International Journal of Engineering & Technology, 7 (3.34) (2018) 900-904

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Machine Learning Algorithm for Efficient Power Theft Detection using Smart Meter Data

Jeyaranjani J[1], Devaraj D[2]

[1] Department of Computer science and Engineering
KARE, Virudhunagar, India
Department of Electrical and Electronics Engineering KARE, Virudhunagar, India
**Corresponding Author E-mail: Jeyaranjani.j@gmail.com, deva230@yahoo.com*

Abstract

Electricity theft is one of the major problems of electric utilities. The dishonest electric power users produce financial loss to the utility companies. It is not possible to inspect the manually. The electricity consumption energy data obtained from the Smart Meter installed at customer premise have the information that is used for identifying the anomaly customers. This paper proposes an approach to identify the suspect customers using the customer power usage pattern. Machine learning algorithm is used for this purpose. The trustworthiness of customers is verified and is selected for theft program. This analysis is carried out by tweaking the actual Smart Meter data to create fraudulent data. The ANN classification model is developed using supervised learning algorithm that helps to discriminate the customers profile based on their genuine activity and fraudulent activity in electricity power usage. Simulation result shows that the proposed system is efficient in identifying the suspects with high accuracy.

Keywords: *Smart Grid, AMI, Smart Meter, Electricity Theft, Machine Learning Algorithm, k-means ANN*

1. Introduction

The electric grid refers to a network of transmission lines, substations, transformers and more that deliver electricity from the power plant to our home or business. Digital technology that allows for two-way communication between the utility and its customers, and the sensing along the transmission lines is what makes the electric grid smart. The smart grid components include Automated Metering Infrastructure (AMI), Phasor Measurement Unit and Communication network. The AMI describe the whole infrastructure from smart meter to two way communication network to control center equipment and all the applications that enable the gathering and transfer of energy usage information in near real-time. The components of AMI include: smart meter, communication network, meter data acquisition system, meter data management system. The AMI improvise the following features: system reliability, energy cost, and electricity theft. The functionality includes service switching, time-based rates, remoteprogramming to control smart devices, power quality measure, and a user interface for real-time monitoring. It is an automated device having the features to collect the consumption data usually in hour basis (may vary).

2. Related Works

Most of the previous research work focus on customer load profile information to expose abnormal behavior that is known to be highly correlated with Non Technical Loss (NTL) activities. In [1] GA provides globally optimized SVM hyper-parameters using a combination of random and prepopulated genomes.The decision tree technique is implemented for finding the potential fraud activity in [2]. The Artificial neural network is built to classify the Non Technical Loss – power tampering for intelligently identifying the losses by selecting the most required features from the customer profile in [3]. The extreme Learning Machine classification technique elucidates the operation of identifying the customer energy consumption pattern that classifies genuine and illegal profiled customers. The classification models are applied on regular energy consumption data as well as the encoded data to compare corresponding classification accuracies andcomputational overhead in [4].

The previous research works on power theft detection focuses on the customer power usage profile data for theft detection. The specific are where there is dissimilarity in supplied power and billed power. All the customers belonging to that area are considered to be suspects. The drawback of the work discussed previously is that power theft identification has been carried out based on the assumption that the customers are suspected to be fraud. This case could identify the potential customer as the fraudulent customers. This provides the motivation for this research work to include the bogus customer power consumption data into the actual power consumption data. The machine learning algorithms are used to analyze the data that literally cluster and then classify the Customer. The customer's data are discriminated as genuine and fraud based on their usage pattern.

The remaining sections are organized as follows: In section 2, methodology of the proposed paper has been discussed. Section 3 presents the information about the dataset. Section 4 presents the results and discussions. Section 5 discusses the conclusion.

3. Proposed Model

Fig. 1, presents the flow diagram of the proposed method for electricity theft detection. The customers profile dataset is given as input to k-means clustering algorithm.

