Journal of Electrical Engineering & Technology Vol. 7, No. 3, pp. 281~287, 2012 **281**
http://dx.doi.org/10.5370/JEET.2012.7.3.281

Repeated Clustering to Improve the Discrimination of Typical Daily Load Profile

Young-Il Kim†, Jong-Min Ko*, Jae-Ju Song* and Hoon Choi**

Abstract – The customer load profile clustering method is used to make the TDLP (Typical Daily Load Profile) to estimate the quarter hourly load profile of non-AMR (Automatic Meter Reading) customers. This study examines how the repeated clustering method improves the ability to discriminate among the TDLPs of each cluster. The k-means algorithm is a well-known clustering technology in data mining. Repeated clustering groups the cluster into sub-clusters with the k-means algorithm and chooses the sub-cluster that has the maximum average error and repeats clustering until the final cluster count is satisfied.

Keywords: Repeated clustering, K-means, Typical load profile, Discrimination, Cosine similarity

1. Introduction

The load analysis of the distribution line of the electric power system is one of the necessary technologies to operate the distribution line efficiently. To analyze the load of a distribution line every 15 minutes, the 15-minute load values of circuits or sections in the distribution line must be collected. In the past, the 15-minutes load values were measured in the substation, and saved to the SOMAS (Substation Operating results MAnagement System), and used to analyze the load of distribution lines. The power quality monitoring system is researched to analyze, diagnose, and improve the power quality of potential risk facilities or area [1]. However, to analyze the load of all the sections, installation of the automatic metering units in all facilities along the entire distribution line is required. Unfortunately, it is hard to achieve due to the installation cost.

Load analysis using the virtual load profile (VLP) has been studied as an alternative [2]. VLP is generated by assigning the monthly usage of a non-automatic meter reading (non-AMR) customer to the real load profile (RLP) of an automatic meter reading (AMR) customer that has a similar load pattern. The load of the section can be analyzed by combining all the RLPs of AMR customers and VLPs of non-AMR customers of that section [3].

By using RLPs and VLPs, a stable operation of the distribution line is provided in the field through management and automation of the distribution network [4, 5], prediction of old facilities, load forecasting [6, 7], and market price forecasting [8]. To generate the VLP of a non-AMR customer, it is necessary to obtain the typical load profile (TLP). TLP is an average load profile (ALP) of a group of customers who have similar load patterns. VLP is calculated by assigning the monthly energy usage of a non-AMR customer into TLP which is similar with the non-AMR customer's load pattern [9].

Fig. 1 shows the example of load analysis using the RLP, TLP, and VLP. Distribution line S has customers C1, C2, and C3. C1 and C2 are AMR customer, then have RLPs. C3 is non-AMR customer, and then has no RLP. Thus, the load profile (LP) of distribution line S is unpredictable. However, if the VLP of C3 is estimated by using TLPs and classification algorithm, the LP of distribution line S is estimated by summing up the RLPs of C1 and C2, and VLP of C3. Therefore, improvement of accuracy of VLP is to be a major factor of distribution analysis performance.

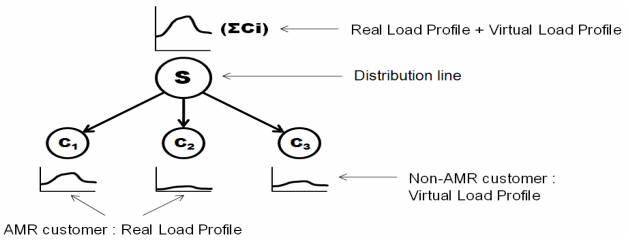

Fig. 1. Example of load analysis using RLP and VLP

To increase the accuracy of VLP of a non-AMR customer, it is important to select the most suitable TLP for the customer. The studies [2, 7], and [8] use error distance to evaluate the performance of TLP. The error distance is calculated by Euclidean distance between TLP of each cluster and RLP of the customer included in each cluster. The main purpose of this method is to minimize the total average error distance by minimizing the error distance of

† Corresponding Author: S/W Center, KEPCO (Korea Electric Power Cooperation) Research Institute, KEPCO, Korea. (yikim@kepco.co.kr)
* S/W Center, KEPCO Research Institute, KEPCO, Korea. (kojm@kepco.co.kr, jjsong@kepco.co.kr)
** Dept. of Computer Engineering, Chungnam National University, Korea (hc@cnu.ac.kr)
Received: December 29, 2010; Accepted: January 31, 2012

each cluster. However, this method has a weakness in discriminating ability. If data set has large amount of LPs with similar load pattern, many TLPs with similar pattern will be generated and it will be hard to choose the TLP related with non-AMR customer.

