Machine learning Techniques for Energy Theft Detection in AMI

Assia Maamar
Faculty of technology
Laboratory of Energetic in arid zones
Tahri Mohammed University of Bechar, Algeria
assia_etud@yahoo.fr

Khelifa Benahmed
Faculty of Exact sciences
Department of mathematics and computer science
Tahri Mohammed University of
Bechar, Algeria

ABSTRACT
Advanced Metering Infrastructure (AMI and smart meter) is considered as the basic building block for the development of smart grid in the power distribution system. a Smart meter is one of the keys elements of Advanced Metering Infrastructure, it provides two-way communication between customer and electricity utility, Smart Meters send consumption data frequently (e.g., every 15 minutes) to the utility for monitoring and billing, therefore, a gold mine of data is generated for utilities. Smart meters have become a major focus for targeted attacks which lead to the energy theft , resulting in losses of billions of dollars per year in many countries. Therefore, multitude of papers have studied the energy theft detection by applying different disciplines on smart meter data. In this paper, we present an overview of machine learning research in energy theft detection using smart meter data. It then surveys these research efforts in a summary and comparison of learning models used, in terms of performance metrics, simulation and analysis environment, and data sets used. It finally highlights the challenges in energy theft detection .We approve that these challenges have not been adequately addressed and considered in previous contributions, also covering them, is necessary to advance the energy theft detection.

CCS Concepts
• Hardware → Power and energy → Energy distribution → Smart grid • Computing methodologies → Machine learning.

Keywords
Advanced Metering Infrastructure (AMI); Energy theft detector (ETD); Machine Learning; Energy theft; Smart meter; Smart Meter Data.

1. INTRODUCTION
Nowadays, the use of electricity has become essential for individuals as well as for the industrial sector. In the current electricity systems, several challenges are confronted such as planning, management, and reliability.

ICSIM2018, January 4–6, 2018, Casablanca, Morocco

ACM ISBN 978-1-4503-5438-7/18/01…$15.00
https://doi.org/10.1145/3178461.3178484

AMI is a system that measures, collect, transfer and analyze energy usage and power quality from the smart meters, and communicate with metering devices. AMI provides intelligent connections between consumers and system operators based on the information exchanged between the two sides. smart meter is one of the keys elements of Advanced Metering Infrastructure, it serves as a gateway between utility and, although the basic purpose of meter is for energy measurement, however smart meters generates lots of data which are considering as "Gold Mine of Data" for utilities, since it enables higher resolution for entire electricity delivery system. Smart meter data has an important role in several smart grid applications and enables novel drive Analytics from data to provide solutions for many concerns in electricity delivery system, such as:

- customer profiling and segmentation.
- Real-time alerts & notifications on energy consumption pattern.
- Identify lost revenue from energy theft or billing problems.
- pricing analysis , load modeling end forecasting.
- demand response evaluation.
- Predictive Analytics for energy usage, distribution planning, leakage, device battery status, theft etc.

ENERGY theft was the major issue in traditional power systems in the global. In the U.S, Individually, the lowest estimate shows electricity theft still costs consumers and utilities well over $1 billion each year , theft of electricity and gas cost UK energy consumers £ 299 million every year[1]. AMI technologies and smart meters have been rolled out by electric utilities to upgrade the power grid and mitigate the risk of energy theft through its monitoring capabilities and the fine grained consummation measurements. In 2010, however, the Cyber Intelligence Section of the FBI reported that an organized energy theft attempt against smart meters in Puerto Rico, causing electricity theft amounting to annual losses for the utility estimated at $400 million.[2].in addition, penetration tests have detected various vulnerabilities in the smart meters. Therefore, It is substantial to develop efficient and credible detection system, which can efficiently reveal the energy theft threats against AMI. Current methods and models proposed for energy theft detection are mainly categorized into three groups [3] : 1) state; 2) game theory; and 3) classification based. This paper highlights methods and models classification based, which adopt machine learning paradigm.

