Short-Term Electric Load Forecasting Based on Data Mining

Han Wu, Dongxiao Niu, and Zongyun Song

***Abstract*—Short-term electric load forecasting is significant for safe and economic operation in power system. In order to improve the accuracy of predicted results, this paper proposes a PSO-SVM model, which is based on cluster analysis techniques and data accumulation pretreatment. Specifically, K-medoids cluster analysis is employed to extract similar days, then original data is divided into k category clusters. In addition, this paper carries on the accumulation of the original data to weaken the effect of the irregular disturbance and enhance the regularity of the sequence. And we use PSO to optimize the parameters of SVM system. In case analysis, the results show that the method proposed in this paper can effectively promote the forecasting performance.**

***Index Terms*—Short-term power load forecasting, data mining, K-medoids (KM), partial swam algorithm (PSO), support vector machine (SVM).**

I. Introduction

In current rapid growth of power industry, developmental level of national economy is closely related to the advantages and disadvantages of the power load forecasting technology. Accurate load forecasting is conducive to the power grid enterprises to set out scientific and reasonable power delivering, deploying plan, to avoid risks effectively, and to ensure the safety and reliability of power supply.

Short-term power load forecasting, which is based on historical load data, taking the impacts of factors such as weather, holidays and economy into consideration, masters load's fluctuation and its internal relation to other factors. Then appropriate prediction technique and mathematical methods are utilized to infer future growing trend of load within a few hours or a day. Short-term power load forecasting needs a large quantity of data and has certain randomness due to noise and other factors. The key to improving accuracy is how to deal with the data scientifically and reasonably. Data mining can extract implied but potential information through enormous, incomplete, noise, blur and random data. Clustering is a kind of data mining technique. Cluster analysis adopts several similar measurements to divide data into a set of subsets. The data has similar nature in each collection, but differs a lot between different collections. The trend of daily load is very similar and fluctuates heavily owing to the type of day, weather, holidays and other factors. The daily load curves of same year and same quarter are similar, of same month in different years are close. Therefore, in order to improve accuracy, this paper takes K-medoids cluster analysis technique to designate similar days as a class.

Manuscript received January 26, 2016; revised May 1, 2016.

The authors are with the North China Electric Power University, Beijing, China (e-mail: wuhan930108@163.com, niudx@126.com, 961056925@qq.com).

Common short term load forecasting methods generally have two categories, classic predicting methods and intelligent ones. The first category includes regression analysis, time series, and grey forecasting method, etc. Regression analysis, whose principle and form are simple, has quick prediction speed and distinct extrapolation performance. But it asks for strict historical data and could not describe a complex problem. Time series analysis [1] needs a small amount of data and calculates rapidly. But it ignores the influential factors. Grey foresting method, relying on less data, can weaken the interference of random factors on the load changes. High precision can be achieved through simple calculation and is easy to be checked. However, this method is only applied to load with exponential trend [2]. As to the second category, there are neural network methods, SVM and so on. Neural network obtains prediction results by self-learning on historical data, but the layers of neural network and the number of neurons are determined with subjective experience. Besides, its convergence rate is slow and the results are vulnerable to fall into local minimum [3].The basic theory of SVM is VC dimension and structural risk minimization principle, according to the limited sample information to find the best compromise between complexity and learning ability of the model, so as to get the best promotion ability. SVM can convert the practical problems to high-dimensional feature space through nonlinear transformation, and deal with dimension problems by means of kernel function to get global optimal solution. It also has fast convergence speed and high accuracy [4].

To improve precision of short term load forecasting, this paper puts forward a PSO-SVM (accumulative PSO-SVM, APSO-SVM) forecasting model based on pretreatment data accumulation, which combines optimizing performance of particle swarm algorithm and predict advantages of SVM. Moreover, this model applies accumulation to weaken the irregular disturbance and enhance the regularity of the sequence.