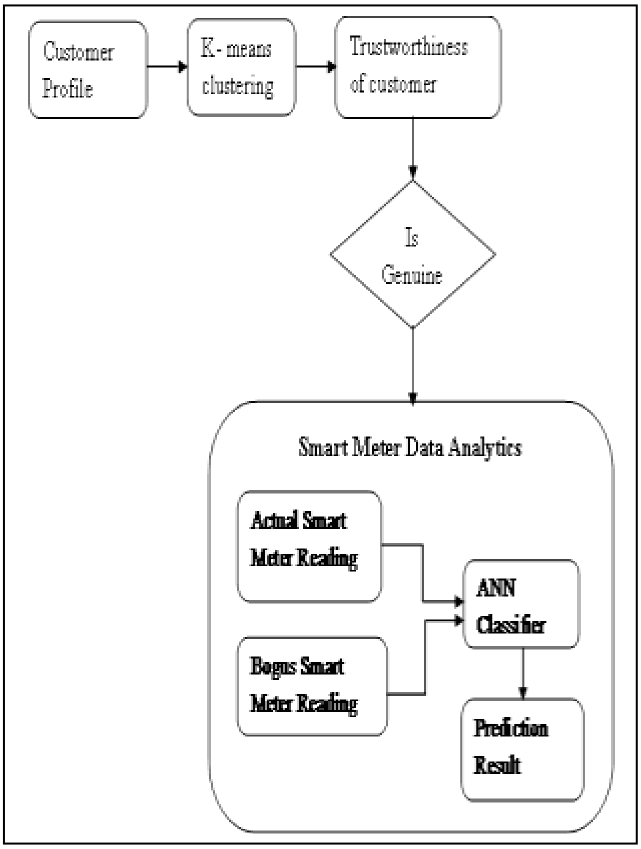
Customer Profile
K- means clustering
Trustworthiness of customer
Is Genuine
Smart Meter Data Analytics
Actual Smart Meter Reading
Bogus Smart Meter Reading
ANN Classifier
Prediction Result

Fig. 1: Proposed Electricity Theft Identification Model

The k-means clustering algorithm will cluster these customers into k clusters. The customer's profile that is close enough to the cluster head is the final selected customers for the theft detection task. These customers are identified as trustworthy customers. The selected genuine customers profile data are the input to the smart meter data analytics system. Three types of bogus data are generated using these selected customers profile dataset. Then the classifier is trained to categories the normal and bogus data that generated and included into the actual genuine dataset. The performance of the classifier is measured and evaluated.

3.1 Trustworthiness of Customers

Define From the available Smart meter data, the customer profile data of a specific area is chosen. It is believed that all the customers of the selected region are genuine profiled. The percentage of trustworthiness of the customers is identified by using machine learning algorithm. For this purpose, unsupervised k-means clustering algorithm is applied. The clustering is performed to select the customer's profile which is more genuine in their power usage pattern by grouping them in clusters. The customers are clustered based on their power usage readings obtained from their smart meter. For this purpose, a sample customer's smart meter reading for every 30 minutes in (KWH) for 28 days is considered.

3.2 K-Means Clustering

Initially the customers profile is clustered based on the feature values (smart meter reading). The k-means clustering algorithm allows us to specify the number of expected cluster. It group the customers based on their 28 days power usage value. '***K'*** cluster are formed to accommodate the customers profile in any one specific cluster. 'K' is the number of clusters determined by ourselves. The number of clusters is determined in correspondence to the number of customer profile taken into consideration (E.g. For 10 customer, 3 clusters is acceptable. The clusters are formed with number of customer's profile, where the cluster with less number of customer profile are not considered. The cluster with the maximum number of customer profile is selected. The customers belonging to the cluster that satisfies these two conditions are considered for the Smart Meter Data Analytics. The prominent customers of the selected cluster are identified by calculating the Euclidean distance of each customer profile to its corresponding cluster head value. The Euclidean distance is used for finding the distance between the cluster member and cluster head. The formula to find the Euclidean distance is as follows:
N

$$D(a,b) = D(b,a) = \sqrt{\sum}(a_i - b_i)^2 i{=}1 \qquad (1)$$

Where a, b are the customers profile, n is the total number of customers considered for analytics. After certain number of iteration, the usage pattern of finalized cluster head originated as the source for the anomaly customer detection. The customers profile which are close to the cluster head is identified as genuine profile. The genuine profiles are separated and their energy consumption data are taken for further investigation. The trustworthiness of the customer is verified by using the clustering algorithm

