Using Classification Data Mining Technique to predict the Non-Technical Losses in Power Systems in Palestine

Rana O Nayfeh 1, Abdallatif Abu-Issa 2

1Master In Computing Program, Birzeit University, rmayfeh@yahoo.com
2Assistant Professor In Computing Program, Birzeit University, abuissa@birzeit.edu

Abstract—Non-technical losses are the problem that concerns all electricity utilities around the world. Data mining algorithms are used to minimize the problem and detect the most proper subscribers for theft; making them the target to the inspection staff periodic detection for the electric meters.

Taking into consideration the specialty of Palestine as the case study, the two major non-technical losses found is tampering with meters or having illegal connection at the network. The second case made the traditional data mining technique that depends on the consumption history of the subscriber little weak. Also the detected frauds admitted that their consumption after fraud was double folds the estimated from the utility; that fact side by side with the fact that most of violators also distribute electricity; makes the consumption record only little weak for detection.

The idea is a classified data mining technique that takes into consideration the general information about the subscribers as influencing factor.

When talking about theft we are talking about illegal electrical consumption despite the way of doing it.

Researchers around the world use the consumption patterns to detect fraud. Locally, people do frauds without having the meter from the beginning that yields to missing patterns records for many cases. The idea is to find the reasons that drive people to theft; that I did and more research can be done also on this cases; and then constructing a database for all the influencing information for theft about the subscriber from the first time apply for the meter. For the cases that have no meter at all and of course the utilities knows nothing about them, the system will help a lot as it will classify any input information to it with either YES or NO fraud.

In this work the RapidMiner software was used to compare the different techniques for classifications: the decision tree with its different types and the Naive Bayes classifier.

Index Terms—Data Mining, Non technical losses, Classification, Fraud Detection.

I INTRODUCTION

500,000 legal subscribers are using electricity in west bank[1] with increment factor of 3.5% yearly; while there were 221,792 subscriber in Gaza in 2014[2] with increment factor of 5% from the previous year.

Electricity is being lost along the system as technical and non-technical losses, the first one depends on the system itself; and it shares a percent of % 10 of the electricity produced [3], such losses could be eliminated using new and well-designed transmission lines and sometimes it can be reduced by transmitting power in the high voltage levels as much as possible.

The second one is the non-technical losses (NTL), which is literally the electricity theft; it is the major problem to electricity utilities not only in the least developed countries but also in the developed countries. For example, in developing countries such as Bangladesh, India, Pakistan and Lebanon, NTL's shares an average between 20% to 30% of the total power generated [4], [5]. Locally many companies deliver electricity to end users; each have its own data for losses, but still the situation is not better; in Gaza; NTL shares 23% of the delivered power, while it is 24% in west bank and 45% in Jerusalem in the areas that follows Jerusalem District Electricity Company (JDECO) [1].

Theft can be done in different ways, the simplest way is just ignoring the bill and by that the amount of energy consumed is known and recorded by the utility; also this way needs only actions from the utility to force people to pay their bills.
The other complicated ways that needs the company estimations to identify the consumption has several ways like meter tampering, bypassing the meter or otherwise making illegal connections.

Most likely subscribers turn from the simple way of theft to the complicated ways as nonpaid bills force the utility to cut the service at some point; and people always prefer the less costly way to turn electricity on.

Manual investigation to detect subscriber’s irregular use of energy is performed by the inception and follows up departments of the different utilities.

In JDECO Company as it is the case study I performed my research on; the meter of the subscriber will have periodic visit every five years, so the inception department at its yearly plan needs to perform a visits to 20,000 meter in its different areas.

As the billing system meters are being checked by the meters readers every month; and they report any misleading behavior of the users or any suspicious conditions around the meter to the inception department, the periodic check focus on the prepaid meter system for their visits.

That way of investigation is slow, hard, costly and is performed randomly as the availability of the human recourses that do it, also it visits every meter without even noticing the subscriber that use it making the process in accurate many times regarding the process. So increasing the speed of investigations and its accuracy is needed to detect the suspicious subscribers and decrease the theft.

In 2014 the inception record of Al-Bireh branch at JDECO company recorded 2,500 electricity theft case in Ramallah and the surrounded villages [6], the strange thing that makes Palestine special case is that 3% of the recorded theft cases was for users that are not registered subscribers at all [6], such cases will never be cached using the recent algorithms that depends on the consumption record as there is no record for these cases.

The new approach of detecting fraud will rely on the general information about the subscribers taken from the subscribers themselves when applying for the service, and also the general information of the building will be taken into consideration.