The rest of this paper is organized as follows. Section 2 describes machines learning paradigms. Section 3, identify the key challenges of this field that need to be accurately studied in order to enhance methods in the future. Section 4 provides a detailed review of energy theft detection research which employing large

scale machine learning algorithms on the smart metering data. Section 5 presents the conclusion.

2. MACHINE LEARNING

Machine Learning (ML) grew out of the field of artificial intelligence (AI) and is the science of getting computers to learn from and make predictions on data, without being explicitly programmed. In other words, machine learning algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions [4].

A subfield of machine learning is Data Mining, where human-understandable knowledge can be discovered from huge datasets such as web click data or medical records. Moreover, there are some applications that cannot be programmed by hand but have to be learned by experience as a human would do. This is, for instance, the case of handwriting recognition, most of Natural Language Processing (NLP), Computer Vision or even an autonomous helicopter. Indeed, it would be unimaginable to write a program for each single customer, especially as people's preferences vary over time. A long-term ambition of machine learning is to build an artificial intelligence able to mimic how the human brain works.

Machine learning's tasks can typically be classified into four categories, depending on the nature of the available data[5]:

Supervised learning: A supervised learning algorithm is presented with a sequence of so-called labeled instances, i.e., the machine gets input variables or features, as well as output variables and must learn to produce the correct output given a new unlabeled input. Examples fed into the learning algorithm are called the training dataset. In the case of classification, the desired qualitative output variables are discrete labels and the goal is to correctly categorize new instances. For example, given some information about patients diagnosed with cancer like a tumor size or the age of the patient, one can predict if the cancer is malignant or benign. Conversely, when talking about regression, quantitative outputs are a function of the input values. The training dataset defines thus the regression function. Estimating housing prices with the help of features like the size, the location or the year of construction falls into this category.

Unsupervised learning: Training examples do not have any labels in unsupervised learning. Instead of classifying instances or building a regression function, the learning algorithm is asked to find some hidden structure in the dataset. The most common case is the clustering algorithm that assembles instances in separate clusters according to their similarities. Clustering is actually used for a bunch of applications, e.g., computer clusters, social network analysis, market segmentation, astronomical data analysis, etc.

Semi-supervised learning: Since the labeling process is often costly, only a small subset of instances can be labeled, where easily accessible unlabeled data support classification or regression. This is a compromise between supervised and unsupervised learning.

Reinforcement learning: The machine can also produce actions that affect the state of a dynamic environment which gives in return rewards (or punishments) that it tries to maximize (or minimize). This kind of machine learning is used to attain a certain goal such as driving a vehicle or playing a game against an opponent.

2.1 Performance measures

The first way to evaluate the learning systems (classifiers) is to compare the observed values of the dependent variable Y with the predicted values Y' provided by the model. Most of performance measures for two-class problems are built over a 2 × 2 confusion matrix as illustrated in Table 1 Several ratios summarizing the performance measures of the classifiers are deduced. From the confusion matrix, four simple measures can be directly obtained: True Positive (TP) and True Negative (TN) denote the number of positive and negative cases correctly classified, while False Positive (FP) and False Negative (FN) refer to the number of misclassified positive and negative examples, respectively. The most widely used metrics for measuring the performance of learning systems are the error rate (ERR) and the accuracy (ACC), defined respectively as:

$$ERR = (FP + FN)/(TP + FN + TN + FP) \quad (1)$$

$$ACC = (TP + TN)/(TP + FN + TN + FP) \quad (2)$$

Table 1. Confusion matrix

Actual Value	**Predicted Value**	
	Positive	***Negative***
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

