J Electr Eng Technol Vol. 7, No. 6: 807-813, 2012
http://dx.doi.org/10.5370/JEET.2012.7.6.807

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Short-term Electric Load Forecasting Using Data Mining Technique

Cheol-Hong Kim[†], Bon-Gil Koo* and June Ho Park**

Abstract – In this paper, we introduce data mining techniques for short-term load forecasting (STLF). First, we use the K-mean algorithm to classify historical load data by season into four patterns. Second, we use the k-NN algorithm to divide the classified data into four patterns for Mondays, other weekdays, Saturdays, and Sundays. The classified data are used to develop a time series forecasting model. We then forecast the hourly load on weekdays and weekends, excluding special holidays. The historical load data are used as inputs for load forecasting. We compare our results with the KEPCO hourly record for 2008 and conclude that our approach is effective.

Keywords: Data mining, K-mean algorithm, k-NN algorithm, Short-term load forecasting, Time series forecasting

1. Introduction

The power industry can no longer over-consume cheap electrical energy. This is because of the need to reduce emission of greenhouse gases that cause global warming and because of the rapid increase in the raw-material prices caused by resource exhaustion. Moreover, the stability of power supply is at risk because of the rapid increase in the power consumption and because of the changes in the usage pattern caused by upgrades in the industrial structure. It is necessary to develop short-term, mid-term, and long-term power load forecasting in order to actively cope with these changes in the power industry and to predict the future demand. This will facilitate economical operation since stable power supply and cost reduction will be ensured.

Power load forecasting can be divided into mid-term and long-term forecasting, which are conducted every year or month, and short-term forecasting, which is conducted every hour. Mid-term and long-term power load forecasting provides basic data for developing long-term plans for system operation, such as a power supply plan, power development plan, power transmission and transformation plan, fuel consumption plan, and investment plans. These forecasts are based on mid-term and long-term indices such as economic growth rate, rate of population increase, and increase in the price of raw materials. Short-term power load forecasting is closely related to the generation costs and reliability; further, it is strongly affected by the day, demand pattern, temperature, wind velocity, etc. The results of forecasting are used for controlling power systems, establishing short-term plans, calculating power tides, etc.

Existing methods for short-term load forecasting include general regression analysis and time series methods [1-3]. Recently, an expert system has been used to predict the load after establishing a database using the knowledge and experience of experts [4, 5]. Another method for predicting the short-term load involves the use of a nerve circuit network, which is an intelligent system [6-9]. In this study, the power loads were predicted using an auto regression integrated moving average (ARIMA) model, a time series method, and a simple exponential smoothing method; the results were then compared.

Data mining has been used to predict power loads more accurately and to classify them according to their patterns. Data mining can be defined as the process of analyzing or extracting new and meaningful information from data, with the aim of finding hidden trends and patterns. In data mining, high-class statistical analysis and modeling techniques are adopted, including clustering, classification, associative rules, and time-series sequential patterns.

This paper the aim of this study was to classify our country's actual power load patterns for 2008 using the K-mean and k-NN data mining techniques, which are representative data mining techniques, and to verify the effectiveness of these techniques by carrying out forecasting.

2. Data Mining

2.1 K-mean clustering

Clustering involves grouping physical or abstract objects into classes of similar objects. It is widely used in fields such as pattern recognition, data analysis, and image processing. It is an independent tool that obtains knowledge of the data distribution, observes the features of

† Corresponding Author: Kyungnam energy co., LTD., Korea. (hongkc@knenergy.co.kr)
* Dept. of Electric and Electronic Engineering, Pusan National University, Korea. (cancer2000@pusan.ac.kr)
** Dept. of Electric and Electronic Engineering, Pusan National University, Korea. (parkjh@pusan.ac.kr)

Received: February 7, 2011; Accepted: May 24, 2012

each cluster, and can focus on a particular cluster set for additional analysis.

