Conference on
Information Technology and
Data Science

A Review on Latest Trends in Non-Technical Loss Detection*

Khawaja MoyeezUllah Ghori[a][b], Muhammad Awais[c], Akmal Saeed Khattak[d], Muhammad Imran[e], Rabeeh Ayaz Abbasi[f], Laszlo Szathmary[g]

[a]Department of Computer Science, National University of Modern Languages, NUML, Islamabad, Pakistan
mghouri@numl.edu.pk

[b]University of Debrecen, Doctoral School of Informatics, Debrecen, Hungary

[c]Department of Computer Science, Edge University, Ormskirk, United Kingdom
mawais@ieee.org

[d]Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
akhattak@qau.edu.pk

[e]College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia
dr.m.imran@ieee.org

[f]Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
rabbasi@qau.edu.pk

[g]University of Debrecen, Faculty of Informatics, Debrecen, Hungary
szathmary.laszlo@inf.unideb.hu

Proceedings of the 1st Conference on Information Technology and Data Science
Debrecen, Hungary, November 6–8, 2020
published at http://ceur-ws.org

Abstract

An increasing interest in digging out the consumption patterns in power and energy sector is observed globally. This includes electrical, gas, and water supply industries. A reason behind analyzing the consumption patterns

*The work/publication is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund.

is the detection of fraudulent attempts which are made for the illegal reduction of bill payments. In the case of electricity, these attempts are made by reversing the meters, by-passing or slowing down the meters or inaccurate readings. The detection of theft attempts in power industry is termed as Non-Technical Loss (NTL) detection. With the increasing demand for electricity, the occurrences of NTL have been reported globally including India, Pakistan, Brazil and China etc. In this paper, we first describe the use of the synthesized and the real datasets in NTL detection. Then, we highlight an interesting characteristic of class imbalance that is exhibited in the datasets used for NTL detection. Moreover, we identify the fruitful areas in NTL detection where the research community has been working on. Lastly, we discuss the need for a relative comparison of the classical machine learning and deep learning over a benchmark dataset for NTL detection.

Keywords: Non-Technical Loss (NTL), Non-Technical Loss detection, machine learning, classification, class imbalance

1. Introduction

Recently, an increasing interest has been observed in recognizing the consumption patterns of the consumers of electricity, gas and water supplies [25]. One of the main objectives of this activity is to identify and forecast the potential theft attempts in order to have reduced bills. This illegal theft attempt has dented the economies of many countries causing a loss of billions of dollars. This includes China [14], Pakistan [10], India, Brazil [18], etc.

Non-Technical Loss (NTL) detection in electric power industry is a term used for the detection of faulty meters or illegal usage of electric units. Losses are bore by the electric supply companies on the account of faulty meters that record fewer units as compared to the consumed electricity. On the other hand, this practice can also be intentional in order to get the electricity bill reduced by a substantial margin. For both cases, the supplier companies look for a solution which can identify them the faulty meters or potential theft instances.

One of the important characteristics of the relevant datasets is that they belong to the class imbalance problem. It is the problem where the dataset is biased towards one class by its heavy representation while the other class is less representative. Interestingly, the problem becomes more challenging when the focus is on the true representation of the less representative class. Naturally, the number of normal electric consumption in a neighborhood area is huge as compared to the number of theft attempts. This gives a clear indication that the real dataset of the consumption of electricity belongs to the class imbalance problem where the number of negative class samples is huge as compared to the number of positive class samples. The techniques used in NTL detection should be able to balance out the positive and the negative class samples before the dataset is used by the machine learning algorithms [9].

One of the techniques used to identify the NTL is applying classical machine learning algorithms to the datasets pertaining to the consumption of electricity.