Discrimination of load curves of TLPs is an important factor when choosing the TLP. In this study, the repeated clustering is proposed to increase the discriminating ability of TLPs. This method groups entire data set into small number of n sub-clusters with k-means clustering and chooses the sub-cluster that has the largest average error distance, and groups the sub-cluster into n sub-clusters with k-means clustering, repeatedly. This method has the advantage of improving the discriminating ability of TLPs by generating the various load patterns of TLPs.

This paper is organized as follows: an introduction is shown in Section 1. In Section 2, the basic approach of customer clustering using 15-minute RLP of an AMR customer is presented and in Section 3, the process of repeated clustering, proposed in this paper is explained. Section 4 presents the experimental result analysis. Statistics of error distance is analyzed among hierarchical, k-means, FCM (Fuzzy C-Means), and repeated clustering. Similarity is analyzed between k-means clustering and repeated clustering using the RLPs of 3183 AMR customers. Classification accuracy is analyzed for four kinds of clustering algorithm. Section 5 is the conclusion.

2. Customer Clustering using RLP

The basic approach of customer clustering is using the k-means algorithm with 15-minute RLP of an AMR customer. The k-means algorithm is one of the clustering methods grouping adjacent data in a specific area and classifying them into several groups. Assuming that data is a dot on a vector space, it minimizes the degree of variance of groups by classifying the dots in each group to minimize the Euclidean distance to the group where a customer belongs compared to the Euclidean distance to the central value of other groups [10].

Fig. 2 shows the example of clustering 3 groups with the RLP of 14 AMR customers. Fig. 2(a) is a clustering result using the k-means algorithm, and Fig. 2(b) is the method proposed in this paper. In the case of the k-means algorithm, its goal is to minimize the error distance between the center node and each RLP. The center node in each cluster is an ALP of RLPs that were grouped in each cluster. In Fig. 2(a), customers A1 ~ A7, A8 ~ A11, and A12 ~ A14 are clustered in each cluster. A1 has a different load pattern compared with A2 ~ A7, but is clustered in C1, because it has less error distance than when making A1 into another cluster. Therefore, Fig. 2(a) has a minimum error distance, but discriminating ability is low. In Fig. 2(b), customer A1 has its own cluster C'1 to improve the discriminating ability, but error distance is increased.

The main purpose of customer clustering is generating clusters and TLPs using the clustering algorithm, and classifying non-AMR customers into clusters and generating the VLP by assigning energy usage into classified TLP. In the case of (a), C2 and C3 have similar load patterns, so it is hard to decide whether to assign a non-AMR customer into C2 or C3. However, in the case of (b), A1 can be easily classified into C'1, A2 ~ A7 into C'2, and A8 ~ A14 into C'3. Therefore, it is efficient to increase the discrimination of TLP to improve the classification performance; though it increases the error distance of clustering results.

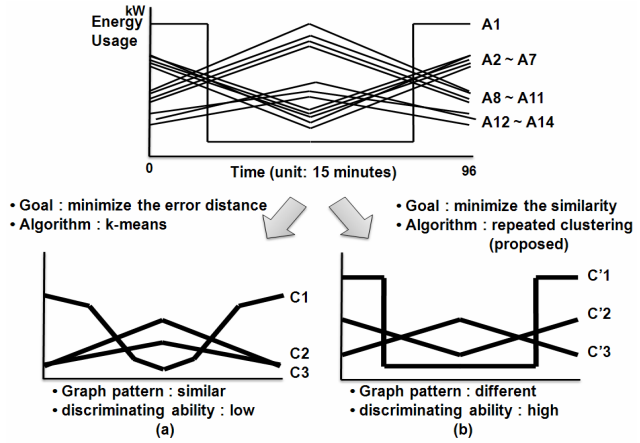

Fig. 2. Example of clustering with RLPs of AMR customers

3. Process of Repeated Clustering

This section describes the process of repeated clustering proposed in this paper; pre-processing, normalization and repeated clustering. The pre-processing process removes empty data or noises from collected AMR data and the normalization process helps the pattern comparison and analysis procedures between RLPs and VLPs by scaling the maximum value of the RLPs of each AMR customers to 1. The repeated clustering process groups the data set into several sub-clusters and chooses a suitable sub-cluster and groups the sub-cluster into several sub-clusters. The choosing and grouping of the sub-cluster are repeated until the sub-cluster count is satisfied the final cluster count.

