Clustering of Smart Meter Data for Disaggregation

Vitaly Ford
Computer Science Department
Tennessee Technological University
Cookeville, TN
vford42@students.tntech.edu

Dr. Ambareen Siraj
Computer Science Department
Tennessee Technological University
Cookeville, TN
asiraj@tntech.edu

Abstract — **This research addresses privacy concerns in smart meter data. Smart meter data is analyzed for learning normal consumer usage of electricity. Clustering technique such as Fuzzy C-Means is used to disaggregate and learn energy consumption patterns in smart meter data. Results of experimentation with real world meter data demonstrate that it is realistically possible to profile the electricity consumption behavior of consumers analyzing their usage captured by smart meters.**

Keywords—Security, Smart Meter, Disaggregation, Fuzzy C-Means Clustering

I. Introduction

Smart grid is the aggregation of digital technologies, electric grid, bulk generation, transmission and distribution lines, service providers, “smart” devices, customers, and markets [3]. Smart devices have several advantageous properties like adaptability according to the customer’s preferences, scalability and collaborative synchronization allowing customers to adjust daytime/nighttime energy consumption in accordance with their inclinations. A smart meter is a control component of advanced and reliable smart grid technologies. Generally, smart meters are energy monitoring devices allowing consumers to have access to their energy consumption 24/7 by means of wireless technologies and two-way communications between the main system and the meters [3]. It is essential to keep smart meter traffic secure because, by compromising smart meters, attackers can replace real data with fraudulent ones or, for example, disconnect customers from electricity supply. This research addresses the privacy concerns in smart meter data. We analyze smart meter data for learning normal consumer usage of electricity. The acquired knowledge is then used to disaggregate energy consumption using clustering technique to identify specific usage behavior. We believe addressing disaggregation techniques can aid in smart meter data forgery problem.

This paper is structured as follows. In section 2: Background describes different attacks and defenses associated with smart meter data, introduces the clustering technique used and describes the experimental dataset. In section 3, we describe the disaggregation approach. Section 4 describes the experiments and results. In section 5, we summarize this research and conclude with future direction.

II. Background

A. Attacks in Smart Meters

There have been several types of notable attacks (both theoretical and demonstrated) aimed at compromising smart meters. Some of these are:

1) <u>Denial of service</u> attacks that compromise smart meters such that they are not capable of responding to any request sent by a customer or energy supplier. It is accomplished through smart grid network exhaustion or tempering with the routing of the smart meter traffic [11].

2) <u>False-data injection</u> attacks that introduce arbitrary and/or certain errors inside a normal smart meter traffic activity causing invalid measurements that are unacceptable in a smart grid network [13]. In addition, the researchers in [13] show that these types of attacks can be implemented even if an attacker is limited to accessing smart meters or has restricted resources for fulfilling the attack.

3) <u>De-pseudonymization</u> attacks that compormise anonymization and privacy of smart meter data [12].

4) <u>Man-in-the-middle</u> attacks where rogue agents can place themselves in between consumer and energy company by compromising Wi-Fi technologies used all over a smart grid network [14].

5) <u>Meter spoof and energy fraud</u> attack can occur by gaining the smart meter ID through physical access [11].

6) <u>Authentication</u> attacks where attackers can authenticate as valid user are possible with physical access to smart meter where user authentication password can be obtained via a direct connection to the EEPROM storage. Also, that fact that utility companies typically use the same password for many smart meters makes this attack more costly [11].

7) <u>Disaggregation</u> attacks for profiling customer energy consumption behavior [8].

B. Clustering with Fuzzy C-Means

Researchers have applied several machine learning techniques to detect attacks in a smart grid network [15, 16, 17]. Clustering is one of the most common methods of unsupervised learning for examining and grouping data with respect to their particular characteristics. In unsupervised learning there are no predefined groups or examples to guide the clustering process. Clusters are generated as groups with unique property. Objects/instances inside a cluster are more similar to each other than to any other objects/instances from

978-1-4799-0248-4/13/$31.00 ©2013 IEEE

the rest of the clusters found in the data set [1]. There are many various kinds of clustering algorithms that have been applied in diverse areas of computer science.

