Journal of Korea Multimedia Society Vol. 20, No. 5, May 2017(pp. 800-807)
https://doi.org/10.9717/kmms.2017.20.5.800

An Adaption of Pattern Sequence-based Electricity Load Forecasting with Match Filtering

Fazheng Chu[†], Sung-Hwan Jung[††]

ABSTRACT

The Pattern Sequence-based Forecasting (PSF) is an approach to forecast the behavior of time series based on similar pattern sequences. The innovation of PSF method is to convert the load time series into a label sequence by clustering technique in order to lighten computational burden. However, it brings about a new problem in determining the number of clusters and it is subject to insufficient similar days occasionally. In this paper we proposed an adaption of the PSF method, which introduces a new clustering index to determine the number of clusters and imposes a threshold to solve the problem caused by insufficient similar days. Our experiments showed that the proposed method reduced the mean absolute percentage error (MAPE) about 15%, compared to the PSF method.

Key words: Load Forecasting, Time Series, Pattern Matching

1. INTRODUCTION

Electricity load forecasting has become necessary for power generation unit schedule. It is the basis of every profit maximization strategy, and it is vital for the security and reliability of the electric power transmission system. Generally, load forecasting aims to extrapolate pattern of load consumption under the effect of factors, such as weather and day-of-week, etc.

Many methods have been employed to load forecasting. As time series, auto-regressive integrated moving average model (ARIMA) is typically used by many researchers as a sophisticated benchmark for evaluating alternative proposals [1,2]. The regression related [3,4] and exponential smoothing [5,6] are always attractive because they are easily implemented. Nowadays artificial intelligence-based techniques also become attractive and widespread due to their flexibility and explanation capabilities, such as artificial neural network (ANN) [7].

However, the load forecasting problem is a complex nonlinear problem linked with social considerations, economic factors, and weather variations. It is difficult to obtain accurate and realistic models for such methods. It is also hard to assess methods since they are different from each other in scenarios and demands.

The pattern sequence-based forecasting (PSF) suggested in paper [8] is an approach to forecast the behavior of time series based on similar pattern sequences. In the approach, a label sequence is generated after clustering the load time series. And then similar predicted day-to-day pattern is searched out by matching to the label sequence. Finally, the average of the similar days is provided as the forecast.

The PSF algorithm is simple and effective, but it does not have a reliable method to determine the

※ Corresponding Author: Sung Hwan Jung, Address: (641-773) 20 Changwondaehak-ro, Changwon, Korea, TEL: +82-55-213-3815, FAX : +82-55-286-7429, E-mail : sjung@changwon.ac.kr
Receipt date: Feb. 20, 2017, Revision date : Mar. 31, 2017
Approval date: Apr. 8, 2017

[†] Economics and Management College of Qingdao Agricultural University
(E-mail: chufazheng@126.com)
[††] Dept. of Computer Engineering, Changwon National University

proper number of clusters during clustering. Another weak point is the high forecasting error rate caused by insufficient similar days in some cases. In this paper, we propose a new clustering index for selecting the number of clusters, and suggest imposing a filter on similar days to lower the forecasting error rate.

The rest of the paper is organized as the following sections: Section 2 presents the Pattern Sequence-based Forecasting (PSF) and its weak points. Section 3 elaborates the proposed method as the adaptation of PSF. Section 4 consists of summary and conclusion.

2. THE PATTERN SEQUENCE-BASED FORECASTING

The Pattern Sequence-based Forecasting (PSF) is a kind of pattern similarity-based methods. A common feature of these methods is learning from the data and using similarities between patterns of the seasonal cycles of the time series. In the PSF, instead of dealing high dimensional data, it converts the time series into a label sequence. Therefore, the method increases the efficiency in pattern matching.

2.1 Description of PSF Method

The PSF method could be described as two phases, and [Fig. 1] shows the flowchart of the method.

Phase 1: Clustering

In this phase, the initial data will be clustered and a label sequence which consists of cluster identifiers will be obtained. A typical clustering k-means is proposed[9] and three indexes is recommended to determine the number of clusters (parameter k). They are Silhouette index [10], Dunn index [11] and Davies-Bouldin index [12].