3.1 Pre-processing and normalization

Fig. 3 shows the clustering step; data collecting, pre-processing and normalization, and clustering. Daily LP of AMR customer (A) of every 15 minutes energy usage (p_i) is as follows:

$$LP_{day}^{A} = [\, p_1^A, p_2^A, \ldots, p_i^A, \ldots, p_{96}^A \,] \quad (1)$$

In the data collecting step, daily LP data is collected from AMR customers. In the pre-processing step, daily LP

data is filtered and those with missing data or outlier values are eliminated. In the normalization step, LP data is normalized to reduce the maximum load value to 1. This normalization method helps the clustering algorithm to group the cluster easier by reducing the scale. A normalized daily LP is as follows:

$$nLP_{day}^{A} = \left[\frac{p_1^A}{p_{max}^A}, \frac{p_2^A}{p_{max}^A}, \dots, \frac{p_i^A}{p_{max}^A}, \dots, \frac{p_{96}^A}{p_{max}^A}\right] = \left[\, l_1^A, l_2^A, \dots, l_i^A, \dots, l_{96}^A \,\right] \quad (2)$$

$$p_{max}^A = max\,(p_1^A, p_2^A, \dots, p_i^A, \dots, p_{96}^A) \quad (3)$$

Data Collecting → Pre-processing & Normalization → Clustering (k=10)

	00:00	00:15	00:30	...	23:45
Cust 1	35.60	37.22	36.18	...	34.96
Cust 2	8.570	8.930	8.350	...	8.710
Cust 3	4.460	4.250	4.460	...	4.680
Cust 4	6.880	6.560	6.120	...	7.030
Cust 5	14.06	13.82	13.44	...	14.45
Cust 6	17.46	17.14	17.03	...	17.82
Cust 7	15.62	14.33	14.51	...	16.27
Cust 8	3.130	2.990	2.700	...	3.460
...	...	...	...	...	...
Cust n	14.69	15.02	14.69	...	15.26

	00:00	00:15	00:30	...	23:45
Cust 1	0.704	0.736	0.715	...	0.691
Cust 2	0.292	0.305	0.285	...	0.297
Cust 3	0.209	0.199	0.209	...	0.219
Cust 4	0.394	0.376	0.351	...	0.403
Cust 5	0.742	0.729	0.709	...	0.762
Cust 6	0.669	0.657	0.652	...	0.683
Cust 7	0.530	0.487	0.493	...	0.552
Cust 8	0.277	0.265	0.239	...	0.306
...	...	...	...	...	...
Cust n	0.420	0.429	0.420	...	0.436

	00:00	00:15	00:30	...	23:45
Cluster 1	0.802	0.807	0.797	...	0.806
Cluster 2	0.210	0.202	0.193	...	0.220
Cluster 3	0.774	0.769	0.765	...	0.773
Cluster 4	0.841	0.896	0.944	...	0.799
Cluster 5	0.548	0.534	0.520	...	0.567
Cluster 6	0.328	0.311	0.294	...	0.351
Cluster 7	0.233	0.226	0.222	...	0.239
Cluster 8	0.350	0.339	0.328	...	0.383
Cluster 9	0.348	0.333	0.320	...	0.364
Cluster 10	0.521	0.492	0.472	...	0.547

Fig. 3. Example of data collecting, pre-processing and normalization, and clustering

3.2 Repeated Clustering

Final cluster count, sub-cluster division count, and minimum customer count per cluster is needed to perform the repeated clustering; final cluster count is the number of clusters in this algorithm; sub-cluster division count is the number of sub-clusters that repeated clustering groups the source cluster into target sub-clusters; minimum customer count is the number of customers that allows clustering of the source cluster, if it has more customers in it.

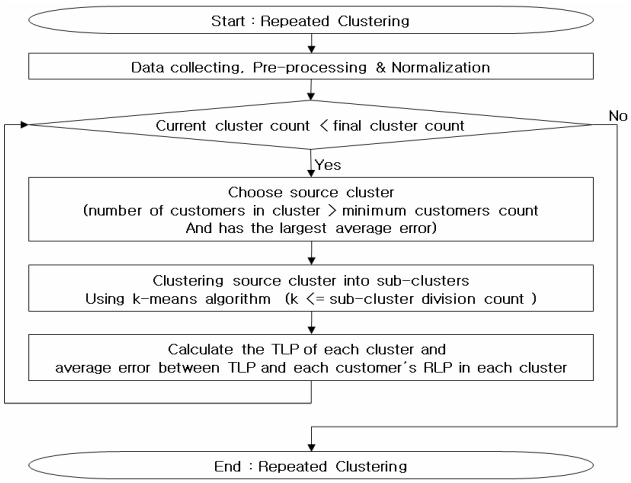

Fig. 4. Flow diagram of repeated clustering

Fig. 4 shows the flow diagram of repeated clustering. After pre-processing and normalization, LPs of the source cluster are grouped into sub-clusters using the k-means algorithm. TLPs of grouped clusters and average errors between TLP and each customer's RLP in each cluster are calculated. Daily TLP of cluster C_k, which has m customers, is as follows:

$$TLP_{day}^{C_k} = \left[\mu_1^{C_k}, \mu_2^{C_k}, \dots, \mu_i^{C_k}, \dots, \mu_{96}^{C_k}\right] \quad (4)$$

$$\mu_i^{C_k} = \frac{\sum_{j=1}^{m} l_i^{A_j}}{m} \quad (5)$$

The error between TLP and RLP of the customer ($d_i^{C_k A_j}$), and average error of cluster ($D_{avg}^{C_k}$) are as follows:

$$D_{avg}^{C_k} = \left[\frac{\sum_{j=1}^{m}\sum_{i=1}^{96} d_i^{C_k A_j}}{96 \times m}\right] \quad (6)$$

$$d_i^{C_k A_j} = \left|\mu_i^{C_k} - l_i^{A_j}\right| \quad (7)$$

After sub-clustering, the current cluster count is compared with the final cluster count to determine whether repeated clustering will be continued or not. If the current cluster count is smaller than the final cluster count, it selects cluster that has more customers than the minimum customer count and also has the largest average error. The selected cluster is clustered into sub-clusters using the k-means algorithm. The repeated clustering is finished when the current cluster count equals the final cluster count. By this algorithm TLPs of all clusters are finally decided.

4. Experimental Result Analysis

4.1 Clustering result

3183 AMR data of high-voltage customers in KEPCO, Korea are collected to compare the proposed algorithm and the k-means algorithm. Fig. 5 shows an example of the proposed repeated clustering process with a final cluster count of 12, sub-cluster division count of 3, and minimum customer count of cluster as 100. 3183 daily ALPs are clustered into 3 sub-clusters with the k-means algorithm. The first 3 clusters have 3.00, 4.72, and 7.94 AED (Average Error Distance) in the first step. 7.94 AED cluster is selected to group into 3 clusters, then 5 clusters are generated in the second step; 3.00, 4.72, 4.67, 4.86, and 6.20 AED. In the third step, 6.20 AED cluster is selected, and grouped into 3 clusters. Clustering is repeated until the sixth step, and finally, 12 clusters and daily TLPs are generated.

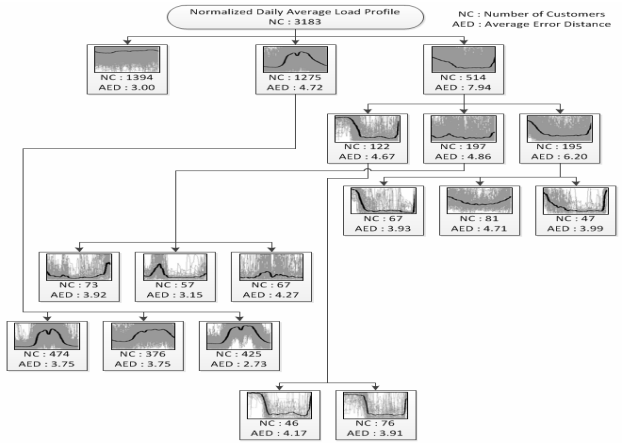

Fig. 5. Example of repeated clustering process

Fig. 6 shows the daily TLPs of 12 clusters and LPs of 3183 AMR customers. Each graph shows each cluster; the daily TLP of each cluster is marked as a bold line and the customer's daily average LP is marked as a gray line. Fig. 6(a) shows the results of k-means clustering, and Fig. 6(b) shows the results of repeated clustering. The k-means clustering (a) has 4 similar patterned TLPs such as TLP 1, 4, 6, and 11. The repeated clustering (b) has 2 similar patterned TLPs such as TLP 2, and 11. The repeated clustering has less similar patterned TLPs than k-means clustering, and therefore is presumed to have more discriminating ability by examining the patterns of the TLPs at a glance.

