and Jose Manuel Benitez, "frbs: Fuzzy Rule-Based Systems for Classification and Regression in R", *Journal of Statistical Software*, vol. 65, issue 6, May 2015.
- [23] Y. Tang, Y.-Q. Zhang, N. V. Chawla and S. Krasser, "SVMs Modeling for Highly Imbalanced Classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, issue 1, pp. 281-288, Feb. 2009.
- [24] Vladimir N. Vapnik, "An overview of statistical learning theory", *IEEE Transactions on Neural Networks*, vol. 10, issue 5, pp. 988-999, Sep. 1999.
- [25] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: cluster computing with working sets", *HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 2010.