Phase 2: Prediction

In this phase, the average of similar day load is provided as the forecast to the day to be predicted. The similar day refers to the day which follows the same sequence segment as the day to be predicted does. In the case, the sequence segment is called query pattern and its length is called window size (parameter w).

A brief description is given as [Fig. 2]. Currently, the window size is 5. With the given window size w, for any day to be predicted if there is no matching results, the process will repeat with $w=w-1$. It is obvious that any day will obtain its similar days when $w=1$.

2.2 Obtaining Parameters in PSF

There are two parameters in the PSF method.

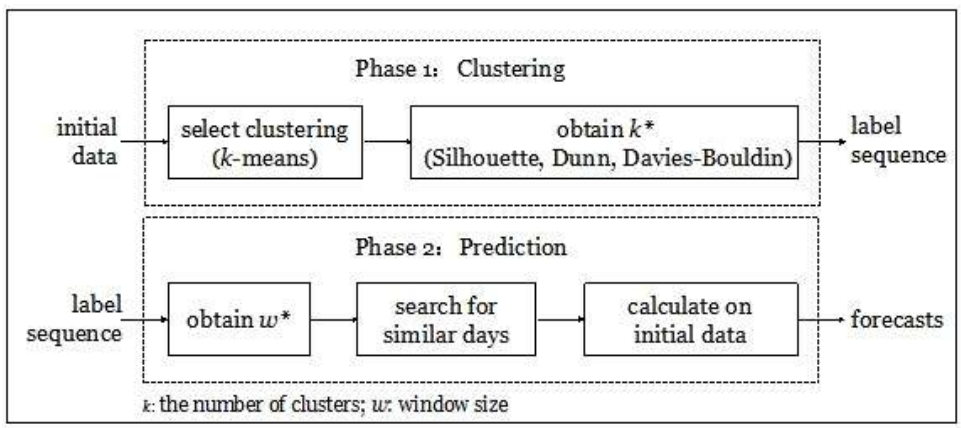

Fig. 1. The flowchart of PSF method.

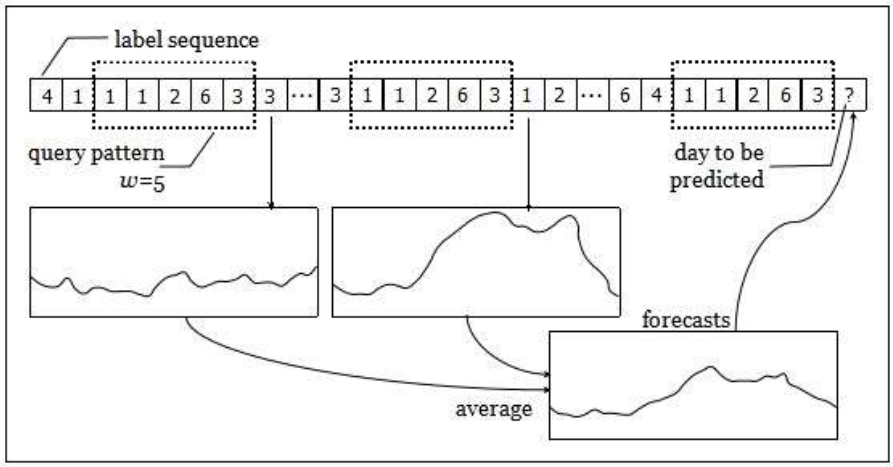

Fig. 2. Pattern matching and prediction calculation.

Both of them have significant impact on the level of forecasting error rate.

(1) Parameter k: The number of clusters. Since the process of how to select the number of clusters becomes the focus for clustering technique, many indexes [13] have been put forward to evaluate the performance of clustering. In PSF method, it recommends selection of appropriate k with three indexes: Silhouette index, Dunn index and Davies-Bouldin index.

(2) Parameter w: Windows size. The parameter should be determined by a training process that minimize the mean absolute percentage error (MAPE).