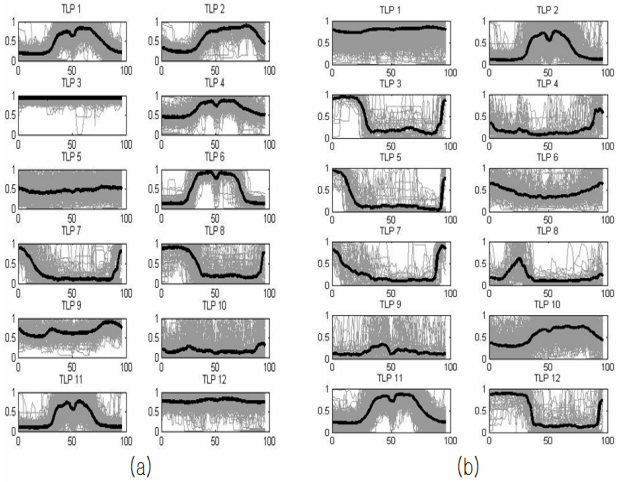

Fig. 6. TLP and LP graph: (a) k-means clustering; (b) repeated clustering

To compare the statistics of error distance, the hierarchical, k-means, FCM, and repeated clustering method are experimented. Table 1 shows the cluster count and the sum, the mean, the standard deviation, and the variance of the error distances between RLPs and TLPs for four kinds of clustering method. As the cluster count increase, the error distance between RLPs and TLPs are decrease because the RLP's candidates for the TLP are increased, and thus statistics of the error distance and discriminating ability of TLPs are increased. When the cluster count is fixed, the k-means clustering is better to minimize the sum and the means of error distance because k-means uses an error distance for an objective function. The repeated clustering has an object to generate the various patterns of TLP, and thus has better standard deviation and variance of error distance. The standard deviation shows how much dispersed from the average. The proposed clustering method has lower standard deviation and variance than the k-means clustering. The lower standard deviation and variance are decreased, the more RLPs are closely clustered into TLPs and the higher discriminating ability is increased.

4.2 Discrimination analysis using cosine similarity

Table 1 shows that repeated clustering has less performance in AED than k-means clustering. However, the main goal of this study is the discrimination improvement of TLP by taking the risk of AED increase.

In this study, the cosine similarity is used to analyze the similarity among TLPs. The cosine similarity is a measure

Table 1. Cluster count and statistics of error distance for four kinds of clustering method

Cluster Count	Clustering method	Statistics of error distance			
		Sum	Mean	Standard deviation	Variance
8	hierachical	5496	1.7268	0.7678	0.5895
	k-means	4753	1.4933	0.7555	0.5708
	FCM	4839	1.5203	0.7967	0.6347
	proposed	5943	1.8671	0.6532	0.4266
10	hierachical	5242	1.6467	0.7299	0.5382
	k-means	4547	1.4285	0.7283	0.5304
	FCM	4735	1.4874	0.8021	0.6434
	proposed	5546	1.7425	0.6135	0.3763
12	hierachical	5220	1.6400	0.7494	0.5433
	k-means	4379	1.3757	0.7356	0.5412
	FCM	4583	1.4398	0.7855	0.6170
	proposed	5406	1.6983	0.5986	0.3584
14	hierachical	5180	1.6274	0.7240	0.5216
	k-means	4282	1.3452	0.7207	0.5194
	FCM	4597	1.4441	0.7979	0.6366
	proposed	5312	1.6688	0.5941	0.3529
16	hierachical	5136	1.6135	0.7209	0.5122
	k-means	4192	1.3170	0.7105	0.5048
	FCM	4554	1.4309	0.8032	0.6451
	proposed	5263	1.6534	0.5936	0.3523
18	hierachical	5033	1.5813	0.7150	0.5022
	k-means	4132	1.2980	0.6986	0.4881
	FCM	4525	1.4215	0.8011	0.6418
	proposed	5199	1.6335	0.5923	0.3508
20	hierachical	4997	1.5699	0.6901	0.4625
	k-means	4064	1.2768	0.6742	0.4545
	FCM	4505	1.4152	0.8028	0.6445
	proposed	5156	1.6199	0.5859	0.3433

of similarity between two vectors of n dimensions by finding the cosine of the angle between them. The cosine similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity [11].