II. Basis of Model

A. K-medoids Clustering Algorithm

K-medoids algorithm, with strong robustness and high accuracy, is a clustering algorithm based on partition [5]. The theory of traditional K-medoids algorithm is as follows. First of all, selecting a representative object for each cluster randomly and each remaining object is clustered with the representative object to the nearest. Then the non-representative object is substituted for the original representative one repeatedly to improve the quality of clustering. The quality of clustering is estimated by a cost function, which can measure the average dissimilarity between the object and its referents [6]. The process is as

DOI: 10.7763/IJET.2017.V9.979

follows [7], [8]:

1) Arbitrarily choose K representative objects in a data set containing M objects as the cluster centers;
2) Calculate the distance from non-representative objects to each cluster center, and assign them to the nearest one;
3) When the distribution is completed, select a data in sequence to replace the original cluster center, update each cluster until quadratic difference (E) is minimum. Quadratic difference is defined as:

$$E=\sum_{j=1}^{k}\sum_{p\in C_i}\left|p-O_j\right|^2 \tag{1}$$

p is the sample in cluster C_i , O_j is cluster center.

4) Compare with the previous clustering, if the quality changes then go to step (2) and vice versa step (5);
5) Output the clustering results.

B. APSO-SVM

The basic idea of APSO-SVM is accumulating the original data to obtain training samples firstly; then train the sample data by inputting them into SVM, exert PSO to optimize SVM parameters (penalty factor C , the width of kernel function σ , insensitive loss function ε) and establish the forecasting model [9]. It is necessary to take inverse accumulated generating operation to get the final results. Detailed steps are as follows:

1) Data preprocessing

The original sequence is $\left\{x^{(0)}(1),x^{(0)}(2),\ldots x^{(0)}(n)\right\}$, one-accumulated sequence is $\left\{x^{(1)}(1),x^{(1)}(2),\ldots x^{(1)}(n)\right\}$,

$$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i),k=1,2,\ldots,n \tag{2}$$

Because SVM is sensitive to data between zero and one and its training speed is fast, we need to normalize the accumulated data, equation is shown as below:

$$x^{(1)}(i)'=\frac{x^{(1)}(i)-x^{(1)}(min)}{x^{(1)}(max)-x^{(1)}(min)} \tag{3}$$

2) Parameter optimization[10]

1) Set the initial value of particle swarm. The maximum iteration is $T_{\max}$, inertia weight is w , c_1 and c_2 are acceleration coefficients. Assume that the iteration time is t , the initial particle swarm is $X(t)$, $x_1,x_2,\ldots,x_s$ represent particles generating from the swarm randomly, the corresponding particle velocity is $v_1,v_2,\ldots,v_s$, the swarm velocity is $V(t)$.
2) Assess the particle swarm. The fitness function is defined as $-F=-\sum_{i=1}^{n}(y_i-y_i)^2$, y_i is the sample and y_i is the output value.
3) After finding the two best values, the particle updates its velocity and positions with following equations.

$$\begin{aligned}v_{id}(t+1)&=wv_{id}(t)+c_1r_1(p_{id}(t)-x_{id}(t))\\&+c_2r_2(p_{gd}(t)-x_{id}(t))\quad(d=1,2,\ldots,n;i=1,2,\ldots,S)\end{aligned} \tag{4}$$

$$x_{id}(t+1)=v_{id}(t)+v_{id}(t+1) \tag{5}$$

S is population size; r_1,r_2 is random number between [0,1] which dominates the effects that front speed has made on current speed; p_{id} is the individual optimal solution pbest; p_{gd} is the globally optimal solution gbest.

1) Detect the loop-termination criteria. The iteration terminates when one of the following conditions is met: ① iteration is greater than the maximum iteration; ②the precision value is less than specification error. If neither can be satisfied, then go back to step two.
2) Extract the optimal parameters and assign them to SVM;
3) Train SVM with sample data and solve the model by the improved minimal sequence value method, the process is expressed as Eq. (6).

$$\begin{aligned}&\max_{\alpha_i^*}W(\alpha_i,\alpha_i^*)=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\alpha_i^*)\cdot(\alpha_j-\alpha_j^*)\cdot\\&K(x_i,x_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\\&subject\ to\begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\\alpha_i-\alpha_i^*\in[0,C]\end{cases}\end{aligned} \tag{6}$$

α_i,α_i^* is Lagrange multiplier and meets the equation: $\alpha_i\times\alpha_i^*=0$, $\alpha_i,\alpha_i^*\geq 0$; $K(x_i,x_j)$ is kernel function.