K-mean algorithm was proposed by Cox (1957) and Fisher (1958) and developed by Hartigan (1975) and MacQueen (1967). It is the method most commonly used in clustering. It divides the data into K relatively similar clusters where k, the number of groups, is determined in advance [10]. The algorithm is as follows:

[Step 1] Select K arbitrary vectors from the dataset $[x_1,\ldots,x_N]$ and create a K initial central set $[y_1,\ldots,y_K]$.

[Step 2] If x_n is closest to y_i, add it to cluster x_i. After all, the dataset is divided into K clusters $[X_1,\ldots,X_K]$.
$X_i=\{x_n|d(x_n,y_i)\square d(x_n,y_i),j=1,\ldots,K\}$

[Step 3] For the new clusters obtained from Step 2, renew each center.
$y_i=c(X_i),\ i=1,\ldots,K$

[Step 4] Obtain the total distortion via the sum of the distances from the cluster centers nearest the data.
$D = \sum_{n=1}^{N} d(x_n,y_{i(n)})$
where $i(n)=k$, if $x_n \square X_K$

[Step 5] If the center values of the new cluster are equal to the values of the previous iteration, stop. Otherwise, repeat Steps 2–4 with the center value of a new cluster.

Fig. 1. shows an algorithm sequence diagram to classify power load types using data mining.

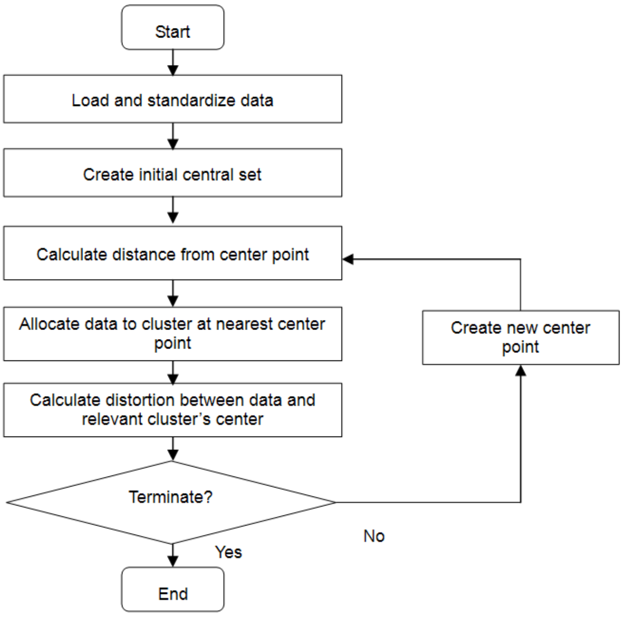
Start
Load and standardize data
Create initial central set
Calculate distance from center point
Allocate data to cluster at nearest center point
Calculate distortion between data and relevant cluster's center
Terminate?
Yes
No
End
Create new center point

Fig. 1. K-mean clustering

2.2 k-NN classification

Classification is widely used in data mining. It involves recognizing a feature of a new object and allocating it to a category according to a predefined classification code.

Classification has been used in fields such as image and pattern recognition, medical diagnosis, and credit approval. Its representative algorithms include k-Nearest Neighbor (k-NN), judgment analysis, and decision trees.

The problem of classification can be described as follows: “There are input data or records that form a training set, and each record has an attribute. The purpose of classification is to analyze the data based on a given feature and to develop an accurate model for each class. The test data are used to classify data whose class label is unknown.”

k-Nearest Neighbor is a simple algorithm that memorizes the training data and performs classification only if the attributes of an object exactly match one of the training examples. It finds the group of k objects in the training set that are closest to the test object and assigns a label to the test set based on the predominance of a particular class in this neighborhood. There are three key elements to this approach: a set of labeled objects, a set of stored records, and a distance or similarity metric to compute the distances between the objects. The algorithm is as follows [10]:

Compute $d(x',x)$, the distance between Z and every object $(x',y') \in D$. Select $D_x \subseteq D$, the set of the k training objects closest to Z.

Input: D, the set of training objects, and the test object $Z = (x',y')$.