After a while, and if the approach approves its accuracy; the different electricity utilities will need to update their database to include more information about the subscribers and their facilities.

The research study depends on the data provided from JDECO Company, and to be more specific the data of al-Bireh branch was the base line for the approach.

The utilities take care of the privacy of the cached theft cases information so unfortunately the names was hidden and also any information that might point to a specific user was also hidden, moreover the information recorded in the violation report was not enough to do the analysis.

With the help of the inception department and their violation report a sample sheet of different 150 subscriber record was established and the same sample was used to do the analysis.

II LITERATURE REVIEW
A Review Stage

Rajesh. R, MrsE.SivaSankari and Matheswaran.P [7], conducted a research that was presented on a paper published on the International Journal for Innovative Research in Science & Technology, Volume on February 2015, their data mining technique was the LMP model that is basically the ANN with one input layer for receiving information, one or more hidden layers for processing and output layer for displaying the results.

The fraud detection of customers who have nonzero consumption on the 12 months monitoring period; was detected by their normality and abnormalities behavior of their load profiles that used for training the classifier and then MLP testing was performed from another same sample of data.

By the end the accuracy of their system was 92.6%.

Iñigo Monedero et al, also used the same data as [8]. Endesa Distribution Company’s subscriber data was the input to the system, and the data mining techniques used two algorithms: the first one is for detecting the usual drop on the consumption directly; it was searched along 16 months period the regular consumption of the customer then the drop for 2 months and lastly they monitored the change of consumption range for six months only.

The second monitoring period was short compared to the first one to ensure the detection of the NTL before it widespread.

As subscribers might be clever enough not to do any eye catching lower consumption, the researchers thought of a more effective algorithm that detect the unusually patterns was with some more better modifications, it was designed to detect the progressive drop with the subsequent stabilization. The experiment also conducted on the first and second window the first window was the consumption of the first 18 months, while the second one with the next 6 months.
The first period was long to make sure of the real normal behavior of the consumption, at the same time; the second was short to detect any up normality as soon as possible. With both algorithms that described previously for NTLs, detection was with drops in consumption but not with other patterns which is not enough as the users might follows energy efficiency advises. Therefore, with Bayesian networks and decision tree techniques a complementary method was developed in order to search patterns with different shapes. As a result the success rate of the system was 38%, but the researchers finds it satisfactory compared to the 10% success of the ordinary inspection department and the little information used in the system which is the consumption record and the type of user only.

Aniruddha Kshirsagar and Lalit Dole [9] performed a research on Recognizing the Theft of Identity, the system is used to avoid the duplicates while applying for a credit card or a loan. Data mining algorithms namely communal detection and spike detection used to detect the multiple applicants.

Bhavani Thuraisingham [10] also searched for Data Mining for Counter-Terrorism, the system is a useful tool for detecting and preventing terrorism it used link analysis as the chosen data mining technique.

Eyd Hashem S.Humaid [11] searched Fraud Detection Model for Water Consumption Billing System in municipality of Gaza, SVM data mining technique was used for the classification of patterns into fraud and no fraud detection. Hsinchun Chen et al [12] searched for Crime Data Mining; the system uses many techniques as Entity extraction, Clustering techniques such as “concept space”, Deviation detection, and Classification and String comparator. But the big work for data mining technique was in education as Baradwaj and Pal [13] in their data mining technique choses the ID3 as it is a “simple” decision tree learning algorithm to analyze the students’ performance in a selected course program.

Pandey and Pal [14] in their data mining research used Naive Bayes classification to predict students as performers or underperformers to predict their learning needs. Naive Bayes classification is a simple probability classification technique that assumes all attributes in the database is independent.

Ahmad and Elaraby [15] performed a research that focuses on classification for predicting students’ performance in a selected course program based on previously recorded students’ behavior and activities. They analyzed enrolled students’ data record in a specific course program across 6 years between 2005 and 2010, with multiple attributes collected from the 3 university database. By the end the prediction system was able to predict in different percent’s of accuracy the student final grade in the chosen courses. And like every other prediction systems it “help the student's to improve the student's performance, to identify those students which needed special attention to reduce failing ratio and taking appropriate action at right time”[15].

Yadav, Bhardwaj, and Pal [16] tried multiple decision tree algorithms on an educational data record to a classification model of the educational performance of students.

