Hierarchical Clustering of Load Profile Database in Understanding of Korean Standard Industrial Classification

Seunghyoung Ryu*, Hongseok Kim*†, Doeun Oh, Jung-il Lee** and Jaekoo Noh****

Abstract – The accurate load prediction has been a long-cherished wish in the electric power industry. To achieve that, understanding load patterns and customer classification should be done in advance. In this paper, we study load clustering from the viewpoint of yearly and daily load pattern and discuss its relation to the Korean Standard Industrial Classification. In doing this we select hierarchical clustering, a well-known method for classifying unlabeled data, to cluster a load database. We find that KSIC can be well characterized by daily load pattern but not by yearly load pattern.

Keywords: Hierarchical clustering, Load profile, Load pattern clustering, Industrial classification.

1. Introduction

Generalization of advanced metering infrastructure (AMI) made it possible to gather huge amount of customer load data. There have been many attempts to classify these load data based on the shape of 24 hour load pattern. To the best of our knowledge, only possible labels for load data are the industrial index and contract classification. Our purpose is to understand the relation between load pattern-based clustering and the K*orean Standard Industrial Classification* (KSIC) label which may seem to have similar load pattern in the same industrial class.

The remainder of this paper is structured as follows. Section 2 shows a brief introduction for database. Section 3 explains the clustering algorithm we use. Section 4 presents clustering results and discuss the relation between *cluster representative load pattern* (CRLP) and KSIC. Finally, conclusions are given in Section 5.

2. Database

From the Korea Electric Power Corporation's (KEPCO) database, we acquire total 980 customers' three years of hourly load data which is classified with 75 different labels of KSIC and provincial locations without revealing the customer's identity. The database consists of the collection of highest electric consumers for almost every province-KISC pair. Before clustering the data, unusual datasets has been removed. To focus on the effect of weekdays, we filter out the weekend and holiday load data. After a filtering process, the database is composed of 762 customers' load dataset and each dataset contains 750 days of 24 hour electric consumption.

Each customers' dataset then yields *representative daily load pattern* (RDLP) and *representative yearly load pattern* (RYLP) to classify customers according to daily and seasonal load variation. RDLP is obtained by averaging hourly consumptions over 3 years, which makes 24 dimensional vector and RYLP is a 250 dimensional vector with daily maximum load of the first one year in each dataset. Both of them are normalized. Length of RYLP is fixed to 250 which is the number of working days in one year. Finally two matrices with RDLP (762 by 24) and RYLP (762 by 250) are used for experiment.

3. Clustering Methods

Clustering is the task of grouping unlabeled data so as to classify each data point into explainable groups. Since 1932, many clustering algorithms such as K-means, hierarchical clustering, expectation maximization (EM), DBSCAN has been developed. Specifically in this paper, we use hierarchical clustering to classify the customer load pattern and determine the proper number of clusters according to validation indices.

2.1 Hierarchical Clustering

For clustering data in a hierarchical manner, we need to choose a metric and linkage criterion. Metric is a measure of distance between two vectors, and we adopt Euclidean distance. On the basis of whole possible pairwise distance,

* Dept. of Electronic Engineering, Sogang University, Korea
** Dept. of Software Engineering, Korea Electric Power Corporation (KEPCO), Korea

† Corresponding Author (hongseok@sogang.ac.kr)

clusters coalesce into a new cluster according to the linkage criterion. The distance (or similarity) between groups is derived from the linkage criterion and arbitrary two clusters are merged together when they have minimum cluster distance.

According to the average distance linkage criterion [1], distance between two clusters A and B is defined as

$$\mathrm{d(A,B)} = \frac{1}{N_A N_B} \sum_{a \in A} \sum_{b \in B} d(a,b), \qquad (1)$$

where N_i represents the number of i cluster member and $d(a,b)$ is the Euclidean distance of vector a and b.

Unlike the average distance linkage criterion, the Ward's criterion is focused on reducing the total within cluster variance [1]. The similarity measure between two clusters A and B can be represented as

$$s(\mathrm{A,B}) = \frac{N_A N_B}{N_A + N_B} d^2(c_A, c_B), \qquad (2)$$

where c_i is the centroid of i cluster. Note that similarity is based on squared distance of centroids.

