Decision Tree Learning for Fraud Detection in Consumer Energy Consumption

Christa Cody*
Computer Science Department
North Carolina State University
Raleigh, NC, USA
cncody@ncsu.edu

Vitaly Ford[1], Ambareen Siraj[2]
Computer Science Department
Cookeville, TN, USA
Tennessee Tech University
[1] vford42@students.tntech.edu, [2] asiraj@tntech.edu

***Abstract*—The electrical grid is transitioning to new smart grid technology. With smart meters becoming an essential feature in smart homes, concerns regarding smart meters and the vast amount of consumer data that it captures are on the rise. While access to this fine-grained energy consumption data captured by smart meters can potentially violate consumer privacy, advanced analysis of this data can help to protect the interest of both the consumer and the utility company by enabling fraud detection at either end. The use of machine learning techniques has been a very common approach to energy fraud detection. Patterns in energy consumption can be recognized and used to detect anomalous behavior. This research reports on a novel application of decision tree learning technique to profile normal energy consumption behavior allowing for the detection of potentially fraudulent activity.**

Keywords-fraud detection; decision trees; smart meter data

I. INTRODUCTION

The outdated electrical grid is undergoing a slow transition to the new Smart Grid technology. The Smart Grid allows the electrical grid to be utilized in numerous ways to become more efficient, reliable, and beneficial to the consumer as well as the service providers. The stream of data generated by the smart meter can provide a log of fine-grained energy consumption that allows the consumers to monitor their energy usage for numerous reasons including financial and environmental. Although the accessibility of the data creates many opportunities for improvements, the same accessibility creates an avenue for potential privacy and security violations [1]. However, security violations related to fraudulent activities can be reduced by intelligent analysis of the fine grained data. By learning the characteristic patterns within the fine-grained data, normal energy usage can be predicted and as such, any anomalies can be reported to be potential energy fraud.

In this paper, we discuss the use of energy fraud detection to address some of these security violations. There are two types of energy fraud to consider [2]:

- Fraud type 1) the consumer's smart meter reports less energy consumption than actually consumed.
- Fraud type 2) the consumer's smart meter reports more energy consumption that actually used due to rogue connections.

These types of fraud can be created by numerous mechanisms such as:

- Unauthorized tapping to electricity line;
- Bypassing the smart meter to report customized energy consumption; and
- Tampering with the smart meter by implanting chips to slow down its readings.

Using machine learning techniques, a consumer's normal energy consumption behavior can be modeled. This modeled behavior can then be used to closely monitor the consumer's activity to detect significant anomalies, which indicate fraudulent activity. Our approach includes using a decision tree machine learning algorithm, M5P, to model consumer's normal energy consumption behavior for the prediction of future energy consumption and the detection of fraudulent activity. We demonstrate the effectiveness of our approach by conducting experiments using anonymized smart meter data from the Irish Social Science Data Archive Center [3].

The structure of the paper is as follows: Section II discusses the related fraud detection work using machine learning techniques and their results. Section III describes the fraud detection approach using pre-processing methods and decision learning. Then, section IV discusses the fraud detection results. Lastly, section V concludes with future work.

II. RELATED WORK

Energy fraud is a very common problem and currently lacks the accuracy of detection that is needed to limit energy fraud to a minimum. Researchers have proposed and applied several machine learning techniques to address this. This section discusses a few of these approaches.

Depuru et al. [4] applied Support Vector Machines (SVMs) to classify fraudulent activity. The proposed fraud detection system includes pre-processing and data classification. Pre-processing of the energy consumption data are used to create noise free data to train and test the SVM. The noise free data are achieved by using a set of criterion, which includes geographical locations, season of the year, and category of customers (agricultural, commercial, and residential). Then, SVM is used to classify the customer's energy consumption data to allow for prosecution or further monitoring. The proposed method

classified data with 98.4 percent accuracy for 220 customers.

Monedero et al. [5] developed a prototype for the detection of fraudulent activities using two data mining approaches: neural networks and statistical techniques. The experiments were conducted with data from Seville, Spain. The results of the data mining processes were given to the electrical companies to carry out inspections, who previously carried out inspections without any filtering. Out of the 13 cases detected as possible fraudulent activity, 50% of the cases were determined to be fraudulent leading to a 50% success rate.