However, it has widely been demonstrated that, when the prior class probabilities are very different, these measures are not appropriate because they do not consider misclassification costs, are strongly biased to favor the majority class, and are sensitive to class skews [6]. For example, consider a problem where only 1% of the instances are positive. In such a situation, a simple strategy of labelling all new objects as negative would give a predictive accuracy of 99%,but failing on all positive cases. In the past few years, several new metrics which measure the classification performance on majority and minority classes independently, hence taking into account the class imbalance, have been proposed. The true positive rate (TPR), also referred to as recall or sensitivity, is the percentage of correctly classified positive instances, defined as :

$$TPR = TP/(TP + FN) \quad (3)$$

Analogously, the true negative rate (TNR)(or specificity), is the percentage of correctly classified negative examples, defined as:

$$TNR = TN/(TN + FP) \quad (4)$$

The false positive rate(FPR), refers to the percentage of misclassified positive examples, defined as:

$$FPR = FP/(FP + TN) \quad (5)$$

The false negative rate(FNR), is the percentage of misclassified negative examples, defined as :

$$FNR = FN/(TP + FN) \quad (6)$$

A way to combine the TP and FP rates consists of using the ROC curve A receiver operating characteristic (ROC). The ROC curve is a two-dimensional graph to visualize, organize and select classifiers based on their performance. It also depicts trade-offs between benefits (true positives) and costs (false positives) [6,7]. In the ROC curve, the TP rate is represented on the Y-axis and the

FP rate on the X-axis. Several points on a ROC graph should be noted. The lower left point (0,0) represents that the classifier labeled all examples as negative the upper right point (1,1) is the case where all examples are classified as positive, the point (0,1) represents perfect classification, and the line y = x defines the strategy of randomly guessing the class. In order for assessing the overall performance of a classifier, one can measure the fraction of the total area that falls under the ROC curve (AUC). AUC varies between 0 and +1. Larger AUC values indicate generally better classifier performance. Several investigations establish that those measures being independent of class priors present a disadvantage in imbalanced environments.

Consequently, it is used the precision (or purity):

$$\text{Precision} = TP/(TP + FP) \quad (7)$$

which is defined as the proportion of positive instances that are actually correct. This measure, in combination with recall, employs a ROC analysis methodology [6,7].

2.2 AMI System and Machine Learning Process

As previously mentioned, the theft of electricity is a global concern, crimes such as meter tampering and multi-party fraud schemes are both easy to commit and hard to detect. By analyzing historic usage data along with customer class and usage profiles, analytic tools and Machine Learning techniques can also identify service failures and probable theft situations.

Figure 1 presents the energy theft detection process , electric utilities are leveraging data collected by the Advanced Metering Infrastructure (AMI) and using machine learning ,data analytics to identify abnormal consumption trends and possible fraud.

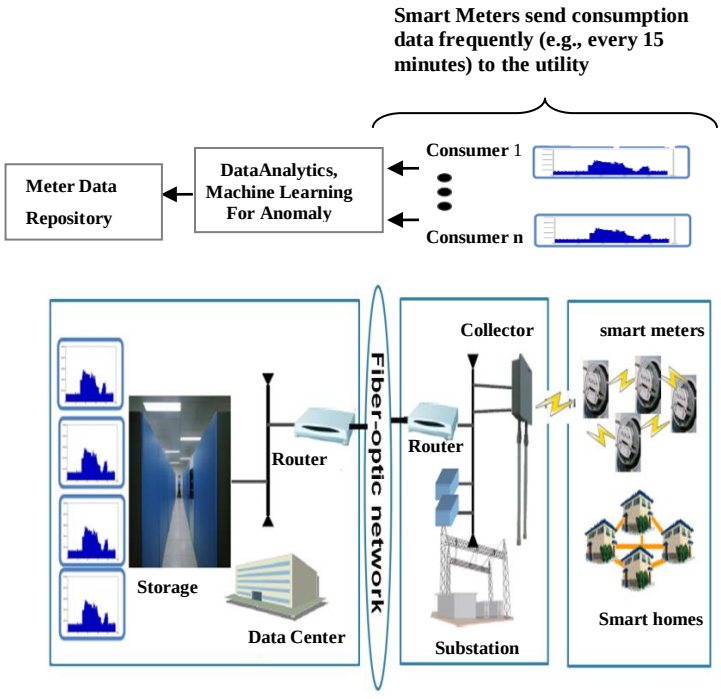

Figure 1. Data collection, transmission, storage, and Data Analytics to Identify energy theft.

