RESIDENTIAL LOAD PROFILE ANALYSIS USING CLUSTERING STABILITY

Fang-Yi Chang[1], Shu-Wei Lin[1], Chia-Wei Tsai[2] and Po-Chun Kuo[3]

[1]Digital Transformation Institute,
Institute for Information Industry, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering
Southern Taiwan University of Science and Technology, Tainan, Taiwan
[3]Department of Information Management
Southern Taiwan University of Science and Technology, Tainan, Taiwan

ABSTRACT

Clustering is an useful tool in the data analysis to discover the natural structure in the data. The technique separates given smart meter data set into several representative clusters for the convenience of energy management. Each cluster may has its own attributes, such as energy usage time and magnitude. These attributes can help the electrical operators to manage their electrical grids with goals of energy and cost reduction. In this paper, we use principle component analysis and K-means as dimensional reduction and the reference clustering algorithm, respectively, and several choices must be considered: the number of cluster, the number of the leading principle components, and whether use normalized principle analysis schema or not. To answer these issues simultaneously, we use the stability scores as measured by dot similarity and confusion matrix as our evaluation decision. The advantage is that it is useful for comparing the performance under different decisions, and thus provides us to make these choices simultaneously.

KEYWORDS

Smart meter; Unsupervised; Nonparameter; Clustering; PCA; Stability; Smart Grid; Value-Add Electricity Services; Energy Saving; Energy management

1. INTRODCUTION

The research of smart meter data has been stimulated by the need for electrical grid operators for energy management as the era of smart grid coming. Smart meter can send the fine grained energy consumption data back near real-time to the electrical operators or the electrical retailers. The amount of smart meter is massive and is accumulated at very faster speed, thus how to utilize and manage the smart meter efficiently has become an important topic worldwide. In Taiwan, Taiwan Power Company, the largest electrical utility in Taiwan, has been aware of the need of clustering of smart meter to better understand the energy usage patterns of low-voltage customers. The company have found that the patterns are so complex, diverse and dynamic that artificial-based methods are inefficient to deal with them. Taiwan now has been undergoing the green

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 73–81, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81707

energy transition since few years ago, and hence Taiwan Power Company or the related bureaus need to know the end user's usage behavior, especially during peak hours, to complete this transition aimed with energy reduction.

Clustering is an useful tool to separate the smart meter data set into several representative groups to reflect the attributes of energy usage time and magnitude for the convenience of energy management. Here we briefly introduce the recent research works of clustering of smart meter. These works can be mainly divided into two approaches. One is to identify the relationship between energy usage patterns and socio-demographics [7][8]. Another is to identify suitable and representative groups using smart meter data only [5][6][9]. Both two approaches are to find energy reduction solutions and hence optimize the electrical network and reduce the energy. Since the amount of smart meter is massive, it is necessary to reduce the dimensionality of the raw data to provide a more robust and efficient clustering. The methods of dimensional reduction among the recent works are principle component analysis (PCA)[9] and artificial-based variable selection [5][7][8]. PCA is an classical technique of dimensional reduction and has achieved a significant success in many field, including bio-statistics, signal process and image process. It reduces dimensionality or selects variables by using the leading high variance principle components to perform data analysis and filtering the rest components which act as noisy signal. Artificial-based variable selection in these research works is to artificially find the representative attributes. For example, Stephen et al. [5] chosen the four time period and other seven attributes by specific mathematical formulas. On the other hand, there have been a variety of clustering techniques, including K-means, finite mixture model (FMM) and Hidden Markov Model (HMM), and evaluation decisions, which have been applied in smart meter in accordance with the experimental data and main purpose. For example, Stephen et al.[5] used FMM and Bootstrap to estimate the validation scores, which is the relative entropy. Adrian Albert and Ram Rajagopal [8] used HMM and BIC score to perform spectral clustering. Charalampos et al.[9] used K-means, Hierarchical Clustering and Hausdorff-based K-medoids, and evaluated these performances by Dunn Index, Calinski Harabasz Index and Energy Variance Index.

