APPLICATION OF IMPORTANCE SAMPLING FOR DETECTION OF NON-TECHNICAL LOSSES IN ELECTRICAL DISTRIBUTION SYSTEMS USING SMART METERS

Mojtaba KHEDEERZADEH
Shahid Beheshti University – IRAN
m_khedezadeh@sbu.ac.ir

ABSTRACT

Electricity theft as one of the primary non-technical losses (NTL) not only causes huge revenue losses but also can result in the surging electricity, the heavy loading of electrical systems, and dangers to public safety. Smart grids help to mitigate the impacts of NTL by integrating information flows with energy flows. Although the availability of massive data generated from smart meters is beneficial in detecting electricity theft, discrimination between honest and fraudulent customers is a difficult task, because fraudulent consumers remarkably outnumber non-fraudulent ones. Therefore, the dataset has an imbalance nature. In this paper, importance sampling method is used to detect irregular consumption, since the accurate estimation of probabilities of rare events is a primary concern of this method. Rare events are almost always defined on the tails of probability density functions. They have small probabilities and occur infrequently in real applications such as the case of benign and fraud consumption. The rare events can be made to occur more often by deliberately introducing changes in the probability distributions that govern their behaviour. As the fraud samples are rare events with very low probability, so importance sampling is well exploited to reduce the variance of these samples. A comprehensive survey is presented regarding the state-of-the-art methodologies to detect NTL. Different methods are proposed to detect NTL. The state-of-the-art methodologies to detect NTL could be categorized into: statistical approaches; expert systems; and data mining and machine learning techniques. Statistical approaches are based on state estimation and analysis of variance to identify consumers with the suspicious of having metering problems. Evidences of NTL are detected quantitatively by estimating the distribution transformer load to detect meter malfunction tampering using smart meters. Different methods are proposed to detect NTL. The state-of-the-art methodologies to detect NTL could be categorized into: statistical approaches; expert systems; and data mining and machine learning techniques. Statistical approaches are based on state estimation and analysis of variance to identify consumers with the suspicious of having metering problems. Evidences of NTL are detected quantitatively by estimating the distribution transformer load to detect meter malfunction tampering using smart meters. Decision trees [6]; artificial neural networks (ANNs) [7]; and Bayesian networks [6] are examples of data mining studies used to identify and predict fraud in distribution systems.

INTRODUCTION

Distribution grid losses could be classified as technical and non-technical losses (NTL). Technical losses include: generation losses, line losses, and copper and iron losses of transformers. NTL refer to a type of loss generally involving electricity theft, errors in meter readings, or defective meters. Electricity theft/fraud can be defined as a dishonest/illegal use of electricity equipment or service with the intention to avoid billing charge. Examples of NTL include meter tampering to record lower consumptions; bypassing meters by rigging lines from the power source; false meter readings by bribing meter readers; faulty/broken meters; un-metered supply; technical/human errors in meter readings, data processing and billing [1]. Power utilities lose large amounts of money each year, so NTL is a major concern in traditional power systems worldwide. Energy theft alone has cost the utility companies in the U.S. around $6 billion per year [2]. In another report [3], the loss is amounted between 0.5% and 3.5% of annual gross revenues in the U.S. Although the absolute percent seems a small amount, considering the U.S. electricity revenues as $280 billion range in 1998, it is realized as between $1 and $10 billion worth of electricity that is stolen. Different methods are proposed to detect NTL. The state-of-the-art methodologies to detect NTL could be categorized into: statistical approaches; expert systems; and data mining and machine learning techniques. Statistical approaches are based on state estimation and analysis of variance to identify consumers with the suspicious of having metering problems. Evidences of NTL are detected quantitatively by estimating the distribution transformer load to detect meter malfunction tampering using smart meters. Decision trees [6]; artificial neural networks (ANNs) [7]; and Bayesian networks [6] are examples of data mining studies used to identify and predict fraud in distribution systems.