3. CHALLENGES

Machine learning technologies are used to learn and train a building a model based on a sample database, which is then utilized to find abnormal patterns. Since these techniques take advantage of the readily available smart meter data, their costs are moderate. However, several challenges in existing energy theft classifiers limit their detection rate (DR) and cause a high false positive rate (FPR). The following diagram presents the most challenges issue that affects the performance of machine learning classifiers, which are necessary in order to advance in energy theft detection.

4. THE STATE OF THE ART

This section, reviews existing electricity theft detector in the literature, which employ smart meters consumption data. Recently, many machine learning models have been proposed to ensure and improve the efficiency of energy theft detection system; these models can be classified into two categories: (1) Simple models and (2) Hybrid models .

Simple models: it uses a single algorithm either supervised or unsupervised[10,11,13,15,16,20,21,23]

Hybrid models :it combines machine learning paradigms (supervised learning and unsupervised learning) [12,14,17,18,19,22,24] , hybrid models usually provide higher performance than models established with a single algorithm .

Table2 presents a summary and a comparison of selected learning models of electricity theft detector in terms of performance measures, simulation environment and data sets used.

Table 2. Summary of performance measures, data set features

Ref	Model	Compared With	Accuracy	Recall	Data set					
					Source	Type	Length	Recording Frequency	Number of customers	Simulation Environment
[10]	Support Vector Machines(SVM)	-	72.60%	53%	TNBD[a] data set	Residential	25 months	Monthly	186,968	LIBSVM v2.86
[11]	Neural networks (NN)	-	83.5%	24.9% 29.8%	Light S.A. Company, Brazil	Mix Residential	4 years	Monthly	-	-
[12]	Neuro-fuzzy	NN	68.2%	51.2%	Light S.A. Company, Brazil	Mix	4 years	Monthly	-	-
[13]	Extreme learning machines (ELM)	SVM	54.61%	-	TNBD data set	Residential	-	30 min	1500	LIBSVM MATLAB Weka
[14]	Genetic- SVM	SVM	-	62%	TNBD data set	Residential	25 months	Monthly	186,968	
[15]	SVM (Gauss)		77.41%	64%	TNBD data set	Residential	25 months	Monthly	186,968	LIBSVM v2.86
[16]	Fuzzy logic	-	-	55%	TNB Metering Services database	Residential	1 month	30 min	-	-
[17]	SVM -fuzzy	-	-	72%	TNBD data set	Residential	25 months	Monthly	36173	-
[18]	Fuzzy- kmeans	-	-	14.2%	-	Mix	6 months	Monthly	20126	Oracle(DBMS) C++
[19]	AutoRegressive Moving Average (ARMA) models	Average CUSUM EWMA LOF	-	62%	-	Residential	6 months	15 min	108	-
[20]	NN	-	93.75	78.94	Dataset collected by CER[b] in Ireland.	Mix	18 months	30 min	5000	WEKA
[21]	Decision tree	-	-	-	Dataset collected by CER in Ireland.	Mix	18 months	30 min	5000	WEKA
[22]	Principal Component Analysis (PCA)	Average Detector	-	-	Dataset collected by CER in Ireland.	Mix	18 months	30 min	5000	-
[23]	Boolean Rules fuzzy logic SVM	-	-	-	Electricity provider in Brazil	-	48 months	Monthly	700K customer	MATLAB LIBSVM
[24]	Kmean-SVM	ARMA models	-	94%	Dataset collected by CER in Ireland.	Mix	18 months	30 min	5000	-

[a]Tenaga Nasional Berhad Division (TNBD)
[b]Commission for Energy Regulation (CER)