This includes the use of Support Vector Machine (SVM), KNN, decision trees, ensemble methods and neural networks [8]. Advances in deep learning has attracted some researchers to test different variants of deep learning for NTL detection. For e.g., the authors of [3] have used deep neural networks along with long short-term memory network to identify the occurrences of NTL in a dataset pertaining to the smart meters of a utility company in Spain. However, there is still a need to compare the performances of the classical machine learning algorithms with the different variants of deep learning architecture. In this paper, we focus on elaborating the importance of a comparative study of the two paradigms for NTL detection in a real dataset.

2. Synthesized Vs. Real Datasets

Two types of datasets are used in NTL detection. One of the types belongs to the synthesized datasets which are randomly generated keeping in mind the requirements of NTL detection. One of the benefits of using the synthesized datasets is that they are easily accessible while on the other hand, they might miss the potentially useful information which otherwise would have been handy in detecting unlawful consumption activities. The other type is the real dataset which is taken from a distribution company. An essential advantage of using such datasets is analyzing the real and unique patterns of the consumption of electricity which might be missing in the synthesized datasets. An associated drawback of using real datasets is that it is hard to get an access of such datasets as distribution companies generally avoid sharing the consumer's information.

3. Class Imbalance: An aspect of NTL Detection

An interesting property associated with the datasets of electric distribution companies is that they belong to the class imbalance problem. Considering two classes; normal consumption and abnormal consumption; the class imbalance problem arises when most of the consumption records belong to the normal consumption class while few records belong to the abnormal consumption class. When the objective is predicting the abnormal consumption class which is rarely represented in the dataset, the success ratio of the classifiers might get deteriorated unless the class imbalance ratio is properly dealt.

To balance out the number of representative records for both the classes, undersampling the majority class can be used in the pre-processing stage of data preparation. This is termed as synthetic minority over-sampling technique (SMOTE) [30]. The other way is over-sampling the minority class, i.e., duplicating or randomly creating new synthesized records of the minority samples [4]. Both techniques have been used in different problem domains belonging to the class imbalance. However, few have tried these techniques in NTL detection.

4. Which Areas in NTL Detection are fruitful?

There is a general pattern followed in the pre-processing of the datasets and application of the classifiers in order to achieve a significant success in NTL detection. This pattern is shown in the Figure 1. The electricity distribution companies provide raw data containing consumption records. Depending on the type of metering infrastructure, the consumption data is either half hourly, hourly, daily, bi-monthly or monthly records. The automated metering infrastructure (AMI) facilitates to record hourly or daily consumption while the manual metering infrastructure uses monthly manual readings done by the meter readers. Not all the records in consumption profile are useful in performing analytics for NTL detection. Similarly, not all the features are important for detecting NTL. For this, record selection and feature selection is performed. Once the dataset is ready, it is scaled for normalizing the values. The next step is training and testing of the classifiers. A range of classifiers of different types is applied. The performance of the classifiers is then measured by performance evaluation metrics.

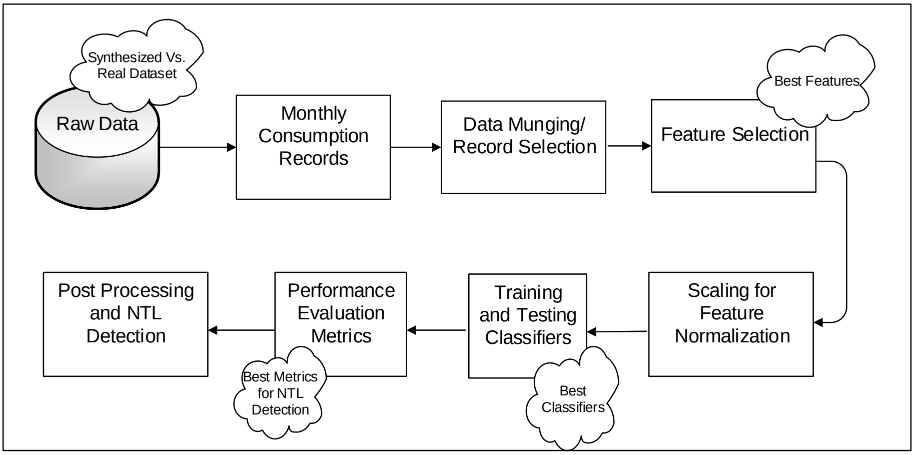
Synthesized Vs. Real Dataset
Raw Data
Monthly Consumption Records
Data Munging/ Record Selection
Feature Selection
Best Features
Scaling for Feature Normalization
Training and Testing Classifiers
Best Classifiers
Performance Evaluation Metrics
Best Metrics for NTL Detection
Post Processing and NTL Detection