In this paper, we use PCA and K-means as our dimensional reduction technique and reference clustering algorithm, respectively. For the optimal decision, we use 'stability' to select to number of clustering and the corresponding clustering and claim that the it is advantageous in reality. The advantages are 'stability' that is naturally led by the evaluation decision, which is convenience for energy management in reality. It provides the 'trade-off' between the number of clusters and the corresponding stability scores, avoiding leading to providing the simplest clustering result only. For example, if the optimal number of cluster is 2, the corresponding clustering is too simple for electrical operators or retailers.

2. DIMENSIONAL REEDUCATION

A. Principle Component Analysis

Principle component analysis is a classical technique of dimensional reduction by orthogonal transformation into a new set of coordinates which are linearly uncorrelated [1]. These new coordinates are called principle components and are arranged such that the kth principle component has the kth largest variance among all principle components. The larger variance implies that the corresponding variable has more information and have more relevant to clustering. In the paper, we denote the PCA relative to the covariance matrix and the correlation

matrix as center PCA and normalized PCA, respectively, and compare the performances between two schemes.

3. CLUSTERING

A. Clustering Algorithm

K-means is attempt to minimize the objective function:

$$Q_K^{(n)}(c_1, ..., C_K) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,...,K} ||X_i - c_k||^2 \quad (1)$$

where K is the fixed number of clusters, X_1, … $X_n \in R$ are the data points and c_1, …. c_k denote the centers of the K clusters.

Several research papers have tried to model energy usage pattern in parameter probabilistic model, including Gaussian, beta, gamma and log-normal distribution [10][11][12]. However, since the distribution of our experimental data are complex and diverse, and no other information is available about our data, we do not assume any parameter probabilistic model to our experimental data in the paper, which motivates us use clustering stability as the evaluation decision. In fact we do not know whether the given data set can be represented as any mathematical models, and the K-means algorithm is used anyway. However, what we concern is that the whole data set can be represented using K-means to split each true cluster in several smaller and representative groups, rather than select the 'correct' number of clusters. It is acceptable even with the fact that the true clusters are split in smaller groups, or to afterwards join these groups to form a bigger group.

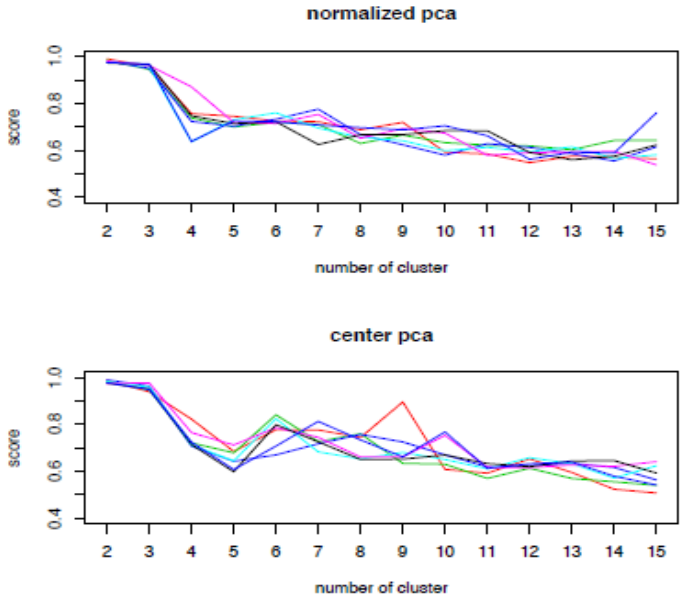

Figure 1. The minimum of dot similarities are estimated for a varying number of principal components p for each number of cluster with center and normalized principle components

B. Clustering Stability

In this work, several issues need to be considered:

- Whether the PCA should be normalized?
- How many leading principle components should be selected?
- How many clusters? or if the optimal decision is two clusters, is there any secondary decision at certain acceptable level?

Since there are more than one decisions that should be simultaneously determined and no mathematical model can be used to describe the complex relationship, using clustering stability may help us to make these decisions. Here is the algorithm to calculate stability score for a fixed number of clusters k [3]:

1) Perform K-means on the original data set (the referencing clustering).
2) Randomly sample a fraction f, larger than 0.5, of the original dataset for t times
3) Perform K-means on the each sub-sample subset (the resampling clusterings)
4) Calculate similarities between the referencing and the resampling clustering

The essential concept in the theory of clustering stability is resampling. It is reasonable to assume that if the decisions respect to the above questions are suitable, the results under the same framework are similar. More specifically, when the structure in the given data is represented well by k clusters, the reference clustering will be similar to the result obtained from the sub-sample data. The similarity measures considered in the papers are dot similarity and confusion matrix.