5. CONCLUSION

Energy theft is a major problem in the traditional grid, end even with the development of advanced metering infrastructure in smart grid, more complicated situation in energy theft has emerged. Energy theft can take many forms: while traditional mechanical meters can only be compromised through physical tampering, in AMI the metering data can be tampered with, both locally and remotely. Based on the consumption data generated by the meters, and by applying machine learning algorithm, it is possible to detect occurrences of fraud, and in particular theft of electricity, in smart grid. In the literature, a vast variety of Energy theft detector which employing machine learning methods are reported. Over the past years, machine learning methods have become more popular. The most commonly used methods are support vector machines and neural networks. The performance of these models is influenced by many factors: feature selection: it consists of finding an optimal subset of "attributes", i.e., features, in the dataset that optimizes the probability of success in the subsequent machine learning task. Also, the nature of dataset affects the choice of evaluation metrics which affects performance evaluation of ETD. Finally, we discuss the challenging issues in energy theft detection : handling imbalanced classes in the training data and choosing appropriate evaluation metrics, describing features from the data, building models scalable to Big Data sets and making results obtained through different methods comparable. Therefore, it is necessary to address these challenges accurately in future research in order to advance in the detection of energy theft.

6. REFERENCES

[1] David McCarty, July2017. *A Study on Video IBM Energy Theft* Identification *and Control*.

[2] https://www.csoonline.com/article/2222111/microsoft-subnet/fbi-warns-smart-meter-hacking-may-cost-utility-companies--400-million-a-year.html.

[3] Jiang, R. et *al.* 2014. Energy-theft detection issues for advanced metering infrastructure in smart grid. *Tsinqhua Sci. Technol.* 19, 2, 105–120. DOI= 10.1109/TST.2014.6787363.

[4] Wikipedia. Machine learning. [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning.

[5] Thierry Z. September 2015. SmartMetering Data Analysis by Machine Learning Techniques. Master's Thesis.The Power Systems Laboratoryof the Department of InformationTechnology and Electrical Engineering at the ETH Zurich.

[6] García,V., Sánchez, J.S. , Mollineda, R.A. , Alejo, R. and Sotoca, J.M. 2007. The class imbalance problem in pattern classification and learning. *In Proceedings of the National Data Mining And Learning Workshop (TAMIDA 2007).*

[7] Martino, M. D. , Decia, F. , Molinelli, J. and Fernández, A. 2012. Improving electric fraud detection using class imbalance strategies. , *In Proceedings of ICPRAM* (Vilamoura, Portugal, 2012). 2.135–141.

[8] Glauner, P., Boechat, A. , Dolberg, L. , Meira, J. , State, R. , Bettinger, F., Rangoni, Y. and Duarte, D. 2017. The Challenge of Non-Technical Loss Detection using Artificial Intelligence: A Survey. *International Journal of Computational Intelligence Systems (IJCIS)*.10, 1. 760-775.

DOI= 10.2991/ijcis.2017.10.1.51.

[9] Axelsson, S. 2000. The base-rate fallacy and the difficulty of intrusion detection. *ACM Trans. Inf. Syst. Security*.3, 3,. 186–205. Doi=10.1145/357830.357849.

[10] Nagi, J. , Mohamad, A. M. , Yap ,K. S. and Tiong, S.K. Non-Technical Loss analysis for detection of electricity theft using support vector machines. *In Proceedings of IEEE 2nd International Power and Energy Conference*.(Malaysia, December 2008).907-912.

DOI: 10.1109/PECON.2008.4762604.

[11] Muniz, C. , Figueiredo, K. , Vellasco, M. ,Chavez, G. and M. Pacheco. 2009. Irregularity Detection on Low Tension Electric Installations by Neural Network Ensembles. In *Proceedings of the IEEE - International Joint Conference on Neural Networks*, (Atlanta, Georgia, USA June 2009), 2176-2182. DOI=10.1109/IJCNN.2009.5178985.

[12] Muniz, C. , Vellasco, M. , Tanscheit R. and Figueiredo, K. 2009. Neuro-fuzzy System for Fraud Detection in Electricity Distribution. *In Proceedings of IFSA/EUSFLAT Conference*. 1096-1101. (Lisbon, Portugal, July 20-24, 2009).