Load profile, i.e., the pattern of electricity consumption of a customer over a specified period of time is widely used as an indicator of fraud; since abrupt changes of the consumption could be considered as electricity theft [8]. A methodology for assigning the typical load profiles to a particular group of consumers based on the consumer’s type of activity using probabilistic neural network is presented. In [9], an integrated expert system for analysis and classification the available useful information of the customer is presented. Customers with NTL are identified. Text mining module, data mining module, and the Rule Based Expert System module are the main modules of the proposed expert system. In [10], support vector machine (SVM) is proposed as an approach towards NTL detection. The fraud detection model preselects suspected customers based on irregularities in consumption behavior. Feature extraction from historical customer consumption data is used in the method. The authors [10] used SVM and the results of real on-field inspections to detect NTL in Malaysia. Onsite inspection toward monitoring fraud activities is a common practice, but increases operational costs with the least success if a preliminary strong screening is not performed. In some utilities, whenever a remarkable increase in the consumption of a customer or null consumption during a certain period is observed, the inspection crew are dispatched. For improvement of this type of inspection, a framework and methodology is proposed in [11], developed as two coordinated modules. The first module filters the customers using a text mining and a complementary ANN. The second module contains a Classification & Regression tree and a Self-Organizing Map neural network. It is claimed that with these modules, the success of the inspections is multiplied by 3. It is worth noting that the availability of the customers’
consumption is the main point in selecting a method for monitoring NTL. For example, there are still many traditional meters in distribution utilities, which are read monthly or bi-monthly by meter readers. Based on this data, features such as: average consumption, maximum consumption, standard deviation, number of inspections and the average consumption of the residential area are extracted and used for fraud detection. Smart meters are the most effective devices that could be used to reduce NTL, as they make fraudulent activities more difficult and easier to detect [12]. With the increasing availability of smart meter devices, consumption of electric energy in short intervals can be recorded. Consumption features of intervals of 15 minutes, 30 minutes and 60 minutes are used in literature. In [12], a methodology for detection of NTL using supervised learning is proposed. This methodology uses the information that are recorded by the smart meters such as energy consumption, alarms and electricity tampering to obtain a profile about the customer’s consumption behavior. The geographical location and technical characteristics of each smart meter is also used in the process of detecting NTL. Gradient boosted tree is used as an efficient classifier. In some methods, instead of directly classifying the customers as having a NTL or not, the energy consumption of the customers is anticipated. If the difference between the actual and forecasted consumption exceeds a pre-specified threshold, the customer is considered to be committing fraud [12]. In [13], in contrast to building a global model that can be used for all customers, a model is built based on a customer-by-customer basis. SVM is used to distinguish between the normal and fraudulent pattern of the customer.

In [14] a method called Extreme Learning Machine is proposed to improve the classification performance. It is shown that this approach is superior to Support Vector Machine (SVM) algorithm. A pattern recognition technique called optimum-path forest is proposed in [15] that includes learning and pruning algorithms. It is mentioned that traditional artificial intelligence techniques could result in a high computational burden in the training phase and problems with parameter optimization, while the proposed pattern recognition technique called optimum-path forest is superior. In [16] a computational technique for the classification of electricity consumption profiles is proposed comprising of two steps. Firstly, a C-means-based fuzzy clustering is used to find consumers with similar consumption profiles. Secondly, a fuzzy classification is performed to find potential fraudsters with irregular patterns of consumption. In [17] the SVM method in [10] is extended with the introduction of a Fuzzy Inference System. It is indicated that with the implementation of the improved method, the NTL detection is increased from 60% to 72%. Reference [18] proposes a criterion based on the amount of not invoiced electricity due to fraud by the estimation of accurate discretization of the original data. Probabilistic classification techniques could extend the role of traditional approaches that only output integer labels as fraudulent/honest customers into monitoring the behavior of consumers. In [19], advanced metering infrastructure (AMI) is used to detect an energy fraud. A mechanism that incorporates the conditional probability into determination of the normality of the prototype for comparison is proposed. A two-dimensional space using similarity and conditional probability is used, so that multidimensional classification methods could be applied. In [20], an analytical model of the Technical Losses variation versus load change is proposed that estimates adequately the sensitivity of the voltages and their respective angles to the variation of the load including the electrical losses. The variation of the Technical Losses is estimated through the variation of the loads which is used in the Probabilistic Energy Balance for calculating the NTL. In [21], a probabilistic-based Optimum-Path Forest classifier to handle the problem of NTL detection is proposed. The proposed algorithm rules a competition process among some key samples (prototypes) and outputs an optimum path forest, which is a collection of optimum-path trees rooted at each prototype. Each Optimum-Path Forest defines a cluster of samples that belong to the very same class, thus labelling the whole dataset with the label of each prototype sample.