2.3 Issues in Application of PSF Method

For application of PSF, there are two issues which need attention.

(1) In the clustering phase, the indexes work incorrectly in selecting parameter k in some case. The label sequence is crucial for PSF method, so the number of clusters is a crucial point. However, with the help of the three recommended indexes, the value selected for parameter k does not work properly. For example, in clustering phase, the value k_1 is better than k_2 according to the indexes, but in prediction phase, the label sequence based on k_2 clusters may achieve a lower forecasting error rate than the label sequence based on k_1.

(2) In prediction phase, scarce similar days increase forecasting error rate dramatically. It is clear that a shorter query pattern will get more matching than a longer one. Generally, the forecast calculated on more similar days are better than fewer similar days. [Table 1] lists the MAPE of forecasts grouped by the number of similar days. It indicates that forecasts based on scarce similar days are susceptible to outliers.

3. THE PROPOSED METHOD

Since the electricity load shows clear seasonal cycles, to take the characteristic into account is a common practice on load forecasting issues [14]. Therefore, concerning on the PSF method, we propose to enhance the feature of weekly seasonality in clustering phase, extend similar days searching

Table 1. The MAPE of forecasts grouped by the number of similar days

Similar days	1	2	3	4	5	>5
MAPE(%)	6.97	5.784	5.569	5.200	4.944	4.207

to avoid matching insufficiency, and finally impose a filtering on similar days to discard the outliers.

3.1 Electricity Load Characteristics

Due to regular daily and weekly activities, electricity load shows a clear multiple seasonal cycles. [Fig. 3] is the load time series of New York City in years 2014 to 2016, got from the New York Independent System Operator [15]. Even though there are many spikes in the summer, it clearly shows yearly seasonality.

Another distinct feature of load is a big gap between weekdays and weekends. [Fig. 4] shows the average of daily load of New York city in 2016 group by day-of-week. There is no much difference among weekdays, but weekends are different definitely.

About the load forecasting issues, researchers have made good use of the characteristic to refine the model [16].

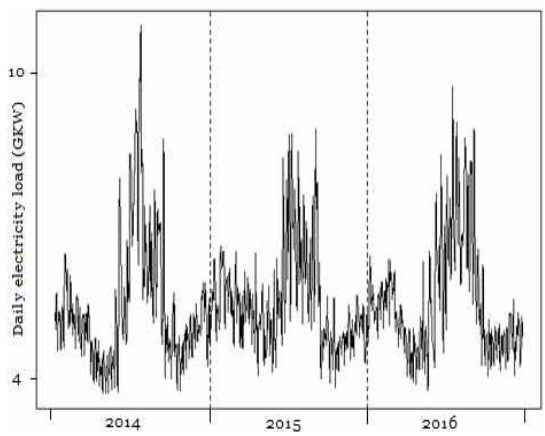

Fig. 3. The load time series of New York city in 2014–2016.

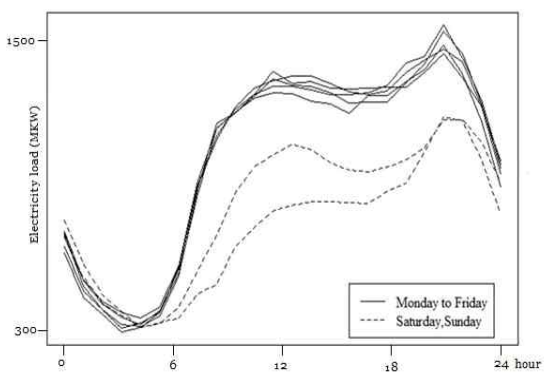

Fig. 4. The daily load of each day-of-week.

3.2 Methodology

With the purpose to improve the PSF method, we intend to obtain more similar days for forecasting when the similar days are insufficient. Meanwhile, impose a filter on the extended similar days to drop the outliers out.

3.2.1 The Diagram of Proposed Method

[Fig. 5] is the flowchart of the proposed method. Compared with the PSF method, two measures are taken. First, we introduce a parameter t as the threshold for the number of matching. If the number of matching is less than t, searching would restart with a short query pattern. Second, filter imposes on the match process to remove the outliers. Additionally, we propose a new index to evaluate the clustering performance and to determine the parameter k in generation of the label sequence.