Fuzzy clustering takes into account uncertainty in real data and allows objects to be part of more than one cluster. In our research, we have employed *fuzzy c-means clustering* technique because of its demonstrated ability to identify clusters in a flexible manner. One of the most unique features of the fuzzy c-means clustering is that all of the instances in the data set belong to each cluster to a certain degree/extent [2]. Fuzzy c-means clustering is typically used for classifying complex data sets like time series data streams [1].

Fuzzy c-means (FCM) clustering split data into clusters in such a manner that each instance belongs to a cluster to an extent defined by a similarity measure that is usually represented by the Euclidean distance. FCM is aimed at minimizing the following function [2, 5]:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2, \quad 1 \le m < \infty ,$$

where u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i-th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by [2, 5]:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{1}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

This iteration will stop if $\max_{ij}\left\{\left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right|\right\} < \varepsilon,$ where ε is a termination criterion between 0 and 1, and k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The algorithm can be represented as the following [2, 5]:

1) *Initialize U=[u$_{ij}$] matrix, U$^{(0)}$*
2) *At k-step: calculate the centers vectors C$^{(k)}$=[c$_j$] with U$^{(k)}$*
3) $c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$
4) *Update U(k) , U(k+1)*
5) $u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{1}{m-1}}}$
6) *If* $\left\| U^{(k+1)} - U^{(k)} \right\| < \varepsilon$, *then STOP; otherwise return to step 2*

C. Cluster Validation with Xie and Beni validation index

Cluster validation techniues play an important role in assessing quality of generated clusters. The main objective of cluster validation is to evaluate clustering results to find the partitioning that best fits the underlying data. Among the validity indices suitable for fuzzy clustering, a notable one is *Xie and Beni validation index* which takes into account membership values of instances in clusters measued by geometric distance measure and distance beetween cluster centroids and the dataset itself [6]. This index demonstrates good performance in comparison with some other validation indices directed to idetifying optimal number of clusters [7].

The Xie and Beni validation index is aimed at determining the minimum of the following function [6], where n – is the total number of instances:

$$V_{XB}(U;V;X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \cdot \left\| x_k - c_i \right\|^2}{n \cdot \left(\min\{c_i - c_j\}\right)}.$$

D. Smart Meter Data

Unfortunately there is little availability of freely accessible smart meter data including energy consumption traces and other pertinent information representing smart meters' behavior inside a network – making research in smart meter security very difficult. As an alternative, researchers have been making use of various kinds of simulators like GridLAB-D, GridSim, or IEEE 300-bus. However, these simulators do not support implementing attacks inside the virtual environment for testing countermeasures [18, 19]. Fortunately, we were able to access and work with real smart meter data – courtesy of the Irish Social Science Data Archive Center [4]. The representative samples contain smart meter data from 5,000 residential consumers and 650 businesses collected over almost two years. The energy consumption information is collected by smart meters every thirty minutes. This massive data is stored in six files with over 24 million lines each. Every line in the files contains three values: (1) smart meter ID, (2) encoded date/time, and (3) value of the consumed energy in kW/h.

III. Approach

For disaggregation of the smart meter data, we conduct clustering with Fuzzy c-means and validate the clusters with Xie and Beni validation index. However to make this process efficient by preprocessing the massive dataset, application of indexing and compressing techniques were necessary.

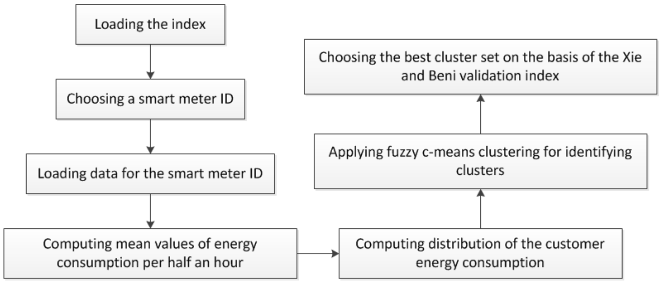
A. Indexing and compressing algorithms

For providing fast access to individual smart meter data, we have employed indexing and compressing algorithms used in the information retrieval systems. As index, combination of meter IDs and line numbers locating the corresponding meter IDs were used. For compressing the index, we have utilized gamma- and delta- codes widely used in information retrieval systems [20].