Given two vector of TLPs, C1($C1_1$, $C1_2$, …, $C1_{96}$) and C2($C2_1$, $C2_2$, …, $C2_{96}$), the cosine similarity of C1 and C2 is represented as

$$Cos(\theta) = \frac{C1 \cdot C2}{\|C1\|\|C2\|} = \frac{(C1_1 \times C2_1 + C1_2 \times C2_2 + \cdots + C1_{96} \times C2_{96})}{\sqrt{C1_1^2 + C1_2^2 + \cdots + C1_{96}^2} \times \sqrt{C2_1^2 + C2_2^2 + \cdots + C2_{96}^2}} \quad (8)$$

Fig. 7 shows the cosine similarity among TLPs for each clustering methods that grouped 3183 AMR data into 12 clusters. The diagonal values represent the cosine similarity between TLP itself, trivially 1. The left-down triangular values represent the cosine similarity between two TPLs for k-means clustering. The right-top triangular values represent for the repeated clustering. The minimum and maximum cosine similarity values are indicated by bold font and bold box. It is difficult to find the most and the worst similar load patterns with the eyes in Fig. 6, and therefore it is needed to calculate the cosine similarity to evaluate the discriminating ability of each TLP and it is shown in Fig. 7. In case of the k-means clustering, TLP 3 and TLP 12 have the most similar load patterns as cosine similarity 0.99. TLP 6 and TLP 7 have the most different load patterns as cosine similarity 0.34.

	TLP1	TLP2	TLP3	TLP4	TLP5	TLP6	TLP7	TLP8	TLP9	TLP10	TLP11	TLP12
TLP1	1.0000	0.8133	0.7442	0.7921	0.6465	0.9691	0.7457	0.7992	0.9267	0.9617	0.8999	0.7552
TLP2	0.9534	1.0000	0.3845	0.4443	**0.3503**	0.6729	0.3799	0.4803	0.9204	0.8793	**0.9754**	0.3972
TLP3	0.8764	0.9143	1.0000	0.7042	0.8926	0.8484	0.8647	0.8220	0.5713	0.5652	0.4743	0.9501
TLP4	0.9651	0.9615	0.9710	1.0000	0.6443	0.8713	0.9170	0.6844	0.6216	0.7150	0.5477	0.6617
TLP5	0.8771	0.9339	0.9953	0.9683	1.0000	0.7710	0.8581	0.5508	0.5057	0.4986	0.4169	0.8236
TLP6	0.9615	0.8550	0.8182	0.9201	0.8000	1.0000	0.8678	0.8140	0.8371	0.8879	0.7771	0.8321
TLP7	0.4186	0.4894	0.7217	0.5885	0.7220	**0.3473**	1.0000	0.6920	0.5791	0.6033	0.4700	0.8088
TLP8	0.4862	0.5154	0.8096	0.6732	0.7781	0.4703	0.8852	1.0000	0.6756	0.6490	0.5850	0.8907
TLP9	0.8602	0.9329	0.9872	0.9575	0.9953	0.7832	0.6965	0.7538	1.0000	0.9205	0.9530	0.6206
TLP10	0.7531	0.8151	0.9487	0.8869	0.9456	0.7100	0.7480	0.8304	0.9493	1.0000	0.9504	0.5556
TLP11	0.9601	0.8558	0.7881	0.9010	0.7765	0.9725	0.3474	0.4170	0.7517	0.6684	1.0000	0.4963
TLP12	0.8931	0.9223	**0.9992**	0.9786	0.9942	0.8369	0.7116	0.7973	0.9840	0.9410	0.8099	1.0000

Fig. 7. Cosine similarity for each clustering methods (left-down: k-means, right-top: proposed)

Fig. 8 shows the cosine vectors of the minimum and maximum similarity for repeated clustering that grouped the AMR data into 12 clusters. Repeated clustering has a maximum discrimination between TLP 2 and TLP 5 because of the minimum cosine similarity as 0.35. TLP 2 and TLP 11 have a minimum discrimination because of the maximum cosine similarity as 0.97 and it can be distinguished with the eye.

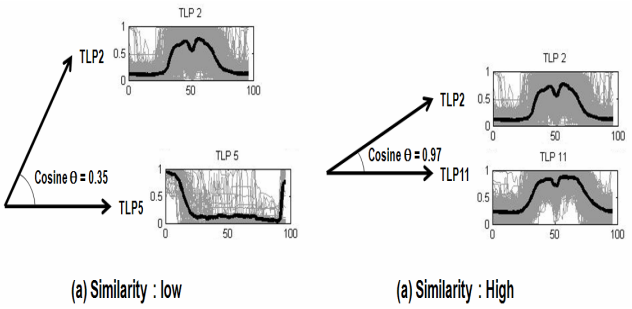

Fig. 8. Minimum and maximum cosine similarity of proposed method

The hypothesis testing method [12], one of the statistical inference theories, is used to prove that the repeated clustering allows more discrimination than the k-means clustering.