The availability of massive data generated from smart grids can help to solve the NTL problem, because of the abnormal consumption pattern of energy thieves. In [22], a method based on Wide & Deep Convolutional Neural Networks (CNN) model is proposed. The Deep CNN component is used to identify the non-periodicity of NTL and the periodicity of normal electricity usage based on two dimensional consumption data. The Wide component is used to capture the global features of one dimensional consumption data. Two dimensional is evaluated as plotting the consumer consumption for different weeks on the same plot. Non-fraudulent consumers obey a relatively similar pattern in different weeks, while fraudulent ones have irregular weekly patterns. Reference [23] deals with NTL in an interesting point of view, i.e., the recent addition of information and communication technologies in electric power distribution systems. It is noted that a new class of NTL is evolved by hiding and changing the consumption data through cyber-attacks. This could be performed by unauthorized access to the application database and digital tampering of smart meters. A strategy is proposed to detect NTL by using a multivariate control chart for monitoring the measured variance. Next, a pathfinding procedure is used to locate the consumption point with NTL. Geographical information system (GIS) could be used to display the consumption point that is the target of the cyber-attack. In [24], it is indicated that a data analytics method for detecting NTL is required due to the increase of tampering with smart meters, hence a labeled dataset or additional system information is needed which is hard to achieve. It is proposed to combine two data mining techniques, i.e., the Maximum Information Coefficient to precisely detect NTL that appear normal in shapes and a clustering technique to find the abnormal users with arbitrary shapes to solve the problem.

Table 1, shows the performance of the models explained above using various metrics such as true positive rate, known also as the recall, false positive rate, the precision and the AUC score. As the detection of NTL has an imbalanced nature, it could be concluded that the AUC score provides more reliable results, because the quality of customers is assessed instead of their classification [12]. Needless to say that the utility prefers to have a ranked list of customers according to their probability of having the tendency toward NTL rather than a list with all the
customers classified as honest/fraudulent, because it could try to enhance the behavior of those customers.

In this paper, importance sampling method [25] is used to detect consumers with abnormal consumption, as discrimination between honest and fraudulent customers is a difficult task. Fraudulent consumers remarkably outnumber non-fraudulent ones, so the dataset have an imbalanced nature. Benign samples are easily available using historic data. On the other hand, NTL samples are rarely available or do not exist for a given customer. Therefore, malicious consumption patterns could be considered as rare events. Hence, this is the motivation to apply rare-event detection methodologies to the NTL detection problem. When we are confronting a rare event, usually Monte Carlo simulations are used to increase the detection problem. When we are confronting a rare event, it could be

a target distribution is approximated by a weighted average of random draws from another distribution. Importance sampling has been introduced as a variance reduction technique in statistical physics. Nowadays, importance sampling is used in a wide variety of application areas and there have been recent developments involving adaptive versions of the methodology. To the best of the knowledge of the author, no previous research on NTL detection has aimed at extracting the most important patterns from consumer load profiles using importance sampling, in order to cope with the effects of imbalanced data in the classification task. As the fraud samples are rare in comparison to the healthy samples, it could concluded that these cases are rare events with a very low probability, so importance sampling could be well exploited to reduce the variance of these samples.

The proposed methodology is performed by:

- Data acquisition and pre-processing in order to remove bad data and interpolate missing data.
- Extraction of the suspicious consumers from the database by using importance sampling method as an imbalanced-data strategy.
- Validation of the results by using different performance metrics such as accuracy and recall.

It is worth noting that accuracy and recall are the most popular performance measures in the literature, ranging from 0.45 to 0.99 and from 0.29 to 1, respectively.