Nagi et al. [6] utilized SVMs for detecting energy fraud in the power grid. A combinations of steps such as data pre-processing, feature extraction, classification, and data post-processing was used for detection of suspicious energy consumption patterns. The experiments yielded a hit rate of 64% using SVMs along with a structured query language (SQL) based decision making system that utilized human expertise in analyzing energy fraud cases.

Costa et al. [7] proposed a classification technique implementing artificial neural networks for energy fraud detection. The method of detecting included data cleaning, data selection/transformation, and data mining. This research resulted in an improvement of 50% over the Brazilian electric power company's previous methods.

It is apparent that existing research in fraud detection increases the detection of fraud and provides many improvements to older detection systems. However, the detection rates are still far from satisfactory leading to high financial loss experienced by both electrical companies and their customers. Furthermore, most of this existing research uses outside knowledge such as fraud detection expertise or highly detailed knowledge about the customer. This outside knowledge might not always be as easily attainable, which would present problems in implementing these methods to real world applications where energy consumption may not be consistent and detailed information about the consumers' may not be readily available.

III. Fraud Detection Approach

It is important to discuss data preprocessing along with the format and the type of data used in our experiments since it impacts the fraud detection method directly. We examined smart meter energy consumption data from a collection of approximately 5,000 residential and 650 business smart meters [3]. The energy consumption data values were recorded in 30 minute intervals during a period spanning roughly two years (2009-2011). The raw data are stored in six different files, each consisting of around 24 million entries. Table I represents a small sample of the data contained within these files and is represented in three columns. The first column represents the smart meter ID which is linked to a particular resident or business. The second column shows the date and time associated with the meter reading and the third column is the energy consumption measurement in kilowatt-hours (kWh).

TABLE I. Raw Data File Structure

Meter ID	Encoded date/time	Energy consumption value kWh
1049	44431	.064
1049	44432	.061
…	…	…
1232	46126	1.092

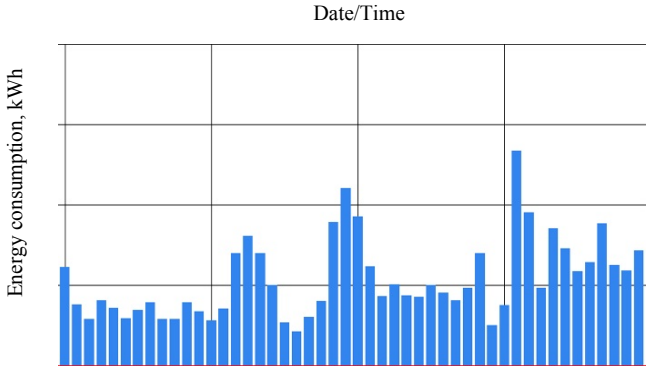

Fig 1. Bar graph representation of energy consumption data in 30 minute intervals.

Fig. 1 shows the energy consumption of a particular consumer over a period of one day.

A. *Pre-processing*

Due to the data being fine-grained, it is possible for some of the data to be inaccurate or missing. It is widely accepted that effective processing of data not only depends on the quality of the selected algorithms, but also the quality of the data. Therefore, a variety of pre-processing techniques were used to ensure the quality of the data. In this section, we discuss the two main pre-processing techniques utilized on the raw energy consumption data before the data are analyzed: *feature selection* and *data aggregation*. For faster access to the data in the files, an information retrieval program [2] was used that compresses and indexes the raw energy consumption data.

1) Feature Selection: The experimental smart meter data contain several classifying parameters such as month, day, week, time, and consumer energy consumption values (kWh). Furthermore, the energy consumption behavior is affected by a multitude of other factors such as holidays, seasons, and special occasions. In order to learn an accurate model, the features selected must be limited to only the relevant ones. Therefore, due to the complexity and possible irrelevancy of other features, we decided to focus on the day, week, time, and energy consumption features excluding the month feature present in the original energy consumption data. The month feature did not add to any significant improvement in prediction results and thus for the sake of simplicity was not taken into account.

2) Data Aggregation: The raw energy consumption data are reported in 30 minute intervals. However, humans typically do not operate on strict daily schedules consistent over such small intervals. Therefore, we aggregated the data

into larger intervals to allow for flexibility in schedules. The data were aggregated into *k* hour intervals, where *k* is the number of hours compressed into one energy consumption value. The new intervals helped to smooth out slight variations in daily activities that could be falsely detected as fraud.