2.2 Validity Assessment Indices

The good clustering has short within-cluster distance and long inter-cluster distance. Many clustering validity indices represent this ratio in different form. We selected *Davies-Bouldin Index* (DBI) and *Calinski-Harabasz Index* (CHI) [2]. The low DBI means well clustered whereas CHI is the opposite. Thus, the proper cluster number K needs to concurrently achieve low DBI and high CHI. But the problem is CHI decreasing in K whereas high CHI means better clustering. So we also plot the *within groups sum of squares* (WSS). The elbow point of WSS curve is considered as a proper cluster number K. The Ward criterion that has higher CHI and WSS value than average criterion is used for clustering. In this way, we choose the number of cluster K to 9 for RDLP and 8 for RYLP, respectively.

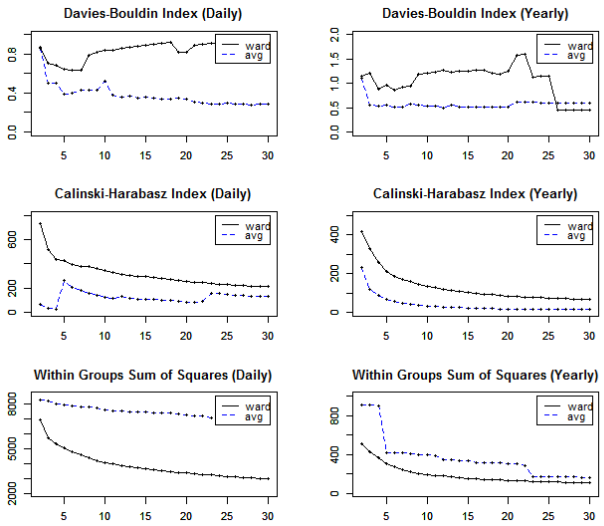

Fig. 1. Clustering validity indices with K = 2 to 30.

4. Results

The results in Fig. 2 and Fig. 3 show *cluster representative daily load pattern* (CRDLP) and *cluster representative yearly load pattern* (CRYLP). As can be seen, each cluster has its own distinct load pattern. For example, the cluster 1 and 2's CRDLP exhibit the time invariant feature whereas other clusters show load concentration during the day time except cluster 6 and 9. The seasonal variation appears in Fig. 3, which also shows various seasonal patterns of different clusters.

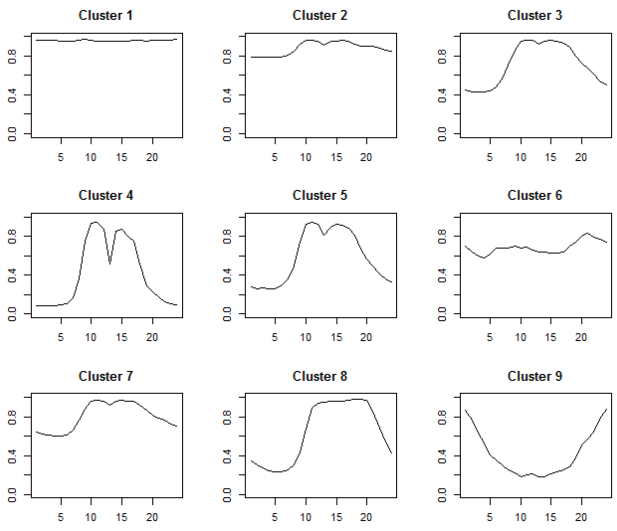

Fig. 2. CRDLP of the hierarchical clustering with K=9.