**PROPOSED METHODOLOGY**

As clearly indicated in [21], probabilistic-driven classification techniques are more beneficial in problems where only detection is not of the main concern. Because punishing the unlawful customers would not completely eradicate the electricity theft, instead monitoring the behavior of consumers in a probabilistic manner is preferable. As the regular customers are much more than customers with NTL, so this class imbalance issue degrades the performance of most supervised machine learning algorithms as a widely-used methodology to identify NTL. This issue is well dealt with in [4] and the inefficient learning of supervised machine learning algorithms when the classes in the training data are highly imbalanced is indicated; because there is only a few samples of the positive class (fraud) in comparison with the negative class (lawful). The underrepresentation of the positive class adversely impacts the classification. The proposed solution to this issue is carried out by resampling techniques, i.e., blending random oversampling with random undersampling. In the reported simulation in [4],

<table>
<thead>
<tr>
<th>Method</th>
<th>NTL Type</th>
<th>Data Source</th>
<th>Customer#</th>
<th>NTL %</th>
<th>Location</th>
<th>Results %</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4]</td>
<td>consumption of each consumer</td>
<td>AMI/881 days</td>
<td>2271</td>
<td>5.2</td>
<td>Honduras</td>
<td>AUC=0.8187, MCC=0.7300</td>
<td>SML/ Wavelet for feature extraction &amp; Random Under-sampling Boosting for NTL detection</td>
</tr>
<tr>
<td>[5]</td>
<td>metering data</td>
<td>Smart Meters</td>
<td>Taiwan</td>
<td>7.504</td>
<td>38.33</td>
<td>AUC=0.91</td>
<td>ANN-based knowledge-discovery process</td>
</tr>
<tr>
<td>[9]</td>
<td>Rule-based</td>
<td>Spain</td>
<td>PRC=77.41</td>
<td>Integrated expert system</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[11]</td>
<td>Null consumption</td>
<td>field inspections</td>
<td>3510</td>
<td>4.67</td>
<td>Spain</td>
<td>AUC=0.75</td>
<td>Text mining, ANN, decision tree, SOM-NN</td>
</tr>
<tr>
<td>[12]</td>
<td>Pattern changes</td>
<td>Smart Meters</td>
<td>5746</td>
<td>5.38-8.37</td>
<td>Spain</td>
<td>AUC=0.91</td>
<td>SVM</td>
</tr>
<tr>
<td>[14]</td>
<td>Pattern changes</td>
<td>database</td>
<td>150</td>
<td>Malaysia</td>
<td>PRC=67.07</td>
<td>extreme learning machine</td>
<td></td>
</tr>
<tr>
<td>[15]</td>
<td>Pattern changes</td>
<td>database</td>
<td>5190+8067</td>
<td>Brazil</td>
<td>PRC=86.62</td>
<td>pattern recognition called optimum-path forest</td>
<td></td>
</tr>
<tr>
<td>[16]</td>
<td>Irregular pattern</td>
<td>database</td>
<td>20126</td>
<td>Brazil</td>
<td>assertiveness = 0.829</td>
<td>Unsupervised cluster-based classification</td>
<td></td>
</tr>
<tr>
<td>[18]</td>
<td>Abrupt changes</td>
<td>billing system</td>
<td>21138</td>
<td>5.6</td>
<td>Serbia</td>
<td>rough set theory</td>
<td></td>
</tr>
<tr>
<td>[19]</td>
<td>abnormality</td>
<td>Smart meters</td>
<td>4245</td>
<td>Korea</td>
<td>PRC=60</td>
<td>inserting conditional probability into normal classification</td>
<td></td>
</tr>
<tr>
<td>[20]</td>
<td>Σ Losses-TL</td>
<td>Feeder-based</td>
<td>Brazil</td>
<td>Probabilistic methodology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>Pattern changes</td>
<td>database</td>
<td>3182+4982</td>
<td>6.19-5.77</td>
<td>Brazil</td>
<td>Probabilistic Optimum-Path Forest classifier</td>
<td></td>
</tr>
<tr>
<td>[22]</td>
<td>Pearson Correlation</td>
<td>dataset</td>
<td>42372</td>
<td>China</td>
<td>AUC=0.7922</td>
<td>Wide &amp; Deep Convolutional Neural Networks</td>
<td></td>
</tr>
<tr>
<td>[23]</td>
<td>Variance</td>
<td>Synthetic meters</td>
<td>834</td>
<td>-</td>
<td>multivariate control chart and A-Star algorithm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[24]</td>
<td>Shape &amp; magnitude</td>
<td>Smart Meters</td>
<td>390</td>
<td>12.8</td>
<td>Ireland</td>
<td>AUC=81.6</td>
<td>Maximum Information Coefficient &amp; clustering</td>
</tr>
</tbody>
</table>