Output: $$y' = \arg\max \sum I(v = y_i),\quad y(x_i,y_i) \in D_z \tag{1}$$

where, v : a class label,
y_i : the class label f or the i th nearest neighbors,
$I(\,)$: indicator function that returns 1 if its argument is true and 0 otherwise.

3. Load Forecasting

We predicted power loads using the simple exponential smoothing model and ARIMA, a simple and practical prediction technique.

3.1 Simple exponential smoothing model

The simple exponential smoothing model is a moving average method that places a greater weight on the recent demand when calculating the average.

$$F_{t+1} = F_t + \alpha(Z_t - F_t) \quad (2)$$

$$F_t = \alpha Z_{t-1} + (1-\alpha)F_{t-1} \quad (3)$$

where, F_{t+1} : Predicted value at time point t+1
Z_t : Observed value at time point t
Z-F : Error
α : Smoothing constant

Larger the smoothing constant α value (0–1), the greater the weight given to the recent data. For data with a particular pattern, more weight should be given to recent data, and for data without a particular pattern or with severe variation, more past data should be included. In addition, α should minimize the mean square error and the mean absolute percentage error (MAPE).

The advantage of this technique is that it needs only three data (the prediction value just obtained, the observed value, and smoothing constants) whereas the weighted moving average method requires the previously observed value and the weight.

3.2 ARIMA Model

3.2.1 Auto regression (AR) model

This model involves performing a regression of the time series itself, and the general p th order AR process$\{Z_t\}$ satisfies

$$(1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Z_t = a_t \phi \quad (4)$$

where, a_t : a white noise process with average 0 and dispersion $\sigma^2{}_a$
ϕ : auto regression factor
B : backward operator satisfying

$$BZ_t = Z_{t-1}, B^2 Z_t = Z_{t-2}, \ldots, B^n Z_t = Z_{t-n} \quad (5)$$

3.2.2 Moving average (MA) model

An MA process is a general linear process of Z_t whose limited number of weights is not 0, and a Z_t th order MA process can be written as

$$Z_t = (1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)a_t \quad (6)$$

where, a_t : white-noise process with dispersion $\sigma^2{}_a$
θ : parameter**.**

3.2.3 Auto regression moving average model (ARMA model)

The AR model and MA model can be combined into a low-order ARMA model when the orders p,q are high, i.e., when there are too many parameters. This will lead to a better approximate value. This time series is referred to as a mixed ARMA(p,q) with a p th order AR part and a q th order MA part and is expressed as follows:

$$\phi_p(B)Z_t = \theta_q(B)a_t \quad (7)$$

The characteristic equations for the AR and MA parts of an ARMA(p,q) model satisfying this equation are defined as follows:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \quad (8)$$

$$\phi_q(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_q B^q \quad (9)$$

3.2.4 Auto regression integrated moving average model (ARIMA model)

The average, dispersion, and autocovariance of time series data are assumed to be independent of temporal changes. However, in actual time series the data are usually abnormal i.e., the average and dispersion depend on the temporal flow. To normalize the data, logarithmic conversion or dispersed stabilization conversion can be applied to the dispersion, and a difference calculation can be applied to the average.

The d th order differential time series of the resulting normalized time series data becomes an ARIMA(p,d,q) model whose general equation is

$$\varphi_p(B)(1-B)^d Z_t = \varphi_q(B)a_t \quad (10)$$

4. Case Study

The electric load classified by season (spring, summer, fall, winter) and type of day (Monday, other weekday, Saturday, Sunday) using K-mean and k-NN data mining. We forecast the load for the next day using ARIMA and a simple exponential smoothing method. To demonstrate the performance of the method we used data for the Korean national electric load from 2008.

4.1 Electric load analysis

We used k-means to classify the load by season and k-NN to classify it by day type. Fig. 2. illustrates the procedure.