3) Indexing and compressing: The data are contained within six files with millions of smart meter ID entries per file. Using an existing information retrieval program created for this dataset [2], the process of accessing the data was greatly improved. This program creates an index for each smart meter ID to its own measurement data to provide quicker data retrieval. Each index is compressed by delta encoding to store or transmit data in the form of differences between the sequential information rather than the individual files.

B. Decision Tree Learning for Fraud Detection

Machine learning techniques are applied in many areas of research for deriving computational intelligence. It allows for the generalization of specific examples that can be used in modelling, predication, and classification of datasets. A decision tree is one such widely used machine learning technique that has been effective for classification or regression. Its usefulness results from the ability to compensate for missing values and having a highly flexible hypothesis space [8].

Decision trees are generated by algorithms that split a dataset into multiple branching segments based on decision rules. These decision rules are determined by identifying a relationship between input attributes and the outputs. In this research, M5P decision tree learning algorithm has been used, which is a regression algorithm reconstructed from Quinlan's M5 algorithm [9]. The M5P algorithm uses a combination of a decision tree and linear regression functions at the leaves. Regression algorithms are used to predict future values based on previously learned data. The key goal of applying the M5P decision tree algorithm in this problem space is to learn individual behavior per costumer to create an energy consumption model. Afterwards, the learned energy consumption model is used to predict future measurements. With the ability to predict future measurements, it is possible to compare the real values with the predicted values and apply statistical measures to detect potential fraudulent activities.

The following is step-by-step description of the energy fraud detection method using Decision Tree.

1) Generation of Training and Validation Datasets: Training sets are generated by selecting a set of measurements from a single smart meter over a certain period of time. For instance, a training set may be generated using the energy consumption measurements collected within a particular time period (for example, October 1st to 30th, 2009) for a particular customer (for example, customer with Meter ID 1000). This selection process is repeated for the generation of the validation dataset. The validation set is generated by selecting another period of time (month/year/season) using the same set of smart meter measurements.

2) Training the Decision Tree: After the selection of the training dataset, this dataset is used generate the decision rules representative of the normal energy consumption behavior model for the customer in question.

3) Prediction: Prediction is achieved by using the generated decision rules to predict the expected energy consumption values based on the feature set (year, day of the week, time) of the validation set.

4) Anomaly Detection: The *Root Mean Squared Error* (RMSE [15]) is a widely used statistical method to measure the differences between predicted values and actual values. This calculation is used as an indicator of deviation between the predicted value as learned from the training dataset and the actual value in the validation dataset and is calculated as in (1):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \left(y_t' - y_t\right)^2}{n}}, \qquad (1)$$

where y_t' is a predicted energy value and y_t is an actual energy consumption value in the validation set, and n is the total number of instances. A threshold for the RMSE value serves as an indicator for fraudulent activity. Any value above the threshold, serves as an indication of possible energy fraud.

IV. Results and Discussion

This section discusses the experimental setup, the datasets, the prediction results, and the use of the prediction for fraud detection.

A. Experimental setup

A variety of experiments were conducted with training and validation sets spanning over varying lengths of time. These experiments were conducted using WEKA [10], a data mining tool that provides access to many machine learning algorithms.

The first step in the experiment is training of the decision tree using the selected training dataset. Since overfitting is a common problem in decision tree learning, the training dataset size must be carefully chosen to ensure that the model does not overfit or underfit the data [11]. To accurately predict future values, the model needs to have an adequate amount of instances to appropriately represent the energy consumption behavior. After trial and error analysis, the training set size was selected to perform optimally when the dataset spanned approximately three weeks. The second step in the experiment included using the model generated from the training dataset to predict expected energy consumption values corresponding to the validation dataset time period. Fraudulent activities were simulated to illustrate real world scenarios in energy fraud [2]. For example, a chip can be inserted into the smart meter to produce lag in smart meter processing, which will result in reporting of lower than average energy consumption by the smart meter (fraud type 1 as described in section 1). Another

case (typical in under developed countries) could be when a customer's energy is being used by an outsider via unauthorized rogue lines, which will result in reporting of higher than average energy consumption by the smart meter (fraud type 2 as described in section 1). To simulate both types of fraud, a random value deviating from 0 to 0.5 kWh was subtracted (imitating fraud type 1) and added (imitating fraud type 2) to every measurement of the original energy consumption data in the validation dataset.