Table 1: Comparison of different methods of NTL detection
nearly two-thirds of the data are initially randomly sampled and used for training the classifiers. The remaining is reserved for testing. Each classification model is trained using different oversampling proportions of the minority class and undersampling of the majority class. In other words, each classifier is trained using zero (no oversampling), five, ten, fifteen, twenty, up to twenty-five times (max oversampling) copies of fraud samples. In this paper, the efficiency of NTL detection using supervised machine learning is improved by:

1) probabilistic-based approaches for NTL identification are used instead of abstract-based machine learning algorithms
2) positive class (fraud) samples are considered as rare events and are identified in the available database using importance sampling to reduce variance

Accurate estimation of probabilities of the rare events is a primary concern of importance sampling. Rare events are almost always defined on the tails of probability density functions. Rare events have small probabilities and occur infrequently in real applications such as the case of benign and fraud consumption. However, these events can be made to occur more often by deliberately introducing changes in the probability distributions that govern their behavior [25].

**IMPORTANCE SAMPLING**

Importance sampling is a procedure that changes the probability density function (pdf) of sampling in such a way that the events with higher contributions to the simulation results have greater occurrence probabilities [25]. The probabilities are altered in order to make the unlikely events more likely. In this way the correct answers are extracted. Electricity thefts are rare events in a distribution system with respect to non-fraudulent consumers. A direct search for these rare events would require an unrealistic amount of computation in the database. In literature 1:25 is mentioned as a real proportion of fraud to non-fraud cases. Under this conditions, the problem of finding fraudulent consumers becomes equivalent to the problem of finding rare-events in combinatorial search spaces. This equivalence suggests that importance sampling techniques, which are vastly successful for solving combinatorial problems, could also be used for efficiently detecting NTL. If we want to compute the expected value of a function $f(x)$ with respect to a probability distribution $p(x)$, in many cases, the integral $\int f(x)p(x)dx$ is intractable due to the complexity of $p(x)$. Importance sampling is a discrete method for approximating $\mathbb{E}_p[f(x)]$ by replacing $p(x)$ with a similar, but easily sampled, distribution $q(x)$ and then correcting for the error introduced by making this switch. Assume that the entire trajectory of a Monte Carlo simulation consists of many paths starting from the bottom of an energy basin and returns to the bottom of the same basin. These paths are undesired paths. Very rarely a path may leave from the bottom of one basin and reach the bottom of another basin; such a path is called a desired path. Desired paths are sampled with very low efficiency. When the problem is formulated in this way, the importance sampling method is used to bias the sampling procedure in such a way that the desired events are sampled with higher efficiency. Figure 1 shows a plot of $p(x)$ and $q(x)$ showing how well the sampling distribution covers $p(x)$ in a classic example. This example is in [https://www.ece.rice.edu/~vc3/elec633/ImportanceSampling.pdf] implementing importance sampling to calculate the expectation of $f(x)$. In Figure 1, $\mu$=0.8. Figure 2 shows the same plot with $\mu$=2.0. As can be seen $q(x)$ is poorly sampling distribution match.

![Figure 1: A plot of $p(x)$ and $q(x)$ indicating the coverage of sampling distribution of $p(x)$.](image1)

![Figure 1: Poor sampling distribution match of $q(x)$.](image2)

**CONCLUSIONS**

The imbalance nature of regular and illegal customers degrades the performance of most supervised machine learning algorithms. This issue is investigated in this paper by using the well-known importance sampling algorithm used to reduce the variance. Under the condition of imbalance proportion of fraudulent to honest customers in the dataset, the training for machine learning would have a weak performance if samples are randomly selected. While selecting an appropriate combination of positive class against negative ones enhances the efficiency of the training. Therefore, the problem of finding fraudulent consumers becomes equivalent to the problem of finding rare-events in combinatorial search spaces. This equivalence suggests that importance sampling techniques, could be used for efficiently detecting NTL. The proposed method is based on probabilistic-based approaches for NTL identification instead of abstract-based machine learning algorithms.
REFERENCES