Table 1: the values for dot similarity and confusion matrix

dot similarity								
		2	3	4	5	6	7	8
2 pc	n	0.99	0.96	0.76	0.75	0.73	0.72	0.69
	c	0.99	0.94	0.82	0.68	0.78	0.78	0.74
3 pc	n	0.98	0.94	0.74	0.70	0.72	0.71	0.63
	c	0.98	0.95	0.72	0.68	0.84	0.73	0.76
4 pc	n	0.98	0.95	0.72	0.70	0.73	0.78	0.67
	c	0.99	0.96	0.71	0.64	0.67	0.72	0.76
5 pc	n	0.98	0.96	0.64	0.73	0.76	0.69	0.66
	c	0.98	0.96	0.71	0.64	0.82	0.68	0.65
6 pc	n	0.98	0.96	0.87	0.72	0.71	0.75	0.65
	c	0.97	0.98	0.76	0.71	0.79	0.74	0.66
9 pc	n	0.97	0.97	0.75	0.71	0.72	0.63	0.67
	c	0.98	0.95	0.71	0.60	0.80	0.73	0.65
12 pc	n	0.97	0.97	0.64	0.73	0.72	0.71	0.70
	c	0.98	0.95	0.72	0.61	0.71	0.81	0.73

confusion matrix								
		2	3	4	5	6	7	8
2 pc	n	0.99	0.96	0.78	0.72	0.73	0.72	0.68
	c	0.99	0.95	0.83	0.65	0.82	0.76	0.72
3 pc	n	0.98	0.95	0.69	0.68	0.71	0.68	0.63
	c	0.98	0.96	0.60	0.65	0.84	0.71	0.78
4 pc	n	0.98	0.96	0.73	0.67	0.71	0.75	0.67
	c	0.99	0.97	0.62	0.70	0.59	0.70	0.67
5 pc	n	0.98	0.97	0.58	0.70	0.73	0.69	0.64
	c	0.98	0.97	0.64	0.67	0.86	0.70	0.68
6 pc	n	0.98	0.97	0.87	0.70	0.71	0.73	0.67
	c	0.97	0.98	0.71	0.67	0.81	0.70	0.64
9 pc	n	0.97	0.97	0.71	0.69	0.72	0.63	0.67
	c	0.98	0.96	0.65	0.58	0.79	0.68	0.67
12 pc	n	0.97	0.97	0.60	0.70	0.72	0.66	0.65
	c	0.98	0.96	0.67	0.64	0.72	0.84	0.73

4. EVALUATION

A. III Dataset

The dataset was collected by Institute For Information Industry which consists of 109 different households in Northern Taiwan between Aug 2017 and September 2018. We treat each record, one day, as an individual points in the process of clustering. Thus, the records in the same customers over periods of time could be belong to different clusters. Although our sample rate is 1 minute, we perform our experiment at sample rate 15 minute to simulate the real condition in Taiwan. The sample rate of smart meter in Taiwan is 15 minute.

B. Experimental result

The stability of the clustering as measured by the minimum among similarity measures for these two measures are plotted for a varying number of principal components in Figure 1-2. These two have similar tendency. The values are briefly presented in table 1. Partitions into 2 and 3 clusters were stable regardless of the number of principal components and normalization, as evidenced by similarity scores being close to 1. Partitions into 4, 6 clusters were most stable for 2, 3 PCs, respectively, with similarity scores above 0.8 with center pca.

Fig 3-6 respectively show the average of each group obtained from the partitions into 2, 3, 4, 6 by K-means with 3 leading principle components. Using center principle components provides more clear separation based on energy usage time and magnitude. On the other hand, although the six clustering is not the optimal decision, it reveals more structures inside the whole data. Intuitively, the six-clustering may be interpreted as the result by splitting the groups obtained from the partitions into 2, 3 or 4, but it need more research work to verify the point.