The original data (noise-free) were used to learn the consumption model and simulated fraudulent data (with the added noise) were used in validating the model. The RMSE calculation is used to distinguish between the normal consumption behavior and possible fraudulent activity. If the RMSE value is below the threshold, the dataset is considered normal. However, if the RMSE is above the threshold (which in the following experiments is 0.4 kWh), then the dataset can be flagged as possible fraud.

B. Experimental Results

The following experiments were conducted to evaluate effectiveness of using decision tree learning to learn the consumer model and application of the model with statistical means to predict energy fraud detection.

Experiment 1: The first experiment was designed to evaluate the extent to which the learned model can predict energy consumption values for the same month a year after the training set. In experiment 1, the decision tree was trained using energy measurements from August of 2009 and validated using energy measurements with and without simulated fraudulent activities from August of 2010. The results of experiment 1 are shown in Fig. 2. This figure shows the resulting RMSE values using a validation set consisting of both normal and fraudulent activities (both fraud type 1 and 2).

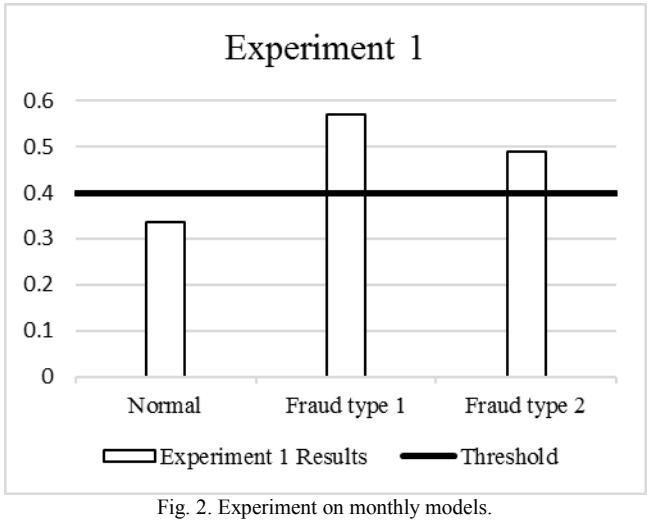

Fig. 2. Experiment on monthly models.

The deviations using the fraudulent validation datasets resulted in values above the threshold (the bold, straight line), which demonstrates that our model was able to identify possible fraudulent behavior, when present. The same experiment was repeated for other months (October and December) and yielded similar results.

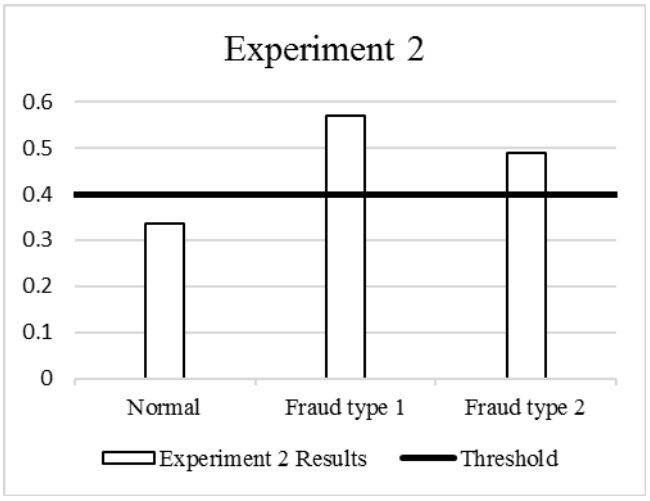

Fig. 3. Experiment on weekly models.

Experiment 2: The second experiment was designed to evaluate the extent to which the learned model can predict energy consumption values for the week following several weeks in the training set. In experiment 2, the decision tree was trained using energy measurements from September 6-26, 2009 and validated using energy measurements with and without fraudulent activities from September 27 – October 2010. The results of experiment 2 are shown in Fig. 3. In this case also both sets fraudulent data were reported to be above the threshold indicating fraudulent activity. The same experiment was repeated for validations sets spanning consecutive weeks and yielded similar results.