The study mainly selects the best decision tree algorithm from among mostly used decision tree algorithms, and provides a benchmark to each one of them. The researchers come to a result that CART (Classification and Regression Tree) decision tree classification method worked better on the tested dataset, which was selected based on the produced accuracy and precision using 10-fold cross validations. In their study they also opened the door of the experimental way of choosing the best classifier technique; as they tried multiple ways before deciding CART as the best one from the outputs of the different ways. And logically; like any other ways of taking decisions, it is highly recommended to try all the possible ways and decides the best one from the results and that’s exactly what I did on my own model.

Bhardwaj and Pal [17] used also the Naive Bayes classification method, on a data record for a group of BCA students (Bachelor of Computer Applications) in Dr. R. M. L. Awadh University, Faizabad, India, who appeared for the final examination in 2010.

With a questioner that was filled from the students by the end of the year about personal, psychological, and social issues; was used to identify relations between these factors and the student’s performance and grades.

The results showed that the most influencing factor to the student’s performance in the tested sample was their grades in senior secondary school; that fact approves students with good performance in secondary school, will also have good performance in their Bachelors study.

It was found also that the living area, student other habits, family annual income, student family status and the mother’s qualification also affects students’ educational performance, that’s why expecting the student performance can be done if a data record containing all previous basic information about them was available.

III METHODOLOGY

The goal of this research is to find the relation between the subscribers age, place of living, its educational, financial and social levels etc., and their theft behavior using data mining techniques. Therefore the system will predict the subscriber’s performance. The system will also classify the subscribers into two groups: Peoples who do the theft and others who don’t. So the best datamining technique that can achieve this goal is classification.
A huge data base containing all the possible information about the subscriber will be needed in order to build up a model with high accuracy. The fact that the different electricity companies knows little information about the subscribers divided the research into two sections: one is detecting the real reasons of electricity theft, and then collecting the related information from a sample of real cached cases involved in theft.

To do that a questioner was performed and the data was collected in order to find the relations between the mentioned factors and the violator’s behavior. Then with the help of the inspection department at JDECO a small database sample was built with real data of violators and commitment subscribers. The fraud performance will be as yes and no fraud; that means the person might be able to do the theft or not.

The model was built and tested using the RapidMiner data mining program; using the same sample sheet the program divided the data into groups consists of 10 subscribers, then from each group 9 subscribers will be used for building the model while the last one will be used to test it.

**A Data Collection**

In an unbiased data collection; data of the reasons that drive peoples to energy theft; used in this study was detected through a survey distributed to different peoples in the society as an online survey using survey monkey. The initial size of the data was 500 survey records. Survey monkey enables visibility to 100 records free only, and with the fact that the questioner was in English and the opinions of local people were important; not too much people shows interest.

Then I moved to traditional paper survey that I filled with the reviewers of the inspection department of JDECO and then I filled the results manually; taking the chance to make interviews with real violators.

By the end the real information about violators I collected; as the survey and the interviews approved were:

- Living Place. Both the questioner and the realviolator’s information approved that some places have the electricity theft as cultural behavior.
- Educational level. As 66% of the respondents approved who steals electricity are not aware of its serious risks
- Home Area. 25% of the respondents think there is a connection between higher energy consumption and electricity theft, but the interviews I made with the real theft cases was different and the real cases records showed higher consumption of them.
- Number of family members.
- The wife is working or not.
- Classification of Area (A, B and C). for the level of security as
- Monthly Income.
- Have a home in any other living place.

**B Data Cleaning and Filtering**

Using the filtering option in excel the wrong writing was detected and corrected. Also any violator having missing data fields was excluded from the final sheet of the data mining source file.

**C Classification**

Different types of decision trees will be used mainly for predicting the theft or no theft behavior. Multiple classification techniques are already found in RapidMiner.

For the data set I have; I preferred to take the decision tree as it’s the simplest way of classifications and the Naíve Bayes that assumes all the variables are independent as the real case of the dataset approved.

RapidMiner that will be used in the data mining techniques divides the data input to the system into 10 records groups, 9 of them are used for learning the model and the left one is used for validating the result.

- ID3 operator setting was customized experimentally as Splitting criterion was set to be information gain ratio, Minimal size for split was 4, Minimal leaf size was 1 and Minimal gain was 0.1.
- C4.5 operator setting was customized experimentally as Splitting criterion was set to be information gain ratio, Minimal size for split was 4, Minimal leaf size was 1, Minimal gain was 0.1, Maximal depth was 20 and the confidence was 0.5.
- CHAID operator setting was customized experimentally as Minimal size of split was 4, Minimal leaf size was 2, Minimal gain was 0.1, Maximal depth was 20 and 0.5 was the confidence.
- Naïve Bayes classification is a simple probability classification technique, its name “Naïve” comes from the assumption that all the attributes in the data set are statistically independent from each other as the algorithm assume. “Bayes classification means that the algorithm is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the property given the set of data as evidence or input.