Figure 1. Pre-processing Pattern of NTL Detection.

The researchers in this area are largely interested in three main modules. Some of them try to find the best combination of features which are most suited for NTL detection. For this, they use feature importance. In one of our previous contributions, we propose the Incremental Feature Selection (IFS) algorithm that selects the best combination of features responsible in detecting the occurrence of NTL [9]. Some authors have taken a keen interest in finding the best individual classifiers or a combination of classifiers for NTL detection. Others have tried to find the best type of the classifiers for NTL detection.

Some authors have focused on the performance evaluation module of the classifiers. They have tried to find out which metrics are best suited to evaluate the classifiers considering the specifications of the NTL detection problem. For example, we have concluded in one of our previous works that recall should be given a higher priority as compared to any other performance evaluation metric considering the special characteristics of the NTL detection problem [10].

5. Machine Learning Vs. Deep Learning in NTL Detection

Machine learning is mainly used for Non-Technical Loss (NTL) detection. NTL is a typical classification task. Most of the researchers have used unsupervised, supervised, semi-supervised, hybrid and network based methods. Some of the research on NTL detection can be seen in [2, 11–13, 15, 17, 19, 23, 24, 26–29, 31–36]. Table 1 summarizes research work on NTL detection. Most researchers used electric supply and distribution companies of specific countries as it is shown in Table 1. The most common evaluation measures used in NTL detection are accuracy, precision, recall, AUC and others as Table 1 shows. A framework to detect NTL is proposed in [17]. The characteristics of consumers were analyzed and data mining techniques were used to recover electrical energy loss. Consumption data, extracted from a Spanish power supply firm known as Endesa Distribution, was used by [17]. Association rule mining was used to find groups of consumers responsible for non-technical loss in the form of electricity theft. Hartmann et. al. [13] used contextual learning on a dataset from Luxembourg electric distribution company. Accuracy, precision, recall and F1 measures were used as evaluation measures. The attributes of consumption were investigated in [26] for NTL detection in two firms. For this purpose, hierarchical clustering, Benford curve and Multi-Dimensional Scaling (MDS) were used on two datasets based on two Colombian companies. The study was evaluated using ROC. Local Outlier Factor (LOF) based on the DBSCAN clustering algorithm was used in the research work of [28]. The technique was evaluated using the Silhouette coefficient and Davies Bouldin index. The research work of [34] used a synthetic dataset and observed unusual profiles of electricity consumers by proposing a distance matrix. Area Under ROC Curve (AUC), F-1 measure and accuracy was used as evaluation measure. The results showed the effectiveness of the proposed technique after comparing it with DBSCAN, GMM and kNN. Yeckle and Tang [33] carried experiments on Irish dataset. Different outlier detection methods are used to detect NTL. The performance is measured by AUC. The research work of Zheng et. al. [36] explored deep convolutional neural networks (CNN) on a dataset based Chinese electricity company. The result from deep convolutional neural networks is compared with SVM, random forest, logistic regression and TSR (Three Sigma Rule). It was observed that deep CNN outperformed the above mentioned classifiers. Another research work of [32] used 4000 Irish household data records collected from smart metering. Gustafson-Kessel clustering algorithm was used in [32] to discriminate non-technical loss. The rate of true positive was observed as 63.6% whereas false positive was 24.3%.