3.2.2 Proposed Index for k Determination

The PSF method suggests three indexes (Davies Bouldin, Dunn, and Silhouette indexes) as indicators to select k. But the recommended indexes do not work properly sometime. By Dun index or Silhouette index, the higher the better.

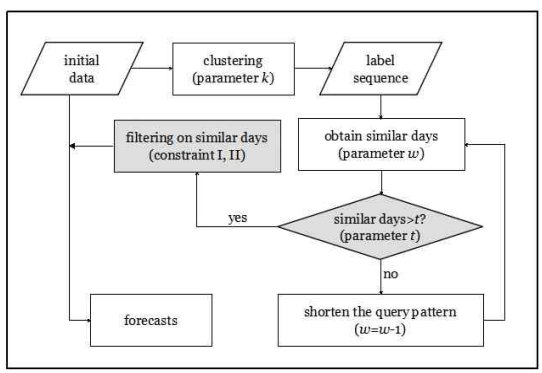

Fig. 5. The general flowchart of the proposed method.

Davies-Bouldin index is the opposite. [Table 2] shows index values, from k=4 to k=9 with k-means algorithm on the load time series of New York City in 2010-2015. According to each index, the number of clusters should be 5, 5 and 4. However, the 8 is better than the other values. The reason will be demonstrated in [Table 3, 4].

Since electricity load has regular weekly seasonality, similar day searching would benefit from an accurate classification of weekday and weekend. Therefore, in clustering phase of PSF algorithm, the effective clustering should differentiate weekdays and weekends as clearly as possible. So, we propose a new index to evaluate the effectiveness of clustering.

Let us consider S days clustered into k clusters, and for the cluster i (i=1, 2, ..., k), n_i and m_i denote the number of weekdays and weekends in the cluster, respectively. Then

$$S = \sum_{i=1}^{k} (n_i + m_i) \tag{1}$$

For a given cluster i, we use Eq. (2) as an indicator to measure the degree of differentiation of weekday and weekend, it is defined as:

$$Z_i = \left| \frac{2}{7} \cdot \frac{n_i}{n_i + m_i} - \frac{5}{7} \cdot \frac{m_i}{n_i + m_i} \right| = \frac{1}{7} \cdot \frac{|2n_i - 5m_i|}{n_i + m_i} \tag{2}$$

Then for all k clusters, we propose the clustering index as:

$$I = \sum_{i=1}^{k} wf_i * Z_i \tag{3}$$

where wf_i is the weighting factor of cluster i, that is $wf_i=(n_i+m_i)/S$.

From Eq. (2) and Eq. (3),

$$I = \frac{1}{7S} \sum_{i=1}^{k} |2n_i - 5m_i| \tag{4}$$

Therefore, according to the proposed index, the higher I value is better. It indicates the weekdays and weekends would be separated more clearly. With the proposed index, the appropriate number of clusters in above case should be 8 as shown in [Table 3].

To evaluate each value of parameter k, [Table 4] shows the MAPE under each given k, w pair. Apparently, the level of MAPE for column k=8 is

Table 2. The indexes of clustering with k = 4 to 9

k	4	5	6	7	8	9
Davies-Bouldin	0.6551	**0.4536**	1.0751	1.2382	0.7707	1.2188
Dunn	0.0226	**0.0323**	0.0299	0.0293	0.0306	0.0318
Silhouette	**0.5526**	0.4889	0.4342	0.3944	0.3692	0.3410

Table 3. The indexes of clustering with k = 4 to 9

k	3	4	5	6	7	8	9
Proposed index I	0.0273	0.0416	0.0559	0.0419	0.0448	**0.0605**	0.0445

Table 4. MAPE under each given k, w pair

MAPE(%)	k=4	k=5	k=6	k=7	k=8	k=9
w=3	14.189	6.549	6.477	6.092	5.778	5.675
w=4	8.044	6.411	6.558	6.314	5.794	5.793
w=5	8.068	6.376	6.665	6.310	5.659	5.764
w=6	7.874	6.511	6.571	6.310	**5.551**	5.719
w=7	7.947	6.509	6.568	6.367	5.597	5.711
w=8	8.176	6.125	6.713	6.384	5.656	5.742
w=9	8.936	6.518	6.233	6.391	5.687	5.765

lower than others, then in the case the proposed index indicates k correctly.