**IV EXPERIMENTAL RESULT**

Different types of decision trees and the naive Bayes techniques mentioned below was tested using the same dataset on the rapidminer program, the experimental results of the tests are shown below:

**A Iterative Dichotomiser 3 (ID3) Decision Tree**

The overall accuracy of the results were 88.95% +/- 8.22%, table (1); still the tree didn’t show logical things and was really huge.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System Accuracy Results Using ID3 Decision Tree.</strong></td>
</tr>
<tr>
<td><strong>Accuracy</strong>: 88.95% +/- 8.22% (mikro: 88.97%)</td>
</tr>
</tbody>
</table>

88
The result shows also that the system worked better with the YES prediction of the fraud rather than the NO prediction as the class Precision shows.

B C4.5 Decision Tree

The overall accuracy of the results were 84.29% +/- 10.77%; Table (2), the precision of the algorithm was lower than the ID3, the recall also was lower but still the tree; figure (2) was logical compared to ID3.

TABLE 2
System Accuracy Results Using C4.5 Decision Tree.

<table>
<thead>
<tr>
<th></th>
<th>True YES</th>
<th>true NO</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. YES</td>
<td>97</td>
<td>9</td>
<td>91.51%</td>
</tr>
<tr>
<td>Pred. NO</td>
<td>7</td>
<td>32</td>
<td>82.05%</td>
</tr>
<tr>
<td>Class recall</td>
<td>93.27%</td>
<td>78.05%</td>
<td></td>
</tr>
</tbody>
</table>

Also the Precision of the No fraud prediction was lower than the YES and the recall also was lower for the NO.

C Chi-squared Automatic Interaction Detection Decision Tree

The overall accuracy of the results were 71.72% +/- 45.03%; Table (3), but with the prediction of no fraud the system unfortunately was failed. So this way failed for the dataset.

TABLE 3
System Accuracy Results Using CHAID Decision Tree.

<table>
<thead>
<tr>
<th></th>
<th>True YES</th>
<th>true NO</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. YES</td>
<td>104</td>
<td>41</td>
<td>71.72%</td>
</tr>
<tr>
<td>Pred. NO</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Class recall</td>
<td>100.00%</td>
<td>0.00%</td>
<td></td>
</tr>
</tbody>
</table>

The YES prediction Precision was low, and the NO was failed which takes this way of classification to scratch.

D Naïve Bayes Classification Results

The overall accuracy of the results were 68.00% +/- 10.04%; Table (4), which is lower than the C4.5 decision tree.

TABLE 4
System Accuracy Results Using Naïve Bayes Classification.

<table>
<thead>
<tr>
<th></th>
<th>True YES</th>
<th>true NO</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. YES</td>
<td>38</td>
<td>12</td>
<td>76.00%</td>
</tr>
<tr>
<td>Pred. NO</td>
<td>15</td>
<td>21</td>
<td>58.33%</td>
</tr>
<tr>
<td>Class recall</td>
<td>71.70%</td>
<td>63.64%</td>
<td></td>
</tr>
</tbody>
</table>
Technical and non-technical losses in power system can be both handled; the first one with the proper design of the power system.

While the second which is related to the energy consumed and not billed is a real problem to the electricity utilities and the national income also.

Several studies was performed on fraud, all of these studies used the load profile of the subscribers. The fact that most of theft locally was with peoples that have no meters and of course the safety is nonexistent; the load profile is not available. The problem of generating a database contains features that maximizes the accuracy rate of the classifiers will need much more work; but still the proposed methods was validated on a set of labelled data of the general information about subscribers that was set by the help of the inspection and monitoring department of JDECO company as they helped in this study.

In future, using different techniques for intelligent classification and optimization it is intended to use the load profile also side by side with more information about the subscribers; also the research I performed on the reasons that drive people to theft needs more work; despite the problem is a society problem which is not related to my study in computing; the output results of my classifier depends on the input information to the system. As I believe the more I knew about the person, the better I can expect its behavior. I will take the summer semester to search on the reasons that drive people to theft energy and of course, I will construct bigger database with the modified information from the general problem research (Fraud reasons) and use to improve the performance of the system in seminar 2.

REFERENCES

[1] Palestine Energy and National Recourses Authority-Ramallah\ AL Yasmeen Building \ Almsayef.
[2] Gaza Electricity Distribution Corporation (GEDCo)
[10] Bhavani Thuraisingham “Data Mining for Counter-Terrorism”.