Table 1: Trustworthiness of Customers

Content	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Customer Profiles	15	14	12	9
Mean of Cluster	47.78571	46.84615	55	48.11111
Normal Profile Customers	10	12	11	9
Anomaly Profile Customers(Might be)	5	2	1	0

3.3 Smart Meter Data Analytics

The genuine customers profile is obtained from the result of k-means clustering. For each customer, the smart meter reading is obtained for every half an hour. This sampling rate is reduced to one reading (average reading of 48 readings of a day) per day per customer. For each of the considered sample in dataset, three types of bogus data samples are generated for every day reading. In type 1, a random value is generated between -0.5 and 0.5. This random value is multiplied with the average reading value calculated for per day. It replicates the acceptable change in meter reading. This is done for all 28 days readings of all sample customers considered for the experiment. In type 2, random days are taken where the actual data values are replaced with zero. This implies the scenario where the meter does not work. In type 3, the mean value of 28 days readings are multiplied with the each of the day reading. Fig. 2, represents an example for the three types of assault generated to the actual data samples. The three bogus data types ($T_{1,}$ $T_{2,}T_3$) are mathematically represented by the following equations:

- $T_1(D_t) = \alpha D_t$ α = Random (-0.5, 0.5)
- $T_2(D_t) = \beta D_t$ β = Random (0.1, 0.8)
- $T_3(D_t) = Y D_t$ Y= mean (D_t)

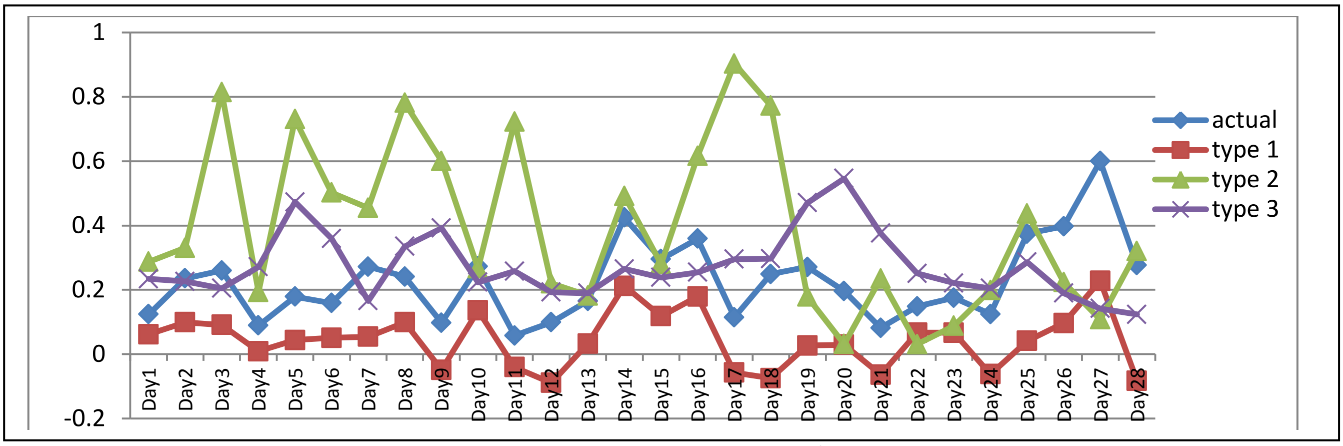

Fig. 2: Sample customer data with three types of bogus data

3.4 Classification

After finding the trustworthiness of the customers, the genuine profiled customers are considered for the classification model by including the bogus data into actual data. The Artificial Neural Network is built to classify the customer's profile. The three types of bogus data along with the actual data are considered to train the neural network. 60% of the dataset is utilized for training the neural network. After required number of iterations, the neural network is trained to predict any new customer profile to genuine or fraud. The remaining 40% of the dataset is used for testing the dataset. The prediction is made by the ANN classification model. The performance of the proposed system is using two parameters namely accuracy and error rate. The difference in actual class value and the predicted class value is considered for the performance evaluation. The performance of the model depends on the number of dataset taken into consideration.