- Null hypothesis H0: the cosine similarity of the repeated clustering(μ) is equal to the cosine similarity of the k-means clustering(μ_0) ($\mu = \mu_0$)
- Alternative hypothesis H1: the cosine similarity of the repeated clustering(μ) is smaller than the cosine similarity of the k-means clustering(μ_0) ($\mu < \mu_0$)

The Mann-Whitney test [12], one of the best-known non-parametric tests, is used because the cosine similarity data of experimental testing results is not a normal distribution. The experimental result is calculated by changing the cluster count from 8 to 20 and the results are shown in Table 2. For cluster counts from 8 to 20, the cosine similarity of the repeated clustering is smaller than the k-means clustering. The p-value is smaller than 0.05, and thus, the null hypothesis is rejected and the alternative hypothesis is accepted.

Table 2. Cosine similarity and p-value for hypothesis testing of each clustering method

Cluster Count	Clustering method	Cosine Similarity				p-value
		Sum	Mean	Standard deviation	Variance	
8	k-means	23.02	0.8220	0.1573	0.0247	0.0281
	proposed	21.26	0.7594	0.1447	0.0209	
10	k-means	36.29	0.8065	0.1705	0.0291	0.0028
	proposed	32.37	0.7194	0.1697	0.0288	
12	k-means	53.54	0.8112	0.1720	0.0296	0.0004
	proposed	47.18	0.7149	0.1770	0.0313	
14	k-means	73.00	0.8022	0.1798	0.0323	0.0005
	proposed	65.96	0.7249	0.1814	0.0329	
16	k-means	97.00	0.8084	0.1711	0.0293	0.0004
	proposed	88.66	0.7388	0.1822	0.0332	
18	k-means	121.94	0.7970	0.1768	0.0313	0.0001
	proposed	110.58	0.7227	0.1871	0.0350	
20	k-means	148.13	0.7796	0.1836	0.0337	0.0093
	proposed	139.01	0.7316	0.2001	0.0400	

The repeated clustering, proposed in this paper, improves the accuracy of VLP classification of non-AMR customers by increasing the discriminating ability of TLP. This method will contribute the load analysis performance of distribution line which is used TLPs and VLPs. Classification accuracies of the hierarchical, k-means, FCM, and repeated clustering are calculated to show up the improvement of classification accuracy. The PNN (Probability Neural Network) classification algorithm [13] with 10-fold cross validation is used to calculate the average accuracy of classification. The 3183 AMR data is partitioned into 10 folds, and each fold have 318 AMR data. Of the 10 folds, a single fold is retained as the validation data for testing the model, and the remaining 9 folds are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 folds used exactly once as the validation data. Table 3 shows that classification accuracy of the repeated clustering is improved 5% than the k-means clustering. This means that the VLP estimation error of non-AMR customers is reduced by 5%. Therefore, the load analysis performance of distribution line is improved by using the repeated clustering.

Table 3. Classification accuracy of clustering methods

Experiment Number	Clustering method			
	hierachical	k-means	FCM	proposed
1st	85.8%	90.3%	87.7%	87.1%
2nd	90.3%	93.7%	89.0%	96.5%
3rd	82.1%	85.5%	74.5%	87.1%
4th	85.8%	88.7%	85.2%	94.3%
5th	81.8%	84.3%	74.2%	94.0%
6th	80.8%	80.8%	74.2%	86.2%
7th	84.0%	85.8%	76.1%	90.9%
8th	87.4%	84.3%	78.6%	89.0%
9th	77.7%	75.8%	72.0%	88.7%
10th	86.5%	85.8%	73.6%	90.6%
Average	84.2%	85.5%	78.5%	90.4%

5. Conclusion

This study proposes the repeated clustering method to increase the discrimination of the typical load profile of each cluster. The repeated clustering groups the cluster into sub-clusters and chooses the cluster that has the maximum average error and repeats clustering until the final cluster count is satisfied.

The average daily load profiles of 3183 customers are used to compare the similarity between the k-means clustering method and repeated clustering method. ALPs are grouped into 20 clusters by each clustering method. The cosine similarities among TLPs of each cluster are calculated and the Mann-Whitney hypothesis testing is used to prove that the discrimination of TLPs is improved. The PNN classification algorithm is used to show that the classification accuracy of VLP of non-AMR customers is increased by increasing the discriminating ability of TLP.

Future studies should examine ways to enhance repeated clustering to determine the optimal sub-cluster division count and minimum customer count automatically and without user interaction. When the proposed method is applied to the load analysis of a distribution line, pattern discrimination of the TLP curve will be increased and performance of load analysis will be improved.

Acknowledgements

This work was supported by the research project (Development of integrated solution for continuous daily demand response based on demand estimation) of KEPCO Research Institute funded by the KEPCO.