B. *Block scheme of our approach*

The procedure starts with loading the existing index of meter IDs into the application. Then, after selecting a particular meter ID, the time series is loaded and analyzed for identifying the mean values of energy consumption per half an hour, distribution of the values over time, and clusters. The minimum number of clusters is set to 4 because, after reviewing the whole data of the customer under consideration, we have noticed that his/her energy consumption values have at least 4 distinct levels. The algorithm splits data into clusters, calculates the Xie and Beni validation index for the current set of clusters, and then decides whether it should increment the current number of cluster and continue dividing the data into clusters or not. Figure 1 illustrates the procedure of analyzing smart meter data as a whole.

Figure 1: Process Map.


Loading the index
Choosing a smart meter ID
Loading data for the smart meter ID
Computing mean values of energy consumption per half an hour
Computing distribution of the customer energy consumption
Applying fuzzy c-means clustering for identifying clusters
Choosing the best cluster set on the basis of the Xie and Beni validation index

IV. Experiments and Results

Several experiments were conducted, implementing both statistical and clustering analysis. We have streamlined an application for conducting the experiments and detecting observable patterns in smart meter data.

A. *Statistical analysis*

When a three-day data sequence for one of the meters (Figure 2) is observed, we can notice that there is a certain pattern of energy consumption behavior. Here axis *X* denotes date/time of the measurement, and axis *Y* denotes energy consumption value in kW/h. From these observations, it can be inferred that the consumer is a service providing business (like a store/eatery) rather than a household as the energy consumption is at its peak consistently between 8:30 A.M. until 10:00 P.M. (Figure 3).

Figure 2: Energy Consumption Profile for One Smart Meter for Three Consecutive days.

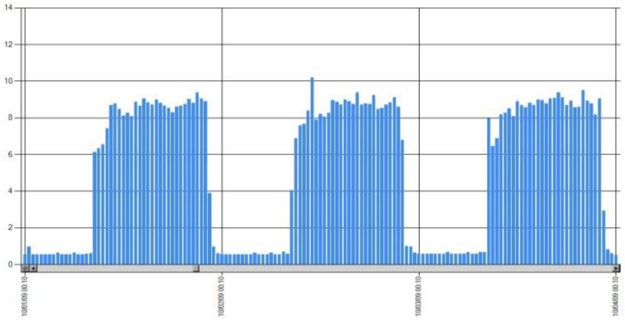
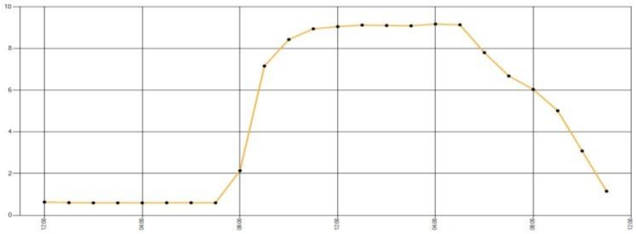


Figure 3: Mean Energy Consumption per Half an Hour.

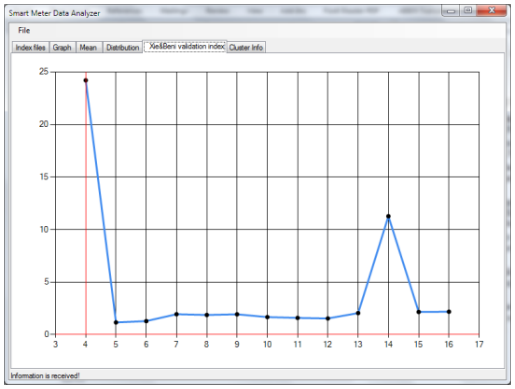


Also, the business company is using certain types of appliances consuming 0.55 kW/h during a nighttime period every half an hour. It is likely that these appliances would be security and/or fire detecting devices with periodic small and persistent energy consumption.

