IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Approach for Detection and Prevention of Electricity Pilferage Using Machine Learning

Satish S. Banait [1], Darshana S. Sarode[2], Kapil M. Shelke[3], Himesh S. Rathod[4], Akash Galande[5]

[1]Assistant Prof, Department of Computer Engineering

[2,3,4,5] Students, Department of Computer Engineering

[1]K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract: Electricity theft is the primary cause of electricity power loss that significantly affects the revenue loss and the quality of electricity power. Nevertheless, the existing methods for the detection of this criminal behaviour of theft are diversified and dimensionality of time-series data makes it challenging to extract meaningful information. This paper addresses these problems by developing a novel electricity theft detection model, integrating three algorithms. The proposed method first applies the synthetic minority oversampling technique for balancing the dataset, secondly we apply pre-processing on smart meter data then do feature selection. Extensive experiments are performed by using real electricity consumption data, and results show that the proposed method outperforms other methods in terms of theft detection.

Index Terms — non technical loss, electricity theft, pipeline, imbalanced dataset, machine learning.

I. INTRODUCTION

Electricity is a basic need of modern life. It is used for lighting, cooling, heating, and powering electric appliances and machines. Modern means of medicines and surgery, entertainment, communication, and transportation have been revolutionized by electricity to comfort people. As demand and use of power are increasing day by day, different measures are being taken to make it capable of fulfilling the requirements. However, the power loss is still the biggest threat to electricity management system the need to reduce power losses and to optimize the useof electricity has led to the development of an intelligent energy system and smart grid (SG). SG is based on advancedmetering infrastructure (AMI) [2]. AMI has introduced smart meters (SM) to the energy system and replaced traditional electric meters [3]. AMI system is able to monitor power consumption concerning the time that helps utility companies to detect anomalies in the network [4], [5]. Anomaly is the unexpected behaviour that makes a customer suspicious [6]. The smart meters generate a huge amount of data, and these data can be helpful in solving many problems [7] such as power loss.

The energy losses in the power system may occur during transmission, distribution, and consumption [1], [8]. These losses are divided into two categories: technical losses (TLs) and commercial loses that are also called non-technical losses(NTLs) [1], [5], [9]. The reason behind TLs is the dissipation of energy in the conductors, transmission lines, and distribution lines. The reasons behind NTLs include installation errors, faulty meters, billing errors, tampering the meters, hacking the smart meters, manipulating the data, direct hooking on other households, etc. [5], [10]. According to utilities, NTLs are defined as the power consumed by the customers that has not been billed by the utility [11].NTLs cause significant harm to the energy system. Fraudulent behaviour of a customer is responsible for the economic problems. NTL is the major cause for revenue loss of power utility [8]. It also affects the quality of supply. Furthermore, NTLs are responsible for the increase in energy tariff that affects all consumers. Because such losses are divided among all consumers during tariff calculation [12]. It also decreases the stability and reliability of the power grid [11]. According to Northeast Group LLC, the world losses $89.3 billion per year due to electricity theft [13]. NTLs are a critical issue not only in developing countries but also for developed countries. As the ratio of NTLs of some countries is tabulated in tableI

TABLE I
RATIO OF NON TECHNICAL LOSS

Country	NTL ratio	Year	Reference
India	$4.5 billion	2013	[2]
US	$6 billion	2016	[5]
Brazil	U.S $4 billion	2011	[12]
Honduras	U.S $13 million	2017-2019	[16]
UK	U.S $23 million	2012	[17]
Puerto Rico	$400 million	2010	[18]

Over the past decade, detecting and reducing NTL is the primary concern of electricity providers as it recovers a higher rate of revenue losses [14]. In literature, several approaches to identify NTL are found, categorized as data-oriented techniques, network-based techniques, and hybrid techniques [15]. These techniques use the data of SG. The information about the network (topology, network measurement, and sensor network) is used in the network-oriented techniques and the information related to consumers such as power consumption and consumer type is used in data-oriented techniques. While, hybrid techniques use the data of both data-oriented as well as network oriented techniques. The main issue that limits the detection rate of any classifier is the imbalanced nature of the data [16]. The data of SG is used to detect NTL. Such as in [1], [3], [4], [12], and [15] data analysis is performed on SG data to detect theft. However, most of these techniques have some limitations including, imbalanced nature of classes, required manual inspection, hardware devices (sensors), handcrafted feature engineering, computational complex, and are not evaluated well in terms of detection rate. In this work, we aim to overcome the problems of existing techniques as mentioned above. A novel electricity theft detection method is proposed by initializing three different algorithms in the pipeline. At the first step the synthetic minority over-sampling technique (SMOTE) is used to balance the dataset, secondly kernel function and principal component analysis (KPCA) for the feature extraction from high dimensional time-series data, and finally support vector machine (SVM) for the classification.