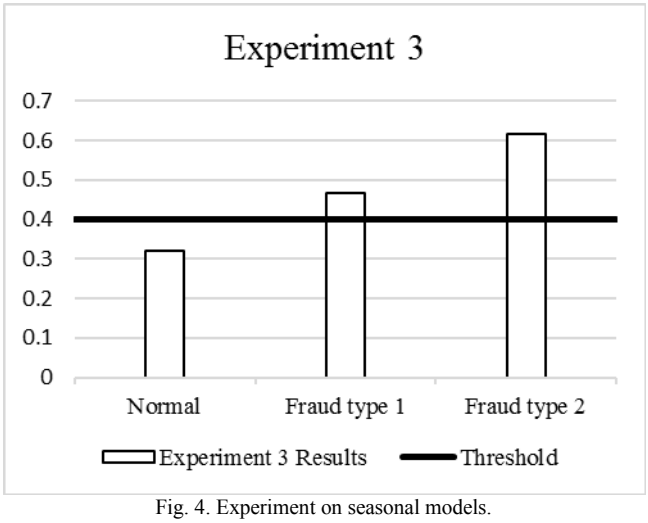

Fig. 4. Experiment on seasonal models.

Experiment 3: The third experiment examines the ability to predict energy consumption values within the same weather season. The training and validation sets consisted of data collected over a time period in the same weather season. Experiment 3 shows the results of training with measurements from June 2010 and validating with a dataset from July 2010. The results of experiment 3 are shown in

Fig. 4. This type of experiment was repeated using testing and validation sets within the same three month weather season period such as June 2010 (for training) and August 2010 (for validation). The results of experiment 3 indicate that both fraudulent datasets were above the threshold, indicating possible energy fraud.

V. CONCLUSION

We demonstrated successful application of decision tree learning for detecting fraudulent activity in fine-grained energy consumption data. The results of the conducted experiments reveal the ability to accurately predict energy consumption values from the same month of a year, subsequent weeks, and within the same weather season. Real historical data were used in these experiments to generate the decision tree model and predict future energy consumption. After statistical analysis, the future energy consumption values were categorized as normal or fraudulent. These results demonstrated how the M5P decision tree learning algorithm can be applied to groups of data under investigation to accurately predict future values.

Future work can be done to improve prediction and detection using this model. To more accurately predict future energy consumption values, pre-processing methods such as feature extraction algorithms can be utilized to provide the model with the most relevant set of features by removing all redundant or irrelevant features. This will allow the model to provide a more comprehensive coverage of the different factors that influence a consumer's energy consumption behavior without making the model overly complex. Furthermore, a comparison to other machine learning approaches can be conducted to identify the positive and negative aspects of each. Such comparison could allow identification of combined application of complementing machine learning techniques to provide a more robust fraud detection system.

ACKNOWLEDGMENT

This work was supported in part by NSF Award #1303441. We are thankful to the Irish Social Science Data Archive Center for providing us with access to the smart meter data archive that was crucial to this research.

REFERENCES

[1] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy Theft in the Advanced Metering Infrastructure", in Critical Information Infrastructures Security, pp. Pp 176-187, 2010.

[2] V. Ford, A. Siraj, W. Eberle, "Smart grid energy fraud detection using artificial neural networks," IEEE Symposium on Computational Intelligence Applications in Smart Grid, 2014, pp.1,6, 9-12 Dec. 2014.

[3] "Commission for Energy Regulation, Irish Social Science Data Archive, ucd.ie. [Online]. [Accessed: March 5, 2013]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulation/"

[4] S.S.S.R Depuru, L. Wang, V. Devabhaktuni., "Support vector machine based data classification for detection of electricity theft," Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES, pp.1,8, 20-23 March 2011.

[5] I. Monedero, F. Biscarri, and C. León, "MIDAS: Detection of Non-Technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques", in Proceedings of the Computational Science and Its Applications Conference – ICCSA, Vol. 3984, pp. 725-734, May 8-11, 2006.

[6] J. Nagi, A. M. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines", in Proceedings of IEEE Transactions on Power Delivery, Vol. 25, No. 2, pp. 1162-1171, April 2010.

[7] B. C. Costa et al., "Fraud Detection in Electric Power Distribution Networks Using an ANN-Based Knowledge-Discovery Process", International Journal of Artificial Intelligence & Applications, Vol. 4, No. 6, pp. 17-23, November 2013.

[8] T. M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[9] J. Quinlan. "Learning with continuous classes." 5th Australian joint conference on artificial intelligence. Vol. 92. 1992.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[11] W.M.P. van der Aalst, V. Rubin, H.M.W. Verbeek, B.F. van Dongen, E. Kindler, and C.W. Günther. "Process mining: a two-step approach to balance between underfitting and overfitting." *Software & Systems Modeling* 9.1 (2010): 87-111.

---

[*] This work was conducted when the author was an undergraduate student at Tennessee Tech University.