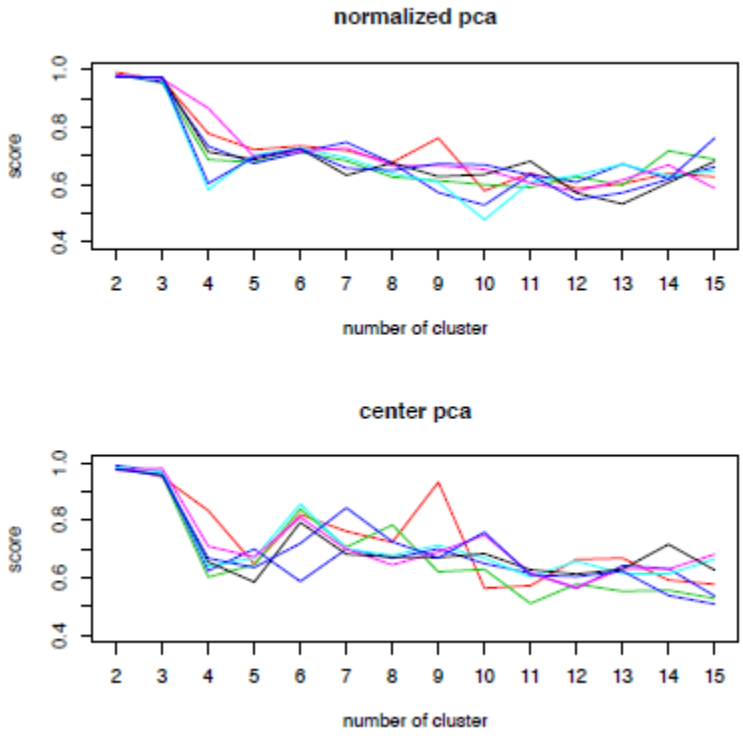

Figure 2. The minimum of confusion matrices are estimated for a varying number of principal components p for each number of cluster with center and normalized principle components