II. RELATED WORK

Electrical power loss is defined as the difference between units generated at the generation side and distributed at distribution side. During the last few years, the research community has paid attention to the problems related to NTL detection [19]. To detect that loss machine learning and deep learning algorithms are used. In this section, approaches for NTL detection and mitigation found in the literature are presented. Two types of solutions are found in the literature, namely: hardware-based and data-driven based [1]. Hardware-based are costly as compared to data-driven approaches; that is whyattention is paid on data-driven approaches. Data-driven approaches consist of classification and game theory [20]. In [1], wide and deep convolutional neural net- work (CNN) is used to capture non-periodicity of power consumption data. However, highly imbalanced data is used. A combination of long short-term memory (LSTM) and multilayer perceptron (MLP) is used for NTL detection in [21]. LSTM is trained on weekly consumption data, while sequential data is used for MLP. Grid search is used for parameters optimization that makes it computational complex and does not guarantee optimal parameters. Authors in [8] used CNN for feature extraction and random forest (RF) is used for finalclassification. The performance of the machine learning algorithm canbe enhanced by using the most relevant features from raw data. In [12], [21], and [22] features are extracted using black hole algorithm, CNN, and maximal overlap discrete wavelet-packet transform (MODWPT) respectively. However, the black hole algorithm gets stuck in local optima and does not guarantee optimal features. MODWPT is computational complex. Moreover, grid search is used for the optimization of parameters that do not provide an optimal solution.MODWPT with random under-sampling boosting (RUS- Boost) is used in [22]. However, RUS causes loss of useful information from raw data. Moreover, the false detection rateis also high that indicates parameters are not tuned well. In [23], the random oversampling technique is used to balance the data. However, it causes overfitting. In [24], three variants of gradient boosting are proposed as theft detector with the modification of theft attacks. However, training time is high and the over-sampling technique used for balancing the dataset causes overfitting. Bagged tree are used in [25]. The results of this technique also lead to overfitting and low detection rate. A hybrid of CNN and LSTM is used in [26] for the detection of electricity loss. CNN was used for the feature extraction. Moreover, synthetic minority oversampling technique (SMOTE) is used to get satisfactory results.

I. OUR APPROACH

In this section, we describe the proposed methodology. Our framework mainly includes two main stages namely: 1) data analysis and preprocessing and 2) the proposed pipeline. Figure. 1 and 2 give overview of the approach usedin this work.

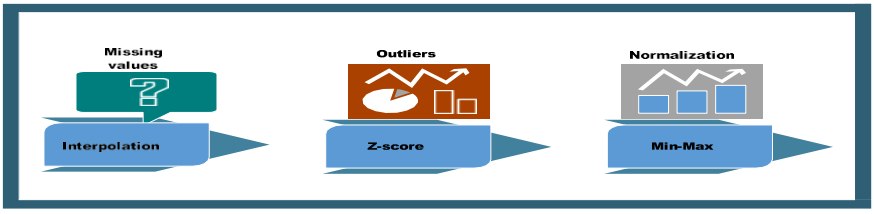

Fig. 1. Overview of data pre-processing

3.1 Data Preprocess

Table II, shows the information about the dataset used to conduct experiments. Data preprocessing method is per- formed to handle erroneous and missing values from the data. Interpolation is a process used to predict missing values. The equation used to deal with missing values is given in [1]. Moreover, outliers are the abnormal patterns that are essential to remove from the data. Z-score is the novel method used to remove outliers from high-dimensional data. After, recovering the missing values from the dataset, erroneous (outliers) values are removed using Z-score as mentioned in [27]. Finally, Min- Max is performed to normalize the data. It is also referred to as feature scaling. In this process, the value of every featureis transformed into a decimal [1].

3.2 Pipeline

In machine learning some tasks are repeated frequentlyto process data or to train a model. Pipeline is used to prevent the use of same task for different values. Several cross validated steps are assembled together while setting different

TABLE II INFORMATION OF SGCC

Total No. of customers	42372
Fraudulent customers	3615
Honest customers	38757
Time duration	2014-01-01 to 2016-10-31

parameters. Several algorithms are chained, composed and scrambled together to process stream of data. Machine learning offers several workflows that can be automated using pipeline. Pipeline is inherited from the scikit-learn. Scikit-learn are a library in python that provides a range of supervised and unsupervised learning algorithms. It is also known as sklearn [28]. Following libraries are used to build:

- **imblearn:** basically, it is used for balancing the data,
- **SciPy:** it is used for technical computing tasks,
- **NumPy:** used for high level mathematical functions, and
- **Matplotlib:** provides plotting functionalities.