4.1.1 Seasonal electric load clustering using K-mean

We used 196 weekdays of data, omitting the summer vacation from mid-July to mid-August. This paper used to classify the electric load data with seasonal day type by use of K-mean. The result of the K-mean algorithm depends on the allocation of the initial center value of the clusters. We tested three initial allocations: allocating randomly,

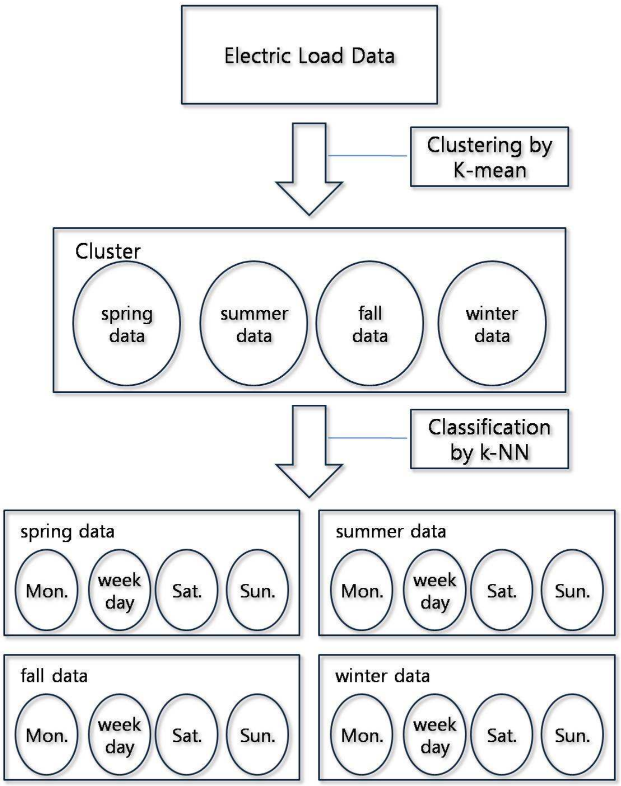

Fig. 2. Procedure for electric load classification

allocating with selected k electric load data, and allocating average values of each day type. The test converges where there is no change in the center values. The results show that the third option is the most efficient in terms of convergence speed and the ability to classify the data by season. Table 1 shows the differences in the results and Table 2 shows the result using the third option.

Table 1. Differences in seasonal load classification for different initial allocations.

	Center of Clusters	Seasonal Classification	Iteration
case 1	Random	Not clear	8.2
case 2	Select load data	Possible	6.64
case 3	Average values that possibly classified to the same cluster	Possible	3.84

Table 2. Seasonal load classification using option 3.

Season	Period
winter	January ~ February
spring	March ~ June 1st week
summer	June 2nd ~ September 3rd week
fall	September 4th ~ November 3rd week
winter	November 3rd ~ December

4.1.2 Daily electric load classification using k-NN

After classifying the data by season, we used k-NN to classify it by day type. When using k-NN, it is usually efficient to set k to an odd number less than 7. We classified the data into five day types (Monday, other weekday, Saturday, Sunday, holiday). We measured similarity using the Euclidean distance. Table 3 shows the results of this classification.

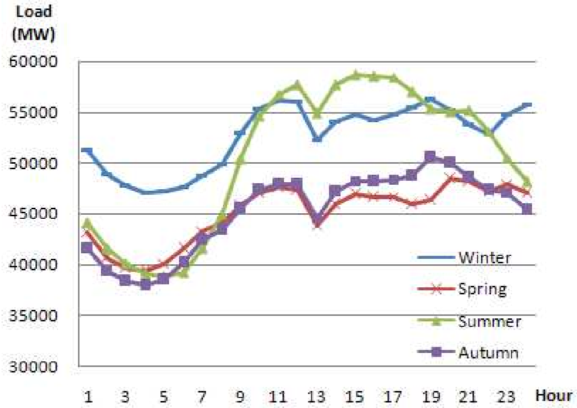

Fig. 3. Result of seasonal load classification

$$D_j = \sqrt{\sum_{i=0}^{N} (x_i - S_1)^2} \qquad (11)$$

where, D : Euclidean distance
i : time
x_i : data to be classified
s_i : training data

Table 3. Day type classification for different values of k.