A dataset collected from a Spanish electrical supply firm was analyzed in the

Table 1. Literature review on NTL detection.

Year	Author-Ref.	Data	Techniques	Evaluation
2015	Hartmann et. al. - [13]	Luxembourg electric distribution company	contextual learning, live machine learning, Gaussian mixture model, profile power consumption	Accuracy, precision, recall, f1 measure
2016	Peng et. al. - [24]	Chinese electric company	Fuzzy clustering and classification	Mean Index Adequacy and execution time (MIA)
2017	Sánchez et. al - [26]	Two Colombian Companies	Hierarchical Clustering, Multidimensional Scaling, Decision Trees	ROC
2017	Sharma et. al - [28]	Data collection from USA and India	Local Outlier Factor (LOF), (DBSCAN) clustering algorithm	Silhouette coefficient, Davies Bouldin index
2017	Zheng et. al - [34]	Synthetic dataset	density-based electricity theft detection	Area Under ROC Curve (AUC), accuracy and F1 measures
2018	J. Yeckle and B. Tang - [33]	Irish dataset	Outlier detection techniques	AUC
2018	Zheng et. al. - [36]	Chinese electricity dataset	Deep CNN	AUC
2018	Viegas et. al. - [32]	Irish household data collected from smart metering	Gustafson-Kessel fuzzy clustering algorithm	True positive and false positive rate
2018	Guerrero et. al. - [11]	Spanish electricity company data	ANN, classification, regression and SOP	Accuracy
2020	Pazi et. al - [23]	Electrical consumption data for a large municipality in South Africa	SVM, kNN and naïve Bayes	Accuracy, detection rate, precision, true negative rate

research of [11]. Artificial neural network, regression and Self Organizing Map (SOM) are used and as a result it was claimed that the accuracy has improved. Statistical learning methods were used in [23] to detect NTL in South Africa. Three classifiers namely Support Vector Machine (SVM), kNN (k Nearest Neighbour) and naïve Bayes were used. The classification models were evaluated using accuracy, detection rate, precision and true negative rate.

Less attention has been paid to deep neural networks and network science based algorithms to detect NTL. From the literature review presented, it can be observed that classical machine learning is the major streamline in solving the problem of NTL detection. However, there is still a need of a thorough testing of different variants of deep learning in a real dataset for NTL detection. Community detection techniques have a key role in computer science, sociology, biology, physics, economics, engineering, marketing, ecology, political sciences and many other fields [7]. Community detection is used for inferring useful information from complex networks. Complex networks are getting attention from researchers of different domains. A network such as biological, information, technological, social and other can be modelled as graphs [6]. Some well known community detection techniques such as Greedy Modularity Maximization [5, 16, 20], Girvan-Newman algorithm [21], Louvain algorithm [1] and k-clique percolation [22] can also be used to identify interesting patterns in NTL.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre: *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment 2008.10 (Oct. 2008), P10008,
doi: 10.1088/1742-5468/2008/10/p10008.

[2] M. Buevich, X. Zhang, O. Shih, D. Schnitzer, T. Escalada, A. Jacquiau-Chamski, J. Thacker, A. Rowe: *Microgrid Losses: When the whole is Greater than the Sum of its parts*, in: Proceedings of the 7th International Conference on Cyber-Physical Systems, IEEE Press, 2016, pp. 46–50.

[3] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, A. Gomez-Exposito: *Hybrid deep neural networks for detection of non-technical losses in electricity smart meters*, IEEE Transactions on Power Systems 35.2 (2019), pp. 1254–1263.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer: *SMOTE: synthetic minority over-sampling technique*, Journal of artificial intelligence research 16 (2002), pp. 321–357.

[5] A. Clauset, M. E. J. Newman, C. Moore: *Finding community structure in very large networks*, Physical Review E 70.6 (Dec. 2004), issn: 1550-2376,
doi: 10.1103/physreve.70.066111,
url: http://dx.doi.org/10.1103/PhysRevE.70.066111.