Sklearn library is used for regression, classification, clustering,preprocessing, dimensionality reduction and model selection.It is widely used in data science. Steps and memory are the two parameters of the pipeline used to set order (sequence) and to cache the fitted transformer respectively. Pipeline is just an abstract notion and is not some existing machine learning algorithm. It combines transformer withclassifier or any other estimator to build a composite estimator. The library of transformers is used to reduce dimensionality, expand (kernel approximation) clean the data (preprocessing), and extract features. Pipeline offers multiple estimators into single pipeline. Pipeline serves number of benefits:

- **Convenience** It helps to make desired number and order of steps to use machine learning algorithms. Coherent and easy-to-understand workflow can be implemented using pipeline.
- **Reproducibility** Different parameters and values in entire pipeline can be reused and to reproduced.
- **Encapsulation** Whole sequence of estimators can be fit and predict once in a pipeline. Pipeline makes it easy to use different types of models.
- **Joint parameter selection** Grid search can be applied on the parameters of all estimators in the pipeline at once.
- **Safety** Pipelines ensures that transformers and predictorsare trained using similar samples. It avoids leaking statisticsfrom training sets into test sets your test in cross- validation.

The proposed pipeline consist of SMOTE as mentioned below, for balancing the dataset, principal component analysis (PCA)and support vector machine (SVM) for feature extraction andas a classifier respectively.

1) ***Synthetic Minority Oversampling Technique:*** SMOTE algorithm uses K-nearest neighbor (K-NN) approach to generate synthetic data. Minority instances are used to generate the data. To introduce the synthetic data, SMOTE takes nearest neighbors from feature vectors, computes the distance between them, multiplies the difference by number (0,1), and adds to feature space. SMOTE is used in the proposed pipeline to balance the dataset.

2) ***Principal Component Analysis:*** PCA is a statistical technique. PCA is used for data analysis. The rate of lossof information during dimensions reduction is minimum in PCA. It identifies useful patterns from the dataset and retains spatial attributes of data. High dimensional electricity consumption data is reconstructed by reducing the dimensions and extracting underlying consumption trends. Data transformed by the PCA has maximum variance. Harold Hoteling wasthe

first person who explained PCA [29]. However, PCAis not efficient for high dimensional non-linear data. KPCAis used for nonlinear generalization of data. Initially, KPCA transformed the input data into infinite dimensional through a nonlinear mapping, and then PCA is performed in the featurespace.

3) ***Support Vector Machine:*** SVM is flexible and a powerful supervised machine learning model. It is used for both classification and regression problem. SVM works by creating one or more hyper planes that separate the data clusters. Hyper plane isa decision plane which divides objects of different classes. Theuse of kernel in SVM makes it equivalent to fed forward neuralnetwork. The kernel trick converts a low dimensional space into high dimensional space that makes it flexible, powerful and accurate. Generally, linear, polynomial and radial basis function (RBF) kernel are used in SVM. In our approach, RBF kernel is used in SVM.

REFERENCES

[1] Zheng, Z., Yang, Y., Niu, X., Dai, H.N. and Zhou, Y.,"Wide and deep convolutional neural networks for electricity- theft detection to secure smart grids." 2017, IEEE Transactions on Industrial Informatics, 14(4), pp.1606-1615. Zheng,

[2] Micheli, G., Soda, E., Vespucci, M.T., Gobbi, M. and Bertani, A.,"Big data analytics: an aid to detection of non- technical losses in power utilities."2019, Computational Management Science, 16(1-2), pp.329- 343.

[3] Maamar, A. and Benahmed, K., A Hybrid Model for Anomalies Detection in AMI System Combining K-means Clustering and Deep Neural Network.",2019, vol.60, no.1, pp.15-39.

[4] Zheng, K., Chen, Q., Wang, Y., Kang, C. and Xia, Q., A novel combined data-driven approach for electricity theft detection. 2018, IEEE Transactions on Industrial Informatics, 15(3), pp.1809-1819.

[5] Razavi, R., Gharipour, A., Fleury, M. and Akpan, I.J.,A practical feature- engineering framework for electricity theft detection in smart grids.2019, Applied energy, 238, pp.481-494.

[6] Fan, C., Xiao, F., Zhao, Y. and Wang, J., Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data.2018, Applied energy, 211, pp.1123-1135.