4. Results and Discussion

The Smart Meter dataset is provided by Irish Social Science Data Archive's (ISSDA), Ireland. The dataset used for the experiment is the subset of smart meter dataset of Ireland in December 2010. The data set considered in this work is residential smart meter data. The data contains the information about the customer id, code for date/time, electricity consumption for every 30 minutes (in KWh). The daily profile for every customer comprises of 48 power consumption readings. Table 2 shows the sample data obtained from a customer. This work requires the aggregated data values. The electricity consumption data obtained for every day per customer is 48 readings. These 48 readings of a customer are averaged for per day power consumption data value. For the experimental purpose, a sample dataset comprises of 4 weeks power consumption reading of the customers is considered. Equ 2 represents the power usage value calculation of the customer per day.

Table: Smart Meter Dataset

Customer Id	Code Date/time	Power consumption units (kwh)
1001		0.654
1001		:
:		:
1001		0.82

$$P_{Ci\_Dk}= \sum_{j=1}^{48} P_{Ci\_Rj} \qquad (2)$$

Where,

Ci = customer ID

Rj = j^{th} reading j=1,2,….48

P_{Ci_Rj}= Power usage of per Meter reading

P_{Ci_Dk}= Power usage of a customer per day

Fig. 3 represents the power usage pattern of a customer for the considered period of 4 weeks. Similarly for all customers, the aggregated power consumption data are obtained. The customer's profiles are chosen based on their regular usage pattern in power consumption. Only the trustworthy customer's profiles in power consumption are considered for the analysis purpose. Also the customer behavior in power consumption may differ based on variety of feature such as season, special occasion, Temperature, working days, Sabbath day. Fig. 4 represents the customer profile cluster using k mean clustering.