The proposed index is more reliable proved by several experiments, the reason is that the label sequence, which separates weekdays and weekends clearly, would contribute to an accurate matching for similar days.

3.2.3 Similar Days Filtering

As mentioned in section 2.3, another issue of the PSF method is that the scarce similar days increase forecasting error rate dramatically. The window size w impacts on the number of similar days directly, then using a shorter query pattern could definitely extend the similar days searching. However, the measure is a double-edged sword. It obtains more similar days, possibly it can include some outliers into account. So to cooperate with an extended similar days searching, we propose imposing a filtering on the similar days to drop the outliers out.

To filter the outliers out, two steps are needed:

Step 1. Pick holidays (except weekends) out before clustering and then insert a special identifier into the label sequence at the location of these holidays.

Step 2. Filter the similar days with one or two of the constraints below:

Constraint I. Both similar day and the day to be predicted are weekdays or weekends.

Constraint II. Both similar day and the day to be predicted are the same day-of-week.

Which constraint should be attached depends on the scale of data set, hence it should be chosen via a training process.

Although it can obtain the forecasts of holidays, conventionally holiday load forecasting is dealt with separately [17].

3.3 Comparison

The load time series was collected from New York Independent System Operator (NYISO). It is of New York city in 2010 to 2016. We took a three-year, four-year and five-year load time series as training dataset respectively. For each case, with n-year training data set, n-fold cross validation had been taken and the final result as shown in [Table 5].

The result shows the proposed method outperforms the PSF method in all cases, and it is about 15% improvement compared with the PSF method. As the training dataset increases, the proposed method will benefit from a longer label sequence more than PSF method does.

4. SUMMARY AND CONCLUSION

In summary, as an adaption of the PSF, we made two contributions. First, we introduced a new clustering index to determine the number of clusters. How to determine the number of clusters is still the crucial issue, because the most processes in the method totally depend on the label sequence. The proposed index has been proven to be more reliable for k selection. Second we imposed a threshold on the number of similar days to solve the problem caused by insufficient similar days. Compared to the PSF method, the label sequence by the pro-

Table 5. The comparison of the PSF and the proposed method

Training Dataset	PSF method		Proposed method	
	Parameters	Avg. MAPE(%)	Parameters	Avg. MAPE(%)
three-year	k=6, w=5	6.125	k=7, w=6, t=2	5.102
four-year	k=6, w=6	5.832	k=8, w=6, t=2	4.763
five-year	k=6, w=6	5.747	k=8, w=5, t=2	4.752

posed method showed that weekly seasonality was more clearly and similar days obtained with match filtering was more accurate. As shown in our experiment, our method reduced the mean absolute percentage error (MAPE) about 15%, compared to the PSF method.

REFERENCES

[1] N. Mohamed, M.H. Ahmad, Z. Ismail, and Suhartono, "Double Seasonal ARIMA Model for Forecasting Load Demand," *MATEMATIKA*, Vol. 26, No. 2, pp. 217-231, 2010.

[2] C.M. LEE and C.N. Ko, "Short-term Load Forecasting Using Lifting Scheme and ARIMA Models," *International Journal Archive: Expert Systems with Applications*, Vol. 38, No. 5, pp. 5902-5911, 2011.

[3] H.S. Migon and L.C. Alves, "Multivariate Dynamic Regression: Modeling and Forecasting for Intraday Electricity Load," *Applied Stochastic Models in Business and Industry*, Vol. 29, No. 6, pp. 579-598, 2013.

[4] A. Goia, C. May, and G. Fusai, "Functional Clustering and Linear Regression for Peak Load Forecasting," *International Journal of Forecasting*, Vol. 26, No. 4, pp. 700-711, 2010.