4) Plug the aforementioned parameters into Equation (7) and confirm the final functional equation.

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{7}$$

III. Instance Analysis

A. Data Collection

Collect a grid's load data and meteorological information from June 1, 2014 to August 31, 2014, totally 92 days. The statistics is shown in Table I.

B. Cluster Analysis on Short-Term Load

We sort the collected data to form a cluster sample. The category number is three, the daily average temperature, type and maximum load seven days ago are selected as related factors that constitute the sample's feature vectors. Cluster analysis on sample is conducted by using K-medoids algorithm. Days are divided into four types: Monday to Thursday, Friday, Saturday and Sunday. They are mapped to a specific range by dimensionless. For the purpose of quantitative, 1 represents for Monday to Thursday, 2 for Friday, 3 Saturday, 4 for Sunday.

Select similar days based on cluster analysis and extract their daily average temperature, type and maximum load seven days ago as the SVM's input vector, then output the

load forecasting value. According to the clustering results, we establish training and testing samples of SVM separately. The data in training samples is from June and July while the testing is from August, statistical results are shown in Table II.

In cluster 1, historical data from June and July is used to train SVM's parameters σ、C and ε, input testing samples' eigenvector to obtain the forecasting results and calculate the prediction accuracy. Cluster 2 has 27 samples (24 training samples, 3 testing samples), cluster 3 has 32 samples (21 training samples, 11 testing samples). Their forecasting value can be computed in a similar way.

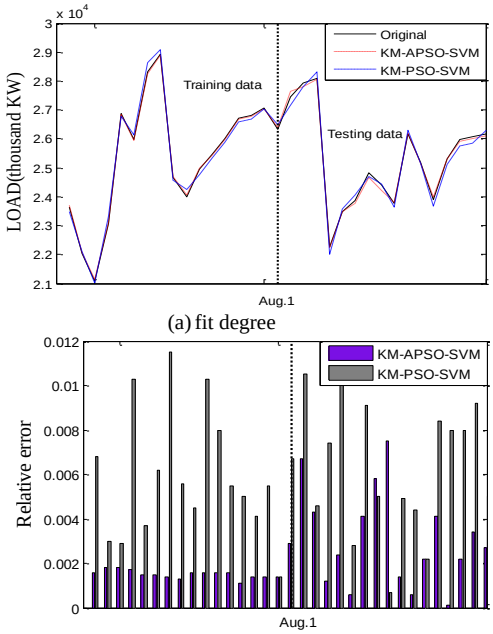
TABLE I: LOAD DATA STATISTICS (UNITS: MW)

Time	Monthly maximum load	Monthly minimum load	Monthly mean load
June	26057.21	13448.77	19535.57
July	28925.94	14604.94	21545.61
August	28109.74	14308.91	20743.61

TABLE II: TRAINING SAMPLE AND TESTING SAMPLE

Clusters	Training set		Testing set	
cluster1	June	6th, 19th to 20th	August	1st to 11th, 15th, 18th to 19th, 22nd, 26th to 27th
	July	10th, 14th to 15th, 20th to 28th, 31st		
cluster2	June	2nd to 5th, 17th to 18th, 26th, 30th	August	16th, 20th to 21st
	July	1st to 7th, 11th to 13th, 16th to 19th, 29th, 30th		
cluster3	June	1st, 7th to 16th, 21st to 25th, 27th to 29th	August	12th to 14th, 17th, 23rd to 25th, 28th to 31st
	July	8th to 9th		

C. Load Forecasting Results and Discussion



(a) fit degree

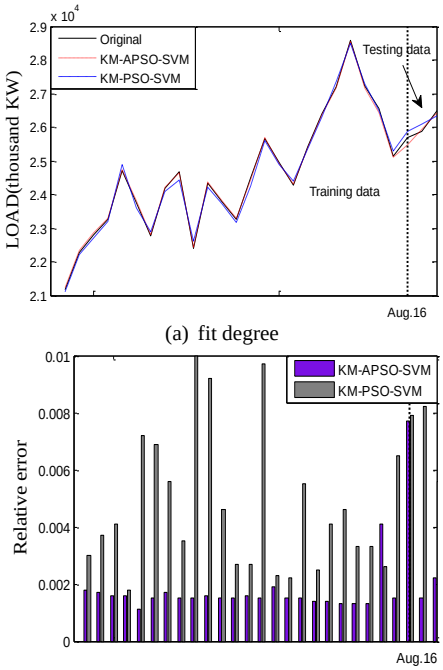
(b) relative error

Fig. 1. The prediction results of cluster 1.