Day type	k=1	k=3	k=5	k=7
Monday	46	45	45	45
Other weekday	184	184	184	184
Saturday	47	46	46	45
Sunday	42	42	40	40
Holiday	46	48	51	51

The weekday results are independent of k, but the other results are not. The k-NN method assumes that the training set has almost perfect accuracy. However, because the load has an annually increasing tendency and depends on the weather conditions, there are no perfect training data. We apply the following procedure: if the data have different day types we confirm this by eye and check that the MAPE is below 5%. We concluded that k=1 gave the best results. Figs. 4 to 7 illustrate the seasonal load patterns.

4.2 Electric load forecasting

We performed the forecasting using the 2008 data classified by K-mean and k-NN. SPSS 14.0K was used for ARIMA and Microsoft Excel 2007 was used for simple exponential smoothing. ARIMA used eight previous data and four day types (Monday, other weekday, Saturday,

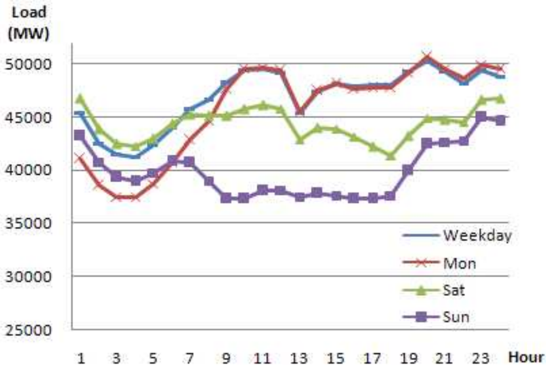

Fig. 4. Load pattern for spring

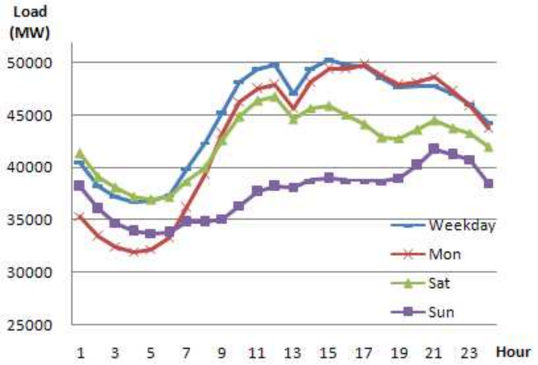

Fig. 5. Load pattern for summer

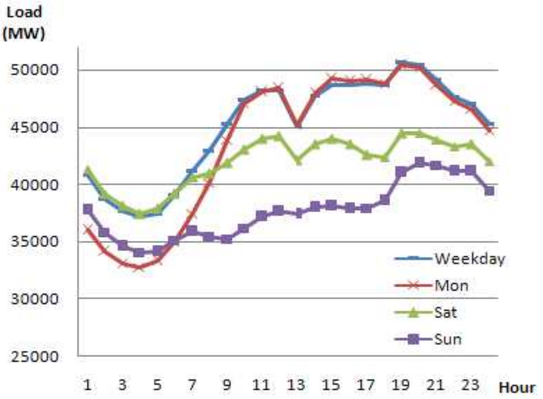

Fig. 6. Load pattern for fall

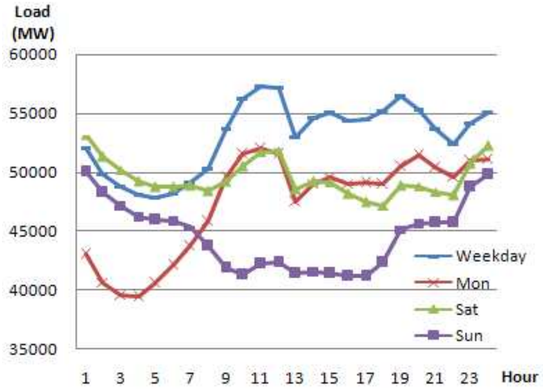

Fig. 7. Load pattern for winter

Sunday). To evaluate the performance of the model, we compared the results with actual electric load data by calculating the MAPE:

$$\mathrm{MAPE}(\%) = \frac{1}{N}\left[\frac{|z_t - x_t|}{x_t}\right]\times 100\% \tag{12}$$

where, z_t : forecast load
x_t : actual load
N : forecasting number

Figs. 8. to 11. give the results for fall. The two models almost match the actual load. Other-weekday forecasting is the most accurate. This is because of the time interval: the