References

[1] Dong-Jun Won, Il-Yop Chung, Joong-Moon Kim, Seon-Ju Ahn, Seung-Il Moon, Jang-Cheol Seo, and Jong-Woong Choe, “Power Quality Monitoring System with a New Distributed Monitoring Structure”, KIEE International Transactions on Power Engineering, vol.4-A. no.4, pp.214-220, 2004.

[2] David Gerbec, Samo Gasperic, Ivan Smon, and Ferdinand Gubina, “Allocation of the Load Profiles to Consumers Using Probabilistic Neural Networks”, IEEE Transactions on Power Systems, Vol. 20, No. 2, May 2005, pp. 548-555.

[3] Young-Il Kim, Jong-Min Ko, and Seung-Hwan Choi, “Methods for Generating TLPs (Typical Load Profiles) for Smart Grid-Based Energy Programs”, IEEE Symposium Series on Computational Intelligence 2011, vol.1, pp.49-54, 2011.

[4] Jeong-Do Park, “Unit Commitment for an Uncertain Daily Load Profile”, KIEE International Transaction on Power Engineering, vol.5-A, no.1, pp.16-21, 2005.

[5] Jong-Young Park, Soon-Ryul Nam, and Jong-Keun Park, “Real-Time Volt/VAr Control Based on the Difference between the Measured and Forecasted Loads in Distribution Systems”, Journal of Electrical Engineering and Technology, vol.2, no.2, pp.152-156, 2007.

[6] J.A. Jardini, “Daily Load Profile for Residential, Commercial and Industrial Low Voltage Consumers”, IEEE Transaction on Power Delivery, vol.15, pp.375-380, 2000.

[7] N.M. Pindoriya, S.N. Singh, and S.K. Singh, “Forecasting of Short-Term Electric Load Using Application of Wavelets with Feed-Forward Neural Networks”, International Journal of Emerging Electric Power Systems, vol.11, no.1, pp.1-24, 2010.

[8] SanJeev Kumar Aggarwal, Lalit Mohan Saini, and Ashwani Kumar, “Electricity Price Forecasting in

Ontario Electricity Market Using Wavelet Transform in Artificial Neural Network Based Model", International Journal of Control, Automation, and Systems, vol.6, no.5, pp.639-650, October 2008.

[9] Young-Il Kim, Jin-Ho Shin, Jae-Ju Song, and Il-Kwan Yang, "Customer Clustering and TDLP (Typical Daily Load Profile) Generation Using the Clustering Algorithm", International Conference of IEEE Transmission and Distribution Asia 2009, vol. 1, pp.1-4, 2009.

[10] Jain A. K. and Dubes R.C., 1988. "Algorithms for Clustering Data", Englewood Cliffs, NJ: Prentice-Hall.

[11] Van Rijsbergen, C. J., "Information Retrieval, 2nd edition", London: Buttersworth, 1979.

[12] Lehmann, E.L., and Joseph P. Romano, "Testing Statistical Hypotheses, 3rd edition", New York: Springer, 2005.

[13] H. Demuth and M. Beale, "Neural Network Toolbox for Use With MATLAB", Natick, MA: MathWorks, Jun. 2001.

Young-Il Kim He was born in Nonsan, Korea, on Nov. 27, 1972. He received his Master's Degree in Computer Engineering from Chungnam National University, Korea in 2000. His employment experience included the K4M Company, Electronics and Telecommunications Research Institute (ETRI), and KEPCO Research Institute. He is responsible for implementation of RTP (Real-Time Pricing) and DR (Demand Response) system. His research interests include RTP-DR, SmartGrid, load analysis, data mining, and RFID/USN.

Jong-Min Ko He was born in Korea, on Nov. 30, 1967. He received his Master's Degree in Computer Engineering from Chungnam National University, Korea in 2004. He is a senior member of S/W Center in KEPCO Research Institute. He is responsible for design and implementation of RTP-DR system. His research interests include RTP-DR, Power IT, SmartGrid, mobile widget, smart phone application.

Jae-Ju Song He was born in Korea, on May. 25, 1967. He received his Master's Degree in Computer Science from Chungbuk National University, Korea in 2004. He is a principal technical staff of S/W Center in KEPCO Research Institute. He is responsible for technical management of RTP-DR application. His research interests include RTP-DR, SmartGrid, IHD (In-Home Display), and RFID/USN.

Hoon Choi He is a professor of the Department of Computer Science and Engineering, the Chungnam National University (CNU), Korea. He received a MS and a PhD in computer science from Duke University in 1990 and 1993, respectively. His research area includes the system software for mobile, distributed computing and the communication middleware.