Carry out cluster analysis on similar days based on K-medoids algorithm, cluster center point K=3. Choose the type of day, daily average temperature and maximum load seven days ago as the input vector. Get the final clustering results, namely, cluster 1, cluster 2 and cluster 3. Input each cluster's original vector stemming from June and July to train APSO-SVM model. The whole load forecasting value would be achieved once each cluster's prediction results are acquired.

Fig. 1, Fig. 2 and Fig. 3 represent the prediction results of cluster 1, cluster 2 and cluster 3 separately. Through observing the fit degree and relative error, we can find that KM-APSO-SVM fits the original load better than the KM-PSO-SVM model. That is to say, APSO, with respect to PSO, can optimize SVM parameters better and improve accuracy.

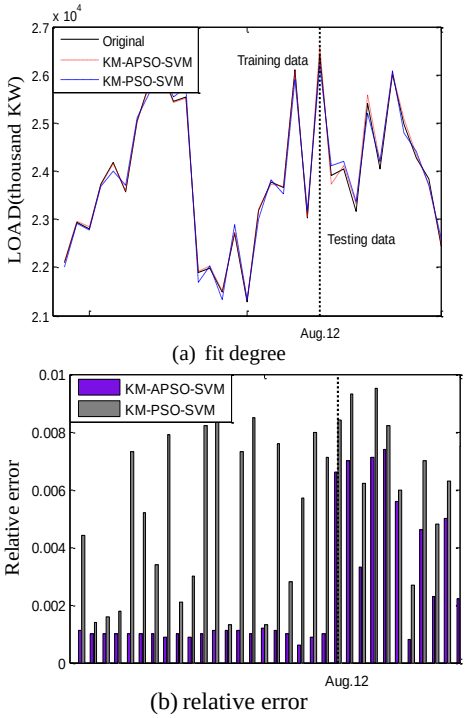
Integrate the results of cluster 1 and cluster 2, cluster 3 to get the whole load forecasting value in August. The final results are shown in Fig. 4.



(a) fit degree

(b) relative error

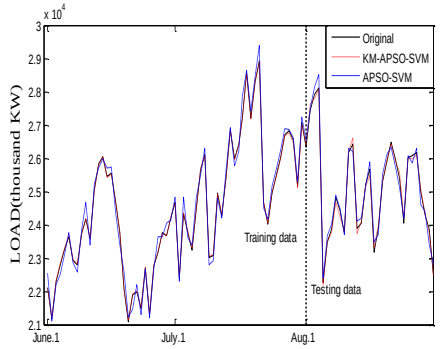
Fig. 2. The prediction results of cluster 2.



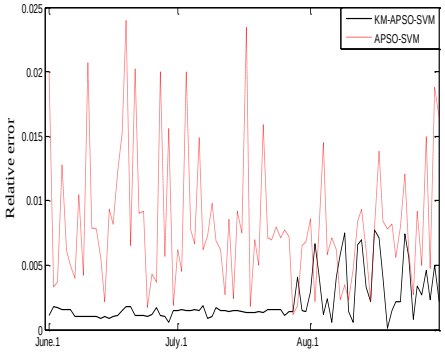
(a) fit degree

(b) relative error

Fig. 3. The prediction results of cluster 3.



(a) fit degree



(b) relative error

Fig. 4 Forecasting value in August.

TABLE III: EACH MODEL'S MAPE VALUE

	KM-APSO-SVM	APSO-SVM	KM-PSO-SVM	SVM	BP	ARIMA
MAPE	0.22%	0.83%	0.56%	1.35%	1.60%	2.44%

Contrasting KM-APSO-SVM results with APSO-SVM results, it is obvious that KM-APSO-SVM has a smaller relative error and higher fit degree, illustrating that clustering can improve accuracy significantly.