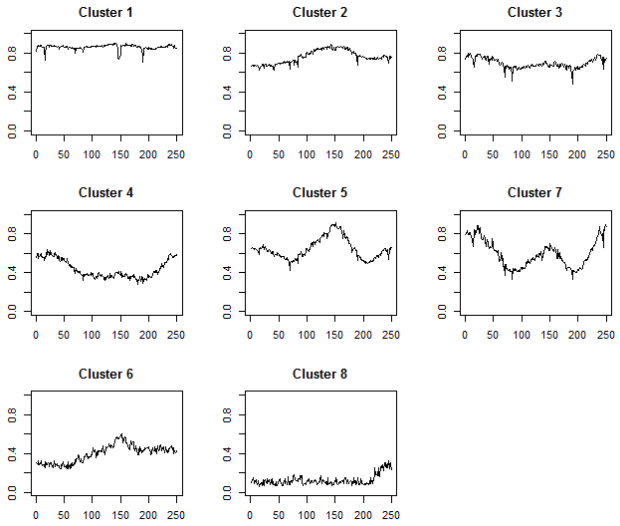

Fig. 3. CRYLP of the hierarchical clustering with K=8.

The results in Fig. 4, 5 show customers distributed in two dimensional space and its classification via hierarchical clustering. To visualize data points into two dimensional space, we adopt *classical multidimensional scaling* (CMDS) [3], which is based on the distance between vectors, so as to resemble hierarchical clustering. We see that customers are well clustered using the proposed method.

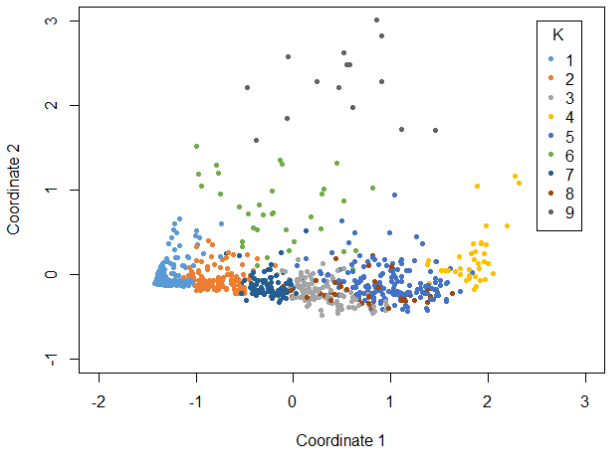

Fig. 4. Two dimensional visualization of RDLP via CMDS.

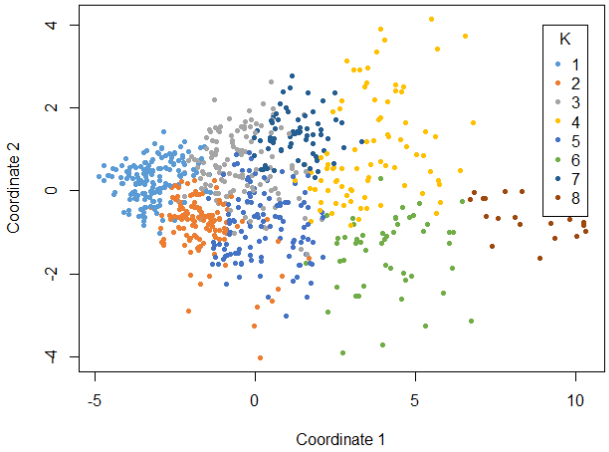

Fig. 5. Two dimensional visualization of RYLP via CMDS

Fig. 6 and Fig. 7 show the relationship between the proposed clustering and KSIC. For the customers in the same KSIC label, we check into which cluster of RDLP and RYLP the customers fall. As can be seen in Fig. 6, some industrial classes have a dominant cluster (occupying more than e.g., 50%) while dominant yearly load pattern is hardly observed in Fig. 7.

The interpretation is as follows. Due to having similar working pattern, the same industrial class tends to have similar daily load pattern. By contrast, the yearly load variation has different aspect. This kind of aspect can be perhaps explained when the yearly load variation is more dependent on a facility's HVAC system rather than working pattern of their industrial fields, but further investigation is needed. Restructuring KSIC in accordance with RYLP may be also required.