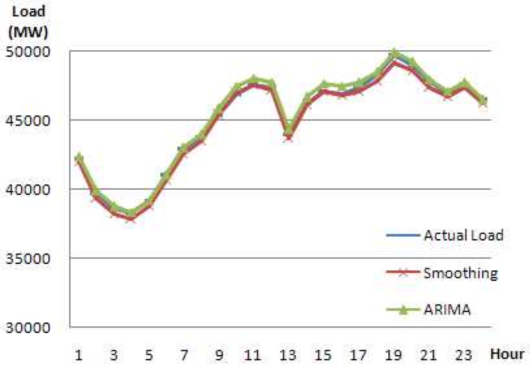

Fig. 8. Result of load forecasting (fall other-weekday)

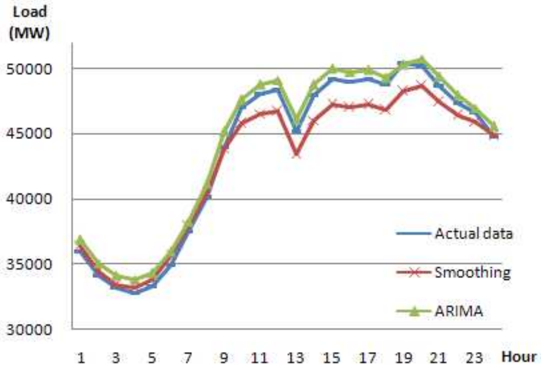

Fig. 9. Result of load forecasting (fall Monday)

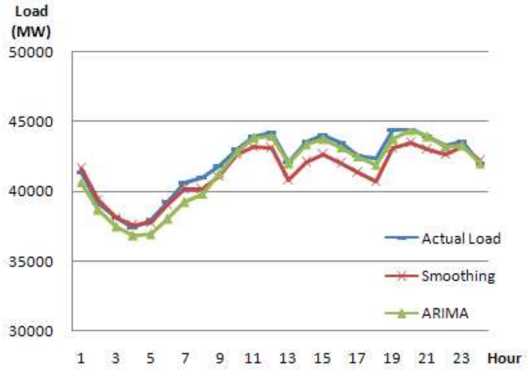

Fig. 10. Result of load forecasting (fall Saturday)

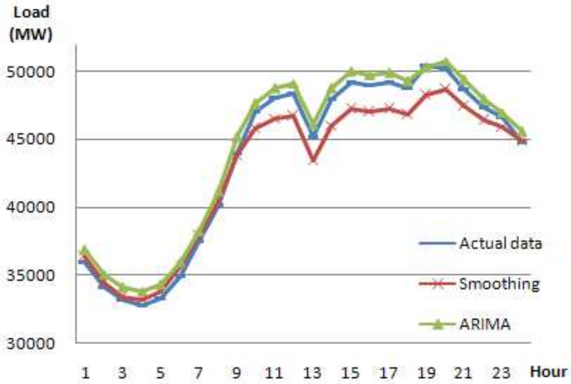

Fig. 11. Result of load forecasting (fall Sunday)

Table 4. shows the MAPE values.

		Min	Max	MAPE
week day	Simple exponential smoothing	0.9	2.1	1.5
	ARIMA	1.0	1.6	1.4
Mon.	Simple exponential smoothing	1.0	2.5	1.6
	ARIMA	1.0	1.5	1.2
Sat.	Simple exponential smoothing	1.0	2.5	2.1
	ARIMA	0.9	2.2	1.5
Sun.	Simple exponential smoothing	1.2	2.6	1.8
	ARIMA	0.1	1.9	1.5

load is dependent on the weather, so naturally there is a difference in accuracy between daily and weekly data. In spite of this difference, two models is about 1.5% accuracy, regardless of the day type.