[7] Amalina, F., Hashem, I.A.T., Azizul, Z.H., Fong, A.T., Firdaus, A., Imran, M. and Anuar, N.B., 2019. Blending big data analytics: Review on challenges and a recent study. IEEE Access, 8, pp.3629-3645.

[8] Li, S., Han, Y., Yao, X., Yingchen, S., Wang, J. and Zhao, Q., Electricity Theft Detection in Power Grids with Deep Learning and Random Forests.2019, Journal of Electrical and Computer Engineering doi:10.1155/2019/4136874.

[9] Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P. and Gómez- Expósito, A.," Hybrid deep neural networks for detection of non- technical losses in electricity smart meters."2019, IEEE Transactions on Power Systems doi:10.1109/TPWRS.2019.2943115.

[10] Lu, X., Zhou, Y., Wang, Z., Yi, Y., Feng, L. and Wang, F., Knowledge Embedded Semi-Supervised Deep Learning for Detecting Non-Technical Losses in the Smart Grid.2019, Energies, 12(18), p.3452.

[11] Guerrero, J.I., Monedero, I., Biscarri, F., Biscarri, J., Millán, R. and León, C., Non-technical losses reduction by improving the inspections accuracy in a power utility. 2017 IEEE Transactions on Power Systems, 33(2), pp.1209- 1218.

[12] Ramos, C.C., Rodrigues, D., de Souza, A.N. and Papa, J.P. On the study of commercial losses in Brazil: a binary black hole algorithm for theft characterization. 2016, IEEE Transactions on Smart Grid, 9(2), pp.676- 683.

[13] PR Newswire. World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets. 2014. Availableonline:http://www.prnewswire.com/news-releases/world-loses-893- billion-to-electricity-theft-annually- 587-billion-in-emerging-markets- 300006515.html (accessed on 10 February 2020).

[14] Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P. and Gómez- Expósito, A., Detection of non-technical losses using smart meter data and supervised learning.2018, IEEE Transactions on Smart Grid, 10(3), pp.2661-2670.

[15] Ghori, K.M., Abbasi, R.A., Awais, M., Imran, M., Ullah, A. and Szathmary, L., " Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection." 2019, IEEE Access, 8, pp.16033-16048.

[16] Jokar, P., Arianpoo, N. and Leung, V.C., Electricity theft detection in AMI using customers' consumptionpatterns.2015, IEEE Transactions on Smart Grid, 7(1), pp.216-226.

[17] Avila, N.F., Figueroa, G. and Chu, C.C., NTL detection in electricdistribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting. 2018, IEEE Transactions on Power Systems, 33(6), pp.7171-7180.

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

[19] Ghori, K.M., Abbasi, R.A., Awais, M., Imran, M., Ullah, A. & Szath- mary, L. Performance analysis of different types of machine learning classifiers for non-technical loss detection. 2019, IEEE Access, 8, pp.16033-16048.

[20] Z. Xiao, Y. Xiao and D. H. Du, March 2013.Exploring Malicious Meter Inspection in Neighbourhood Area 376 Smart Grids, in IEEE Transactions on Smart Grid, vol. 4, no. 1, pp. 214-226.

[21] N. F. Avila, G. Figueroa, and C.-C. Chu, NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random under sampling boosting, 2015, IEEE Transactions on Power Systems, vol. 33, pp. 7171-7180.

[22] Jokar, P., Arianpoo, N. and Leung, V.C., Electricity theft detection in AMI using customers' consumption patterns.2015, IEEE Transactions on Smart Grid, 7(1), pp.216-226.

[23] R. Punmiya and S. Choe, Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing,2019, IEEE Transactions on Smart Grid, vol. 10, pp. 2326-2329.

[24] M. S. Saeed, M. W. Mustafa, U. U. Sheikh, T. A. Jumani, and N. H. Mirjat, ,Ensemble Bagged Tree Based Classification for Reducing Non- Technical Losses in Multan Electric Power Company, 2019, Electronics, vol. 8, p. 860.

[25] Hasan, M., Toma, R.N., Nahid, A.A., Islam, M.M. and Kim, J.M., Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach, 2019 Energies, 12(17), p.3310.

[26] Yang, X., Zhou, W., Shu, N. and Zhang, H., A Fast and Efficient Local Outlier Detection in Data Streams , February, 2019. In Proceedings of the 2019 International Conference on Image, Video and Signal Processing (pp. 111-116).

[27] Hao, J. and Ho, T.K., Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. 2019, Journal of Educational and Behavioural Statistics, 44(3), pp.348-361.