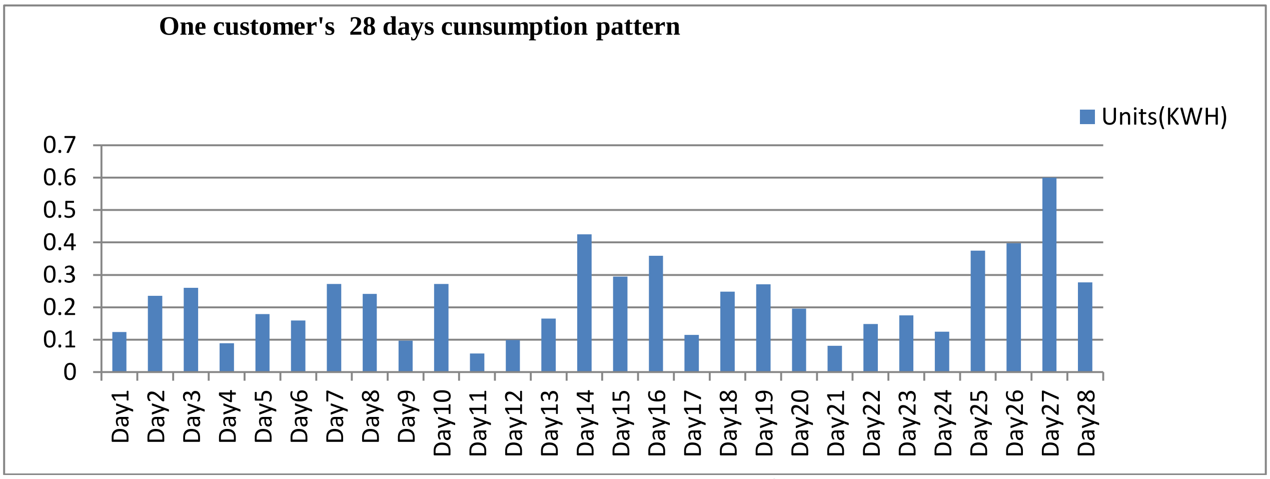

Fig. 3: Sample customer power consumption for 4 weeks

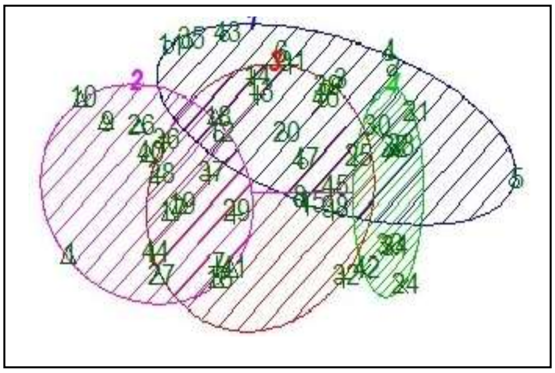

Fig. 4: K-menas cluster based on 28 days power usage data

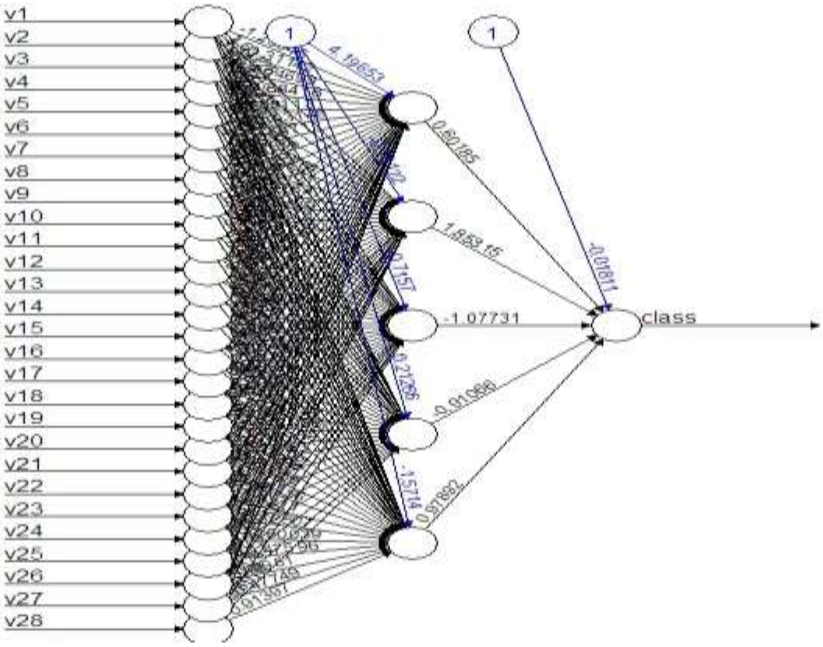

Fig. 5: Neural Network model with 28 days smart meter reading per customer

Fig 5 represents the neural network formed by our proposed system. The smart meter readings of the customer are the input neurons. The neural network is modelled with one hidden layer. The weights are assigned to each of the input neuron. It is adjusted during the number of iterations in training phase. The neural network predicts the class value of all the data. The root mean square error is calculated to evaluate the performance of neural network.Accuracy Percentage of correctly classified records.Accuracy is Sum of correctly classified value divided by total number of classification. The Accuracy (AC) depends on the two factors: Sensitivity (SE) and Specificity (SP). The accuracy of the proposed system is the percentage of correctly classified records. By considering equation, the proposed system provides **97%** accuracy.

AC = (TP+TN) / (TN + FP + TP + FN) (4)

Sensitivity (SE): Recall is commonly called as Sensitivity. It is the True Positive Rate (TPR).

SE = TP / (TP + FN) (5)

Specificity (SP): Specificity is True Negative Rate (TNR).

SP = TN / (TN +FP) (6)

Precision: Precision is Positive Predictive Value (PPV).

Precision = TP / (TP+FP) (7)

Where,
TP (True Positive) is number of records correctly classified
TN (True Negatives) is number of records correctly classified

as abnormal
FN (False Negatives) is number of records misclassified as abnormal
FP (False Positives) is number of records misclassified as normal

The performance of the system is calculated using equation 3, 4 and 5. Table 3 represents the output performance measure of each parameter to its respective class value.

Table: performance measure calculation

Class	Precision	Sensitivity	Specificity
0	0.85	0.85	0.96
1	0.87	0.87	0.95
2	0.81	1	0.92
3	1	0.90	1

Table III infer that the performance is calculated for the trained ANN model using the testing dataset. Class 0 is the actual dataset, class 1 is the type 1 bogus dataset, class 2 is type 2 bogus dataset and class 3 is type 3 bogus dataset. The accuracy percentage is obtained as 97%.This paper proposes an efficient way of using machine learning algorithm to address the power theft problem. In smart grid data analytics system, it is necessary to know the real time electricity consumption data to forecast the exact future demand of electricity and plan accordingly. The identification of power theft will also extend its support for load forecasting that permits the utilities to exactly predict the power demand for future specific to individual customer. The information produced through this analytics, increase knowledge of customer usage pattern and the requirement of power for the future.