[13] Nizar, A. H. , Dong Z. Y. and Wang, Y. 2008. Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method. *IEEE Transactions on Power Systems*. 23, 3.946-955. DOI=10.1109/TPWRS.2008.926431.

[14] Nagi, J. , Yap, K. S. , Tiong, S. K. , Ahmed, S S. K. and Mohammad, A. M. 2008. Detection of abnormalities and electricity theft using genetic support vector machines. *In Proceedings of TENCON IEEE Region Conference.*

(Hyderabad, India, 19-21 Nov. 2008*), 1-6.* DOI=10.1109/TENCON.2008.4766403.

[15] Nagi, J. , Yap, K. S., Tiong, S. K. , Ahmed, S. K. and Mohamad, M. 2010. Nontechnical loss detection for metered customers in power utility using support vector machines, *IEEE Transactions on Power Delivery*.25, 2. 1162-1171. DOI: 10.1109/TPWRD.2009.2030890.

[16] Nagi, J. , Yap, K. S. , Nagi, F. , Tiong , S. K., Koh, S. P. Ahmed, S. K. 2010. NTL Detection of Electricity Theft and Abnormalities for Large Power Consumers In TNB Malaysia. *In Proceedings of IEEE Student Conference on Research and Development* (Putrajaya, Malaysia, 2010). 202-206. DOI=10.1109/SCORED.2010.5704002.

[17] Nagi, J. Yap, K. S. , Tiong, S. K. , Ahmed, S. K. and Nagi, F. 2011. Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System. *In Proceedings of IEEE Transactions on Power Delivery*.26,2.1284-1285.
DOI=10.1109/TPWRD.2010.2055670.

[18] dos Angelos, E. W. S. , Saavedra, O. R. , Carmona Cortes ,O. A. and Nunes de Souza, A. 2011. Detection and identification of abnormalities in customer consumptions in power distribution systems. *IEEE Transactions on Power Delivery*.26,4.2436-2442.
DOI: 10.1109/TPWRD.2011.2161621.

[19] Mashima D. and Cárdenas, A. A. 2012. Evaluating electricity theft detectors in smart grid networks.*In Proceedings of* Research in Attacks, Intrusions, and Defenses. *Springer*.210-229.

[20] Ford, V. , Siraj A., Eberle, W. 2014. Smart Grid Energy Fraud Detection Using Artificial Neural Networks. *In Proceedings of IEEE Symposium on Computational Intelligence* Applications *in Smart Grid* (Orlando, FL, USA, 2014). DOI: 10.1109/CIASG.2014.7011557.

[21] Cody C., Ford, V. , Siraj, A. 2015 . Decision Tree Learning for Fraud Detection in Consumer Energy Consumption. *In Proceedings of IEEE, 14th International Conference on Machine Learning and Applications* (Miami, FL, USA 2015). DOI: 10.1109/ICMLA.2015.80.

[22] Krishna, V. B. , Weaver, G. A. , Sanders, W. H. 2015. PCA-Based Method for Detecting Integrity Attacks on Advanced Metering Infrastructure.*In Proceedings of 12th International Conference on Quantitative Evaluation of Systems* (Madrid, Spain,September 01 - 03, 2015). Springer. 9259. 70-85. Doi=10.1007/978-3-319-22264-6_5.

[23] Glauner, P. , Boechat, A. , Dolberg, L. , State, R. , Bettinger, F. , Rangoni, Y. and Duarte, D. 2016. Large-Scale Detection of Non-Technical Losses in Imbalanced Data Sets. *In Proceedings of Seventh IEEE Conference on Innovative Smart Grid Technologies*.

[24] Jokar, P. , Arianpoo, N.and Leung, V. C. M. 2016. Electricity Theft Detection in AMI Using Customers' Consumption Patterns, *IEEE Transactions on Smart Grid.*7, 1. 216 - 226 . DOI: 10.1109/TSG.2015.2425222.