5. Conclusion

In this paper, we suggested that electric load data should be classified by using K-mean and k-NN methods rather than by using a calendar, in order to improve forecasting accuracy. The load data are classified by season using k-means; then, the data are classified according to four day types. Using the classified data, we performed short-term electric load forecasting using the simple exponential smoothing and the ARIMA model.

The error for March and June was relatively large compared to that for winter. However, the error for each model was small (about 1.5%) for all day types, and the forecasting was mostly successful.

By performing a case study using actual electric load data, we proved the effectiveness of our model in practical applications. Nonetheless, there is a need to improve the accuracy of load forecasting. The inclusion of additional data, such as data on the weather, temperature, and time of day, or the detection of faulty data may improve the forecasting accuracy.

References

[1] G. Juberias, R. Yunta, J. Garcia, Moreno C. “Mendivil, A New ARIMA Model for Hourly Load Forecasting”, *Transmission and distribution Conference*, 1999 IEEE Vol.1, 11-16 April 1999, pp:314-319 vol.1

[2] Dae-Yong Kim, Chan-Joo Lee, Yun-Won Jeong, Jong-Bae Park, Joong-Rin Shin, Development of System Marginal Price Forecasting Method Using SARIMA Model, *Proceeding of conference KIEE* Nov. 2005, pp.148-150

[3] W. R. Christiaanse, “Short Term Load Forecasting Using Genera Exponential Smoothing,” *IEEE Trans. on Power Apparatus and Systems*, Vol. PAS-90, No. 2, pp. 900-910, March/April, 1971

[4] T. S. Dillon and M. A. Laughton, “Expert System Applications in Power Systems,” *Prentice Hall*, 1990

[5] S. Rahman, and R. Bhatnagar, “An Expert System Based algorithm for Short-Term Load Forecast,” *IEEE Trans. On Power Systems*, Vol. 3, No. 1, pp. 50-55, 1987

[6] Hyeonjoong Yoo, Russel L. Pimmel, “Short-Term Load Forecasting using a Self-Supervised Adaptive Neural Network”, *IEEE Transaction on Power System*, Vol.14, No.2, May 1999

[7] K. Y. Lee, Y. T. Cha. J. H. Park, “Short-Term Load Foreacasting using an Artificial Neural Network”, *Transaction on Power System vol.7*, No1, February 1992

[8] Koh Hee-Soek, Lee Chung-Sik, Kim Hyun-Deok, Lee Hee-Chul, “Short-term Load Forecasting using Neural Network”, *Proceeding of conference KIEE* 1993, pp.29-31, Nov. 1993

[9] Byoung-Su Kim, Ho-Sung Shin, Kyung-Bin Song, Jung-Do Park, “Short-Term Load Forecasting of Pole-Transformer Using Artificial Neural Networks”, KIEE: *Proceeding of conference KIEE* 2005, pp.810-812, July 2005

[10] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg “Top 10 algorithms in data mining”, *Knowledge and Information Systems*, vol.14, pp1~37, 2007

[11] William W. S. Wei, Time Series Analysis(Univariate And Multivariate Methods) 2nd Ed., Addison-Wesley, 2005

Cheol-Hong Kim He received his B.S. and M.S. Degrees from Pusan National University, Busan, Korea in 2003 and 2005, respectively. He is currently PH.D candidate at Pusan National University. He is also with Kyungnam Energy co., LTD. His research interests include power system operation and smart grid.

Bon-Gil Koo He received his B.S. and M.S. Degrees from Pusan National University, Busan, Korea in 2008 and 2010, respectively. He is currently PH.D candidate at Pusan National University. His research interests include electric load forecasting and intelligent systems applications to power systems.

June Ho Park He received the B.S., M.S. and Ph.D. degrees from Seoul National University, Seoul, Korea in 1978, 1980, and 1987, respectively, all in electrical engineering. He is currently a Professor at the School of Electrical Engineering, Pusan National University, Busan, Korea. His research interests include intelligent systems applications to power systems. Dr. Park has been a member of the IEEE POWERENGINEERING SOCIETY.