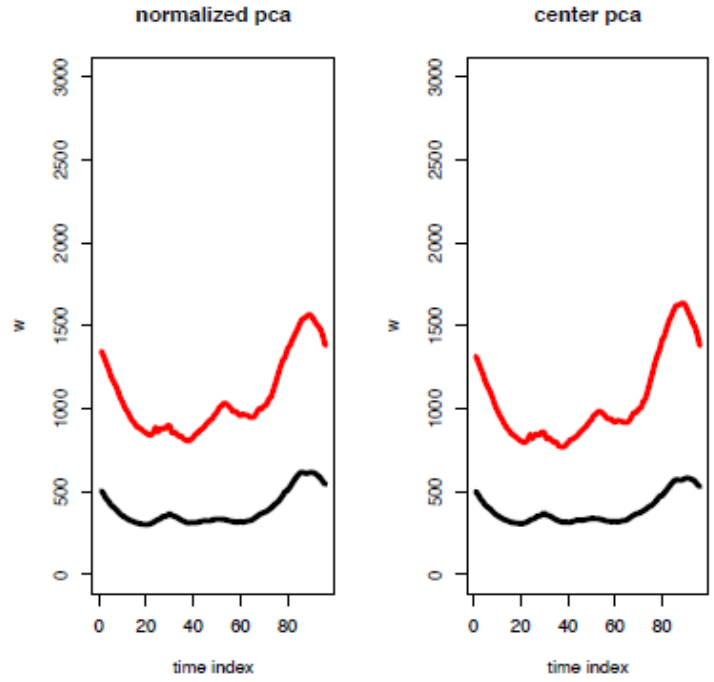

Figure 3. The average energy patterns as the number of cluster is two

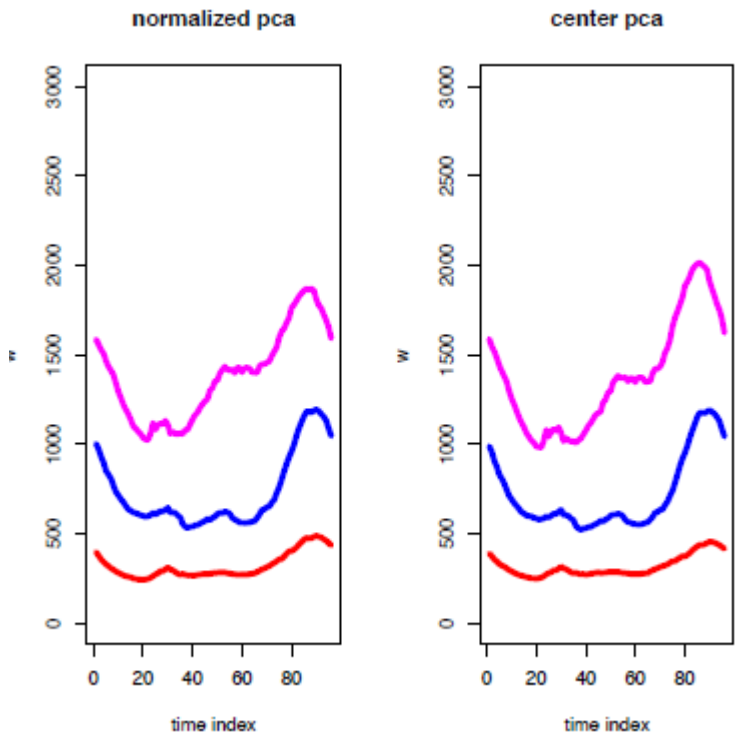

Figure 4. The average energy patterns as the number of cluster is three

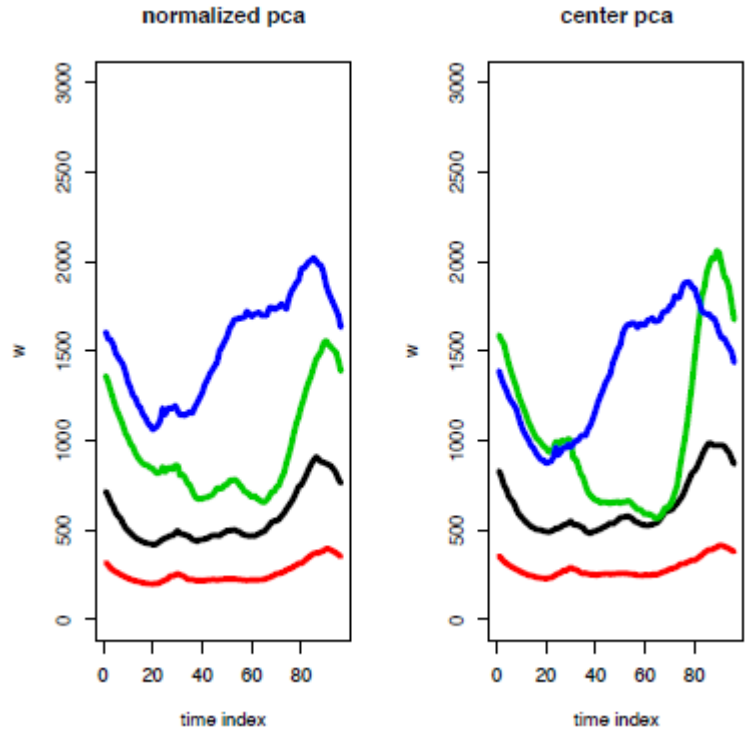

Figure 5. The average energy patterns as the number of cluster is four

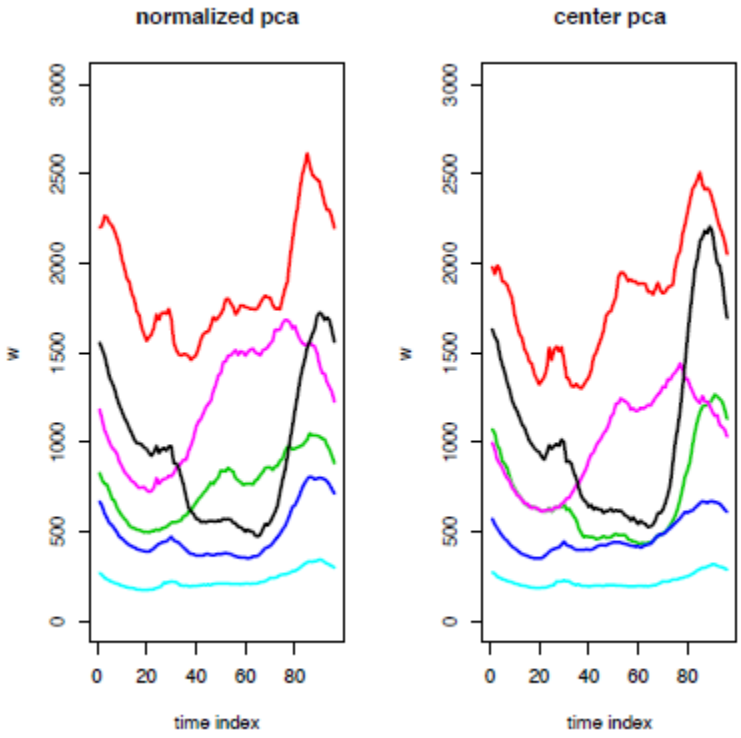

Figure 6. The average energy patterns as the number of cluster is six

5. CONCLUSION

It is the first try to cluster smart meter data with the evaluation by stability, and this paper shows the empirical results. We think that there has space either to develop theory of clustering stability or to use the evaluation in more smart meter data set.

Back to our problems as mentioned in section II, clustering stability indeed help us to make these decisions at the same time. There are no big different performance between center and normalized pca in term of stability, but the center one can provide more clear separation clustering based on energy usage time and magnitude. According to our empirical results, 2 or 3 cluster are the optimal decisions. The transition from average 0.9 level to average 0.7 level occurs between k=3 and k=4. However, the secondary decisions may consider 4, 6 with 2, 3 pcs, respectively, with the stability scores above 0.8. As mentioned in the section I, partitions into 2 or 3 are too simple for electrical operators or retailers to make any management decision, and thus we must to find the secondary solutions with certain reliability.

The clustering of smart meter is a nonparameter or unsupervised learning problem. There are no suitable mathematical model describing the complex relationship mentioned in section II up to now. Hence, we think that clustering stability is a nice try for clustering of smart meter.

ACKNOWLEDGMENT