B. *Clustering analysis*

While experimenting with FCM clustering algorithm, for a similarity measure, time and the value of the energy consumed is used. Figure 4 depicts the Xie and Beni validation index metric (*Y* axis) depending on the number of clusters (*X* axis). This graph represents energy consumption for a single customer randomly chosen from the dataset.

Figure 4: Xie and Beni Validation Index for a Single Customer.

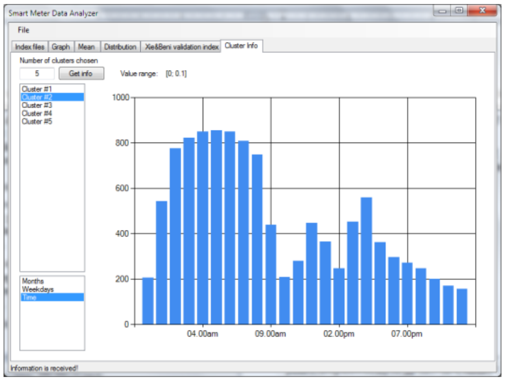


Information is received!

Analyzing Figure 4, it can be inferred that the optimal number of clusters for that particular customer equals 5 because this measure has the minimum value of the Xie and Beni validation index. Figure 5 represents the second cluster found for that particular customer, whereas Figure 6 shows the forth cluster (the value of the consumed energy varies from 0.358 to 0.548 kW/h) for the same. We are unable to share results of all 5 clusters found here due to space constraints.

Figure 5 shows that the value of the consumed energy varies from 0 to 0.1 kW/h between 1 A.M. to 8 A.M. Therefore, it can be inferred that during that time the customer does not usually use any appliances maybe because, if it is a residential home, customer sleeps at that time or, if it is a business, nobody is active at that period of time.

Figure 5: The Second Cluster.



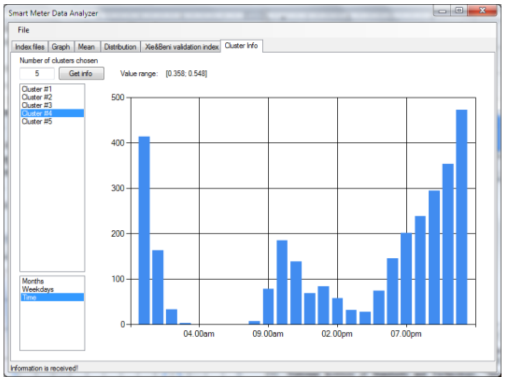
Number of clusters chosen: 5

Value range: [0; 0.1]

Information is received!

Taking further observation into account, like the fact that the customer usually consumes from 0.358 to 0.548 kW/h during 8 P.M. – 12 A.M. (Figure 6), it can be deduced that we are looking at a typical working household (where people sleep at night, go to work all day and come back to have dinner, watch TV and then go to bed again), and not a business.

Figure 6: The Forth Cluster.



Number of clusters chosen: 5

Value range: [0.358; 0.548]

Information is received!

V. Conclusion

In this research, disaggregation analysis on smart meter data was conducted by applying fuzzy c-means clustering. We have demonstrated that with access to customer energy consumption data, one can apply disaggregation techniques for inferring consumers' energy usage profile. Knowledge gained such way can be abused in undesirable ways, for example, time frame between when the costumer leaves and returns home offers opportunities for home invasion, marketing by phone, or even children behavior profiling [8, 9, 10]. On the other side, utility companies can use such knowledge to detect abrupt changes in consumer profile, suspecting energy fraud.