[6] S. Fortunato: *Community detection in graphs*, Physics Reports 486.3 (2010), pp. 75–174, issn: 0370-1573,
doi: https://doi.org/10.1016/j.physrep.2009.11.002.

[7] S. Fortunato, D. Hric: *Community detection in networks: A user guide*, Physics Reports 659 (2016), Community detection in networks: A user guide, pp. 1–44, issn: 0370-1573,
doi: https://doi.org/10.1016/j.physrep.2016.09.002.

[8] K. M. Ghori, R. A. Abbasi, M. Awais, M. Imran, A. Ullah, L. Szathmary: *Performance analysis of different types of machine learning classifiers for non-technical loss detection*, IEEE Access 8 (2019), pp. 16033–16048.

[9] K. M. Ghori, A. R. Ayaz, M. Awais, M. Imran, A. Ullah, L. Szathmary: *Impact of Feature Selection on Non-technical Loss Detection*, in: 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), IEEE, 2020, pp. 19–24.

[10] K. M. Ghori, M. Imran, A. Nawaz, R. A. Abbasi, A. Ullah, L. Szathmary: *Performance analysis of machine learning classifiers for non-technical loss detection*, Journal of Ambient Intelligence and Humanized Computing (2020), pp. 1–16.

[11] J. I. Guerrero, I. Monedero, F. Biscarri, J. Biscarri, R. Millán, C. León: *Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility*, IEEE Transactions on Power Systems 33.2 (2018), pp. 1209–1218.

[12] W. Han, Y. Xiao: *NFD: a practical scheme to detect non-technical loss fraud in smart grid*, in: Communications (ICC), 2014 IEEE International Conference on, IEEE, 2014, pp. 605–609.

[13] T. Hartmann, A. Moawad, F. Fouquet, Y. Reckinger, T. Mouelhi, J. Klein, Y. Le Traon: *Suspicious electric consumption detection based on multi-profiling using live machine learning*, in: Smart Grid Communications (SmartGridComm), 2015 IEEE International Conference on, IEEE, 2015, pp. 891–896.

[14] W. Hu, Y. Yang, J. Wang, X. Huang, Z. Cheng: *Understanding Electricity-Theft Behavior via Multi-Source Data*, arXiv preprint arXiv:2001.07311 (2020).

[15] L. A. P. Júnior, C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. da Costa, J. P. Papa: *Unsupervised non-technical losses identification through optimum-path forest*, Electric Power Systems Research 140 (2016), pp. 413–423.

[16] H. Kwak, Y.-H. Eom, Y. Choi, H. Jeong, S. Moon: *Consistent Community Identification in Complex Networks*, 2009, arXiv: `0910.1508 [physics.soc-ph]`.

[17] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, R. Millán: *Variability and trend-based generalized rule induction model to NTL detection in power companies*, IEEE Transactions on Power Systems 26.4 (2011), pp. 1798–1807.

[18] J. A. Meira, P. Glauner, R. State, P. Valtchev, L. Dolberg, F. Bettinger, D. Duarte: *Distilling provider-independent data for general detection of non-technical losses*, in: Power and Energy Conference at Illinois (PECI), 2017 IEEE, IEEE, 2017, pp. 1–5.

[19] R. Mutupe, S. Osuri, M. Lencwe, S. D. Chowdhury: *Electricity theft detection system with RF communication between distribution and customer usage*, in: PowerAfrica, 2017 IEEE PES, IEEE, 2017, pp. 566–572.

[20] M. E. J. Newman: *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences 103.23 (2006), pp. 8577–8582, issn: 0027-8424,
doi: `10.1073/pnas.0601602103`, eprint: `https://www.pnas.org/content/103/23/8577.full.pdf`,
url: `https://www.pnas.org/content/103/23/8577`.