We thank the Bureau of Energy, Ministry of Economic Affairs of Taiwan, ROC for the financial support under Contract No.107-E0215.

REFERENCES

[1] I. Jolliffe. Principle component analysis. Wiley Online. 1967.

[2] Ulrike von Luxburg. Clustering stability: an overview. Found. Trends Mach. Learn., vol. 2, 235-274. 2010.

[3] Asa Ben-Hur, and Isabelle Guyon. Detecting Stable Clusters Using Principal Component Analysis. In Functional Genomics: Methods and Protocols. M.J. Brownstein and A. Kohodursky (eds.) Humana press, pp.159-182, 2003.

[4] Yi Wang, Qixin Chen, Tao Hong, Chongqing Kang. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid, 2018.

[5] Stephen Haben, Colin Singleton, Peter Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Transactions on Smart Grid, 235-274. 2015.

[6] Simmhan, Yogesh, and Noor, Muhammad Usman. Scalable Prediction of Energy Consumption using Incremental Time Series Clustering. In IEEE International Conference on Big Data, 2013.

[7] Christian Beckel, Leyna Sadamori, Thorsten Staake, Silvia Santinic. Revealing household characteristics from smart meter data. Energy, 78:397-410, 2014.

[8] Adrian Albert and Ram Rajagopal. Smart Meter Driven Segmentation: What Your Consumption Says About You. IEEE Transactions on Power System, 2013.

[9] Charalampos Chelmis, Jahanvi Kolte, and Viktor K. Prasanna. Big data analytics for demand response: Clustering over space and time. In IEEE International Conference on Big Data, 2015.

[10] Viktoria Neimane . Distribution network planning based on statistical load modeling applying genetic algorithms and Monte-Carlo simulations. In 2001 IEEE Porto Power Tech Proceedings, 2001.

[11] Schalk W. Heunis and Ron Herman A probabilistic model for residential consumer loads. In IEEE Trans. Power Syst., vol. 17, no. 3, pp. 621625, Aug. 2002.

[12] E. Carpaneto and G. Chicco Probabilistic characterisation of the aggregated residential load patterns. In IET Gen., Transm. Distrib., 2007.