Acknowledgment

We like to acknowledge with thanks to Irish Social Science Data Archive Center for providing us with the access to the Smart Meter Data Archive that had a vital role for this research work.

References

[1] Patrick Glauner et al. "Large-scale detection of non-technical losses in imbalanced data sets". In: Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society. IEEE. 2016, pp. 1–5

[2] [2]S.S.S.R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: overview,issues, prevention and a smart meter based approach to control theft," Energy Policy, vol. 39, pp. 1007–1015, Feb. 2011.

[3] Blue Yonder GmbH Maximilian Christ. How to add a custom feature. tsfresh. May 28, 2017. url: http : / / tsfresh . readthedocs . io / en/latest/text/how_to_add_custom_feature.html

[4] Patrick Glauner et al. "The Challenge of Non-Technical Loss DetectionUsing Artificial Intelligence: A Survey". In: International Journal of Computational Intelligence Systems 10.1 (2017), pp. 760–775

[5] S. McLaughlin , B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz , "A multi-sensor energy theft detection framework for advanced metering infrastructures," IEEE J. Sel. Areas Commun., vol. 31, no. 7, pp. 1319–1330, Jul. 2013.

[6] J. I. Guerrero , C. Leon, I. Monedero, F. Biscarri, and J. Biscarri , "Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection." Knowl.-Based Syst. , vol. 71, pp. 376–388, 2014.

[7] P. Jokar, N. Arianpoo, V.C.M. Leung, "Electricity theft detection in AMI using customers' consumption patterns", *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216-226, 2016.

[8] Srihari Mandava, Vanishree J, Ramesh V, "Automation of Power Theft Detection Using PNN Classifier" International Journal of Artificial Intelligence and Mechatronics, Volume 3, Issue 4, ISSN 2320 – 512.

[9] J. R. Filho, E. M. Gontijo, A. C. Delaiba, E. Mazina, J. E. Cabral, and J. O. P. Pinto, "Fraud Identification in Electricity Company Customers Using Decision Trees" in Proc. of 2004 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, pp. 3730-3734, Oct. 2004.

[10] Technique for identifying abnormal energy usage pattern, in Proc. IEEE North American Power Symposium (NAPS), 2012, pp. 1-6

[11] Sanujit Sahoo, Daniel Nikovski, Toru Muso, and Kaoru Tsuru. Electricity theft detection using smart meter data. In Innovative Smart Grid Technologies Conference (ISGT), 2015 IEEE Power & Energy Society, pages 1–5. IEEE, 2015.

[12] Vladimir N Vapnik. An overview of statistical learning theory. IEEE transactions on neural networks, 10(5):988–999, 1999

[13] Soma Shekara Sreenadh Reddy Depuru, Lingfeng Wang, Vijay Devabhaktuni, and Robert C Green. High performance computing for detection of electricity theft. International Journal of Electrical Power & Energy Systems, 47:21–30, 2013

[14] Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Syed Khaleel Ahmed, and Farrukh Nagi. Improving svm-based nontechnical loss detection in power utility using the fuzzy inference system. IEEE Transactions on power delivery, 26(2):1284–1285, 2011

[15] Breno C Costa, Bruno LA Alberto, Andre M Portela, ´ W Maduro, and Esdras O Eler. Fraud detection in electric power distribution networks using an annbased knowledge-discovery process. International Journal of Artificial Intelligence & Applications, 4(6):17, 2013.

[16] AH Nizar, ZY Dong, and Y Wang. Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 23(3):946–955, 2008.

[17] Cyro Muniz, Marley Maria Bernardes Rebuzzi Vellasco, Ricardo Tanscheit, and Karla Figueiredo. A neuro-fuzzy system for fraud detection in electricity distribution. In IFSA/EUSFLAT Conf., pages 1096– 1101. Citeseer, 2009.

[18] Cyro Muniz, Karla Figueiredo, Marley Vellasco, Gustavo Chavez, and Marco Pacheco. Irregularity detection on low tension electric installations by neural network ensembles. In 2009 International Joint Conference on Neural Networks, pages 2176–2182. IEEE, 2009.

[19] Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Syed Khaleel Ahmed, and Malik Mohamad. Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE transactions on Power Delivery, 25(2):1162–1171, 2010.