Currently, we are investigating solutions to address privacy issues of smart meter data communication, secure data communication, particularly storage protection of cryptographic keys. We hope to contrive a suite of different protection mechanisms for effective smart meter security. In addition, we have been working on applying machine learning techniques for learning normal behavior of energy consumption and identifying abnormal activities, heading towards the energy fraud detection.

Acknowledgements

This work is supported by the Center for Energy Systems Research of TTU. We are thankful to the Irish Social Science Data Archive Center for allowing access to the Smart Meter Data that was instrumental to this research. We would like to thank Dr. Martha Kosa and Dr. Doug Talbert for their insights into indexing, compressing and clustering algorithms.

References

[1] T. W. Liao, "Clustering of time series data – a survey", *Journal of the pattern recognition society 38*, pp. 1857-1874, January 7, 2005.

[2] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York and London, 1987.

[3] National Institute of Standards and Technology, "Smart Grid Cyber Security Strategy and Requirements", *NISTIR 7628*, Vol. 1, Aug., 2010.

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

[5] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57.

[6] Xie, X. L, Beni, G.. "A Validity Measure for Fuzzy Clustering", *IEEE Trans. on Pattern Analysis and machine Intelligence*,Vol.13, No4, 1991.

[7] M. R. Rezaee, B.P.F. Lelieveldt, and J.H.C. Reiber, "A new cluster validity index for the fuzzyc-mean", *Pattern Recognition Letters 19*, pp. 237-246, October 16, 1998.

[8] Jack Kelly, "Smart Meter Disaggregation", *jack-kelly.com*. [Online]. Available: http://jack-kelly.com/smart_meter_disaggregation [Accessed: March 8, 2013].

[9] B. J. Murill et al., "Smart Meter Data: Privacy and Cybersecurity", in Congressional Research Service Report for Congress, February 2, 2012.

[10] E. McKenna et al., "Smart meter data: Balancing consumer privacy concerns with legitimate applications", *Energy Policy journal 41*, pp. 807-814, 2012.

[11] S. McLaughlin et al., "Multi-vendor Penetration Testing in the Advanced Metering Infrastructure", in *Proceedings of the 26th Annual Computer Security Applications Conference*, New York, NY, 2010, pp.107-116.

[12] M. Jawurek et al., "Smart Metering De-Pseudonymization", in *Proceedings of the 27th Annual Computer Security Applications Conference*, New York, NY, 2011, pp.227-236.

[13] Y. Liu et al., "False data injection attacks against state estimation in electric power grid", in *Proceedings of the 16th ACM conference on Computer and Commun. Security*, New York, NY, 2011, pp.21-32.

[14] U.S. Department of Energy, Office of Electricity Delivery and Energy Reliability, "Study of Security Attributes of Smart Grid Systems – Current Cyber Security Issues", INL/EXT-09-15500, Apr. 2009.

[15] R. Berthier, W. H. Sanders, "Specification-based intrusion detection for advances metering infrastructure", in *Proceedings of the 2011 IEEE 17th Pacific Rim International Symposium on Dependable Computing*, Washington, DC, 2011, pp.184-193.

[16] M. A. Faisal et al., "Securing AMI using intrusion detection system with data stream mining", in *PAISI'12 Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics*, Berlin, Heidelberg, 2012, pp.96-111.

[17] R. Berthier et al., "Intrusion Detection for Advanced Metering Infrastructures: Requirements and Architectural Directions," in *Smart Grid Communications, 2010 First IEEE International Conference*, pp.350-355, Oct. 4-6, 2010, doi: 10.1109/SMARTGRID.2010.5622068.

[18] D. M. Nicol, "Testing the Edge: Cyber Security Testing in the Smart Grid", Information trust institute.

[19] R. Goodfellow et al., "First steps toward scientific cyber-security experimentation in wide-area cyber-physical systems", in *Cyber Security and Information Intelligence Research Workshop '12,* Oak Ridge, TN, Oct. 30 – Nov. 01, 2012. ACM 978-1-4503-1687-3/12/10.

[20] C. D. Manning, P. Raghavan, and H. Schütze, *An introduction to information retrieval*. 2009.