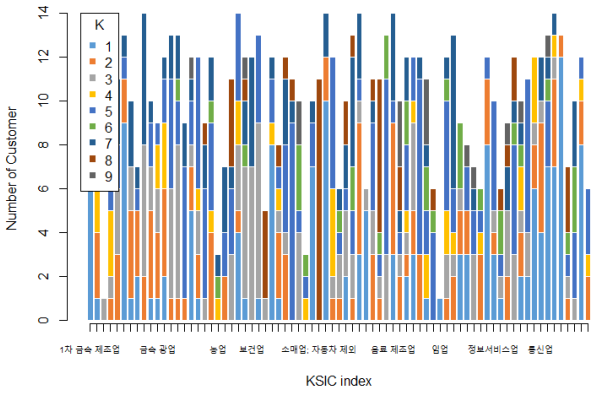

Fig. 6. RDLP Cluster distribution of KSIC.

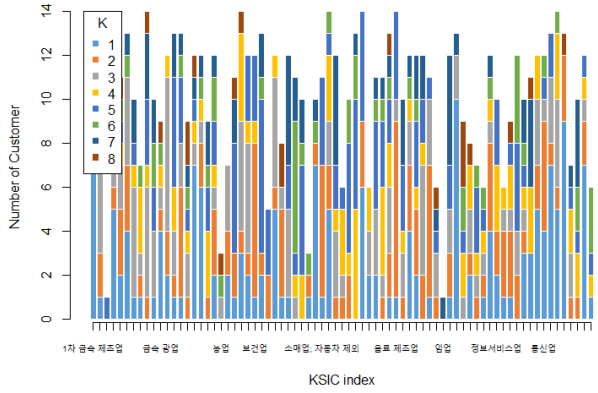

Fig. 7. RYLP Cluster distribution over KSIC

6. Conclusion

So far in this paper, we clustered load data with the viewpoint of yearly and daily variation and also investigated the relationship between the KSIC and load pattern. We found the tendency such that if customers are in the same industrial class they have similar daily load pattern, but interestingly, their yearly load patterns are diverse. In this study, we used the shape of pattern by normalizing the actual load, so the real amount of electricity consumption is not considered this time, which remains for further research. In addition we plan to use the result of load classification to improve the accuracy of load prediction.

Acknowledgements

This research was supported by Korea Electric Power Corporation through Korea Electrical Engineering & Science Research Institute. (grant number : R14XA02-50)

References

[1] Gianfranco Chicco, Roberto Napoli and Federico Piglione, “Comparisons among Clustering Techniques for Electricity Customer Classification”, *IEEE Trans. Power Systems*, vol.21, no.2, pp.933-940, May. 2006.

[2] Ujjwal Maulik and Sanghamitra Bandyopadhyay, “Performance Evaluation of Some Clustering Algorithms and Validity Indices,” IEEE Trans. *Pattern Analysis and Machine Intelligence*, vol.24, no.12, pp.1650-1654, Dec. 2002

[3] Trevor F Cox and Michael A. A. Cox, *Multi-dimensional Scaling*: CRC Press, 2000.

Seunghyoung Ryu
He received B.S degree in electronic engineering from Sogang University. His research interests are machine learning, smart grid and demand forecast.

Hongseok Kim
He is an Associate Professor in the Department of Electronic Engineering at Sogang University, Korea. He received his B.S. and M.S. degrees in electrical engineering from Seoul National University in 1998 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering at the University of Texas at Austin in 2009. His research interests include energy big data analysis, optimization and control in smart grid with ESS and renewable, demand response, etc.

Doeun Oh
He is a Principal Researcher in the Department of Software Engineering at Korea Electric Power Corporation (KEPCO). He receives his M.S degrees in computer engineering from Chungnam University in 2002. His research interests are energy big data platform and analysis.

Jung-il Lee
He is a Senior Researcher in the Department of Software Engineering at Korea Electric Power Corporation (KEPCO). He receives his B.S degrees in computer engineering from Chonbuk University in 2004. His research interests are demand response and forecast.

Jaekoo Noh
He is an Associate Researcher in the Department of Software Engineering at Korea Electric Power Corporation (KEPCO). He receives his B.S degrees in computer engineering from Konkuk University in 2006. His research interests are database management and demand forecast.