[5] P.S. Kalekar, "Time Series Forecasting Using Holt-Winters Exponential Smoothing," *Kanwal Rekhi School of Information Technology*, pp. 1-13, 2004.

[6] J.W. Taylor, "Short-Term Load Forecasting with Exponentially Weighted Methods," *IEEE Transactions on Power Systems*, Vol. 27, No. 1, pp. 458-464, 2012.

[7] F.J. Marin, F. Garcia-Lagos, G. Joya, and F. Sandoval, "Global Model for Short-term Load Forecasting Using Artificial Neural Networks," *IEEE Proceedings-Generation, Transmission and Distribution*, Vol. 149, Issue 2, pp. 121-125, 2002.

[8] F. Martinez-Alvarez, A. Troncoso, and J.C. Riquelme, "Energy Time Series Forecasting Based on Pattern Sequence Similarity," *IEEE Transaction on Knowledge and Data Engineering*, Vol. 23, No. 8, pp. 1230-1243, 2011.

[9] S.-H. Jung, "Clustering for Analysis of Raman Hyperspectral Dental Data," *Journal of Multimedia Society*, Vol. 16, No. 1, pp. 19-28, 2013.

[10] P.J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics*, Vol. 20, No. 4, pp. 53-65, 1987.

[11] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics,* Vol. 3, Issue 2, pp. 32-57, 1973.

[12] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, Issue 3, pp. 224-227, 1979.

[13] B. Desgraupes, *Clustering Indices*, University Paris Ouest Lab Modal'X, 2013.

[14] J.W. Taylor, "Triple Seasonal Methods for Short-term Electricity Demand Forecasting," *European Journal Operational Research*, Vol. 204, No. 1, pp. 139-152, 2010.

[15] The New York Independent System Operator, http://www.nyiso.com (accessed Jan., 13, 2017).

[16] S.L. Zhu, J.Z. Wang, W.G. Zhao, and J.J. Wang, "A Seasonal Hybrid Procedure for Electricity Demand Forecasting in China," *Applied Energy*, Vol. 88, No. 11, pp. 3807-3815, 2011.

[17] Y.M. Wi, S.K. Joo, and K.B Song, "Holiday Load Forecasting Using Fuzzy Polynomial Regression With Weather Feature Selection and Adjustment," *IEEE Transaction on Power Systems*, Vol. 27, No. 2, pp. 596-603, 2012.

Fazheng Chu

He received the B.S. and M.S. degrees in Economics from Qingdao University, China in 1999 and 2004, respectively. He is a lecturer in Economics and Management College of Qingdao Agricultural University, China since 2004. He is currently doing a Ph. D. course in Department of Computer Engineering, Changwon National University. His research interests include econometric modelling and forecasting, and image processing.

Sung-Hwan Jung

He received the B.S., M.S., and Ph. D. degrees from in Electronic Engineering (information and communication major) from Kyungpook National University, Korea in 1979, 1983, and 1988, respectively. He had worked for the Electronic and Telecommunication Research Institute in Korea as a research staff, where he had experienced some national research projects including developing a portable computer.

In 1988, he joined the faculty of the Department of Computer Engineering at Changwon National University in Korea, where he is currently working as a full professor. From 1992 to 1994, he was a post-doctoral research staff of the Department of Electrical and Computer Engineering at the University of California at Santa Barbara (UCSB). From 1999 to 2000, he also worked for the Colorado School of Mine (CSM) in Golden, Colorado as an exchange professor. From 2008 to 2009, he had experience on the medical information processing at the Dental School of the University of Missouri at Kansas City (UMKC), as a visiting professor. He is an Information System Auditor and P.E. in the area of Information Processing and Electronic Computer.

His research interests include content-based image retrieval, steganography, watermarking, medical image processing, computer vision and pattern recognition, etc. He is a co-author of many image processing related books including "Visual C++ Digital Image Processing Using Open Source CxImage," "Practical Computer Vision Programming Using VC++ and OpenCV," and Image Processing and Its Application with OpenCV."