[21] M. E. J. Newman, M. Girvan: *Finding and evaluating community structure in networks*, Phys. Rev. E 69 (2 Feb. 2004), p. 026113,
doi: `10.1103/PhysRevE.69.026113`,
url: `https://link.aps.org/doi/10.1103/PhysRevE.69.026113`.

[22] G. Palla, I. Derényi, I. Farkas, T. Vicsek: *Uncovering the overlapping community structure of complex networks in nature and society*, Nature 435 (2005), pp. 814–818, issn: 1476-4687,
doi: `10.1038/nature03607`,
url: `https://doi.org/10.1038/nature03607`.

[23] S. Pazi, C. M. Clohessy, G. D. Sharp: *A framework to select a classification algorithm in electricity fraud detection*, en, South African Journal of Science 116 (Oct. 2020), pp. 1–7, issn: 0038-2353,
url: `http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0038-23532020000600010&nrm=iso`.

[24] B. Peng, C. Wan, S. Dong, J. Lin, Y. Song, Y. Zhang, J. Xiong: *A two-stage pattern recognition method for electric customer classification in smart grid*, in: Smart Grid Communications (SmartGridComm), 2016 IEEE International Conference on, IEEE, 2016, pp. 758–763.

[25] Q. A. Al-Radaideh, M. M. Al-Zoubi: *A data mining based model for detection of fraudulent behaviour in water consumption*, in: 2018 9th International Conference on Information and Communication Systems (ICICS), IEEE, 2018, pp. 48–54.

[26] C. C. Sánchez-Zuleta, J. P. Fernández-Gutiérrez, C. C. Piedrahita-Escobar: *Identification of the characteristics incident to the detection of non-technical losses for two Colombian energy companies*, Revista Facultad de Ingeniería Universidad de Antioquia 84 (2017), pp. 60–71.

[27] D. D. Sharma, S. Singh: *Aberration detection in electricity consumption using clustering technique*, International Journal of Energy Sector Management 9.4 (2015), pp. 451–470.

[28] D. D. Sharma, S. Singh, L. Jeremy, E. Foruzan: *Identification and characterization of irregular consumptions of load data*, Journal of Modern Power Systems and Clean Energy 5.3 (2017), pp. 465–477.

[29] J. V. Spirić, S. S. Stanković, M. B. Dočić: *Identification of suspicious electricity customers*, International Journal of Electrical Power & Energy Systems 95 (2018), pp. 635–643.

[30] M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan: *A multiple expert approach to the class imbalance problem using inverse random under sampling*, in: International workshop on multiple classifier systems, Springer, 2009, pp. 82–91.

[31] E. Terciyanli, E. Eryigit, T. Emre, S. Caliskan: *Score based non-technical loss detection algorithm for electricity distribution networks*, in: Smart Grid and Cities Congress and Fair (ICSG), 2017 5th International Istanbul, IEEE, 2017, pp. 180–184.

[32] J. L. Viegas, P. R. Esteves, S. M. Vieira: *Clustering-based novelty detection for identification of non-technical losses*, International Journal of Electrical Power & Energy Systems 101 (2018), pp. 301–310.

[33] J. Yeckle, B. Tang: *Detection of Electricity Theft in Customer Consumption Using Outlier Detection Algorithms*, in: Data Intelligence and Security (ICDIS), 2018 1st International Conference on, IEEE, 2018, pp. 135–140.

[34] K. Zheng, Y. Wang, Q. Chen, Y. Li: *Electricity theft detecting based on density-clustering method*, in: 2017 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), 2017, pp. 1–6,
doi: `10.1109/ISGT-Asia.2017.8378347`.

[35] K. Zheng, Y. Wang, Q. Chen, Y. Li: *Electricity theft detecting based on density-clustering method*, in: Innovative Smart Grid Technologies-Asia (ISGT-Asia), 2017 IEEE, IEEE, 2017, pp. 1–6.

[36] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, Y. Zhou: *Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids*, IEEE Transactions on Industrial Informatics 14.4 (2018), pp. 1606–1615.