A Research on Power Load Forecasting Model Based on Data Mining

Fuyu Sun[1] and Yunshi Yang[2]

[1] Department of Physics & Electronic Information Engineering, ChiFeng College, ChiFeng 024001, Inner Mongolia, P.R. China sfyzi@163.com
[2] College of Computing & Communication Engineering, Graduate University of the Chinese Academy of Sciences, Beijing 100049, P.R. China yysh@court.gov.cn

Abstract. Utilizing the advantage of data mining technology in processing large data and eliminating redundant information, the system mines the historical daily loading which has the same meteorological category as the forecasting day in order to compose data sequence with highly similar meteorological features. With this method it can decrease SVM training data and overcome the disadvantage of very large data and slow processing speed when constructing SVM model. Comparing with single SVM and BP neural network in short-term load forecasting, this new method can achieve greater forecasting accuracy. It is denoted that the SVM learning system has advantage when the information preprocessing is based on data mining technology.

Keywords: *Data mining, Meteorological factor, Power system, Support vector machines. Short-term load forecasting*

I. INTRODUCTION

The short-term power load forecasting is very significant to electric network's reliable and economic running. It can plan the open or stop of generators and keep the electric network running safely and stably with the exact load prediction. With the development of electric market, people have been paying more and more attention to load prediction. How to make the prediction on the short-term power load exactly becomes a hot point [1].

In the short load forecasting with 96 points, the short load forecasting be affected by many random interference factors and it need several kinds of forecasting model to fit, these are the most difficulty ones which affect the accuracy of forecasting. One factor of these is the weather, it cause the history short loading data which be arranged on date to be affected by noise and breakdowns the intrinsic variable rule. If it doesn't consider the factor of weather, the percentage error would become wider no matter it uses any model, this will influence the stable operation of the power network.

At present, researchers are using the technology with abnormal value processing technique or different smooth method to force the sequence of history data to consume to a new one of smaller amplitude. Using this method, they build a neural

Please use the following format when citing this chapter:

Sun, F., Yang, Y., 2007, in IFIP International Federation for Information Processing, Volume 255, Research and Practical Issues of Enterprise Information Systems II Volume 2, eds. L. Xu, Tjoa A., Chaudhry S. (Boston: Springer), pp. 1369-1377.

net model to forecast. However, there are some disadvantages with this method: it can't show the change of the short load forecasting which has feature with specific weather; Meanwhile, some new problems are happened when use the method artificial neural net to forecast. For example: BP net emphasizes to overcome the leaning excessively and has weak generalization performance, it is difficult to define the number of implicit unit, so the final weight will be more influenced by the initial data, and so on.

Support Vector Machines (SVM) is based on the statistic theory which is raised by Vapnik. It is a machine learning algorithm which began to be used in the middle of ninetieth century. Statistic theory employs the criterion that minimizes the structure risk, meanwhile it can low the global error of model. It raises the generalization capability of the model, which is more prominent in the small-sample learning.

Through above analysis, a new thought which can improve the accuracy of load forecasting be presented, it believes the key which can improve the accuracy are the preprocessing of the history data and the improved forecasting model, so it present a new method which is Support Vector Machines based on data mining technology in power load forecasting model. First it comprehend the weather character of forecasting data through weather forecasting; Second it takes advantage of the merit in disposing large data size and erasing redundant information to search a number of history load data which similar with forecasting date in weather character. Basing on the correlation analysis technology extract the data and compose data sequence which has very similar weather characters, it can decrease the training data of SVM. At last, build the forecasting model with these data. From the actual study, forecasting accuracy is higher than BP neural network and single SVM.

2. DATA MINING PREPROCESSING

2.1 Data Mining Introduction

Data mining technology is a process to search new relation or pattern and trend with pattern recognition technology, statistic and mathematic technology that is to mine the information which has potential value. It has function as following:

(1)Classification: It can build variant categories to describe things according to attribute and character of the object. The data which category identification sample event belongs to be are contained in sample data.

(2)Clustering: Recognize intrinsic regulation of data analysis, and classify the factors into variant categories according to these regulations.

(3)Association: Search the relative events or records, deduce the potential relations among the events, and recognize the pattern which probable occur repeatedly. It can collate according to the extent of the relation.

(4)Sequence pattern: Like association analysis, it expands correlation analysis to the relation among item set in a period of time, and it believes the sequence pattern to be the relation of time variable.

(5)Forecasting: Analyze law of development of objects and forecast coming trend.

2.2 Data Mining Preprocessing Technology

Fuzzy Classification **Disposing Weather Factors.**

At present, as to short load curve, weather is a very important influencing factor. It must consider influencing factors in order to improve accuracy of short load forecasting. In order to obtain the relation between weather factors and load swing, it needs to draw some chats of points (One can be got from Fig.1). A conclusion can be known through these chats: The factors which have more influence to loading and can be acquired from weather forecasting are daily highest temperature, daily lowest temperature, daily average temperature and daily rainfall [2-5].

These four factors need to be assigned and classified fuzzily, the result is represented by vector (Z_1, Z_2, Z_3, Z_4), Z_1, Z_2, Z_3 be classified to be lowest, middle separately, highest, they are assigned to be 1, 2, 3; Z_4 be classified to be no rainless, light rain, moderate rain, heavy rain, and assigned to be 0, 1, 2, 3 separately. The history daily loading can be classified as follows:

$