In order to further verify the validity of KM-APSO-SVM, we use SVM, BP neural network and ARIMA to forecast. Comparing each model's mean absolute percent error (MAPE), results are shown in Table III.

We can come to the conclusion easily from Table III that MAPE value of SVM model, BP neural network and ARIMA model is greater than 1%, far greater than the improved SVM model. The MAPE of KM-APSO-SVM is the smallest, merely 0.2172%, less than the error of KM-PSO-SVM and APOS-SVM. Table III further demonstrates that accumulative PSO-SVM model based on cluster analysis has a better goodness-of-fit and higher forecasting accuracy.

IV. CONCLUSIONS

This paper combines K-medoids of cluster analysis in data mining with SVM prediction method, proposing a KM-APSO-SVM prediction model based on cluster analysis and pretreatment of data accumulation. KM-APSO-SVM prediction model sorts out similar days by means of K-medoids and divides the original series into three clusters. Combining optimizing performance of particle swarm algorithm and predict advantages of SVM, we forecast every cluster separately and get the results. The model is applied to predict a grid's load in China. The results indicate that the accuracy of KM-APSO-SVM is clearly higher than that of SVM model without cluster analysis, SVM model without data accumulation and other classic prediction models. In general, the proposed short-term load forecasting method based on cluster analysis greatly improves forecasting accuracy and is feasible in practical applications.

REFERENCES

[1] X. Ma, A. A. El-Keib, H. Ma *et al*., "Short-term load forecasting," in *Proc. the IEEE*, December 1987, vol. 75, pp. 1558-1573.
[2] K. H. Kim, H. S. Youn, and Y. C. Kang, "Short-term load forecasting for special days in anomalous load conditions using neural networks and fuzzy inference method," *IEEE Trans Power Syst.*, vol. 15, pp. 559-565, February 2000.
[3] "Application of artificial neural networks in global climate change and ecological research: An overview," Chinese *Science Bulletin*, vol. 34, pp. 3853-3863, January 2010.
[4] S. J. Kiartzis, A. G. Bakirtzis, and V. Petridis, "Short-term load forecasting using neural networks," *Electric Power Systems Research*, vol. 33, pp. 1–6, January 1995.
[5] F. Informatik, L. Viii, K. Intelligenz *et al*., "Making large-scale SVM learning practical," *Technical Reports*, vol. 8, pp. 499-526, March 1998.
[6] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 17, no. 1, pp. 113-26, 2004.
[7] A. P. Reynolds, G. Richards, and V. J. Rayward-Smith, "The application of K-Medoids and PAM to the clustering of rules," *Lecture Notes in Computer Science*, vol. 3177, pp. 173-178, 2004.
[8] W. Sheng and X. Liu, "A genetic K-medoids clustering algorithm," *Journal of Heuristics*, vol. 12, pp. 447-466, June 2006
[9] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya *et al*., "Improvements to the SMO algorithm for SVM regression," *IEEE Transactions on Neural Networks*, vol. 11, pp. 1188-1193, May 2000.
[10] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, pp. 3336-3341, February 2009.

Han Wu was born in Nantong, Jiangsu, in 1993. She is a postgraduate student at School of Economics and Management of North China Electric Power University in China and mainly works on the short-term electric load forecasting.

She has researched in short-term electric load forecasting for one year. She also participated in the evaluation of wind power project in 2015. During college, she awarded numerous scholarships.

Dongxiao Niu was born in Suxian, Anhui, in 1962. He received his Ph. D. degree in technology economy and management from North China Electric Power University in 2002 and is currently a professor, doctoral supervisor at North China Electric Power University. He is majoring in the research of project forecasting and decision theory and its application, project comprehensive evaluation method and its application.

He has engaged in education for many years and cultivated a lot of talent.

Prof. Niu has achieved the Yangtze River scholar and published many articles in some international journals, such as SCI and EI.

Zongyun Song was born in Shandong, in 1991. She is a postgraduate student at School of Economics and Management of North China Electric Power University in China and mainly works on the short-term electric load forecasting.

She has researched in short-term electric load forecasting for three years. She also participated in the evaluation of wind power projects in 2015. During college and postgraduate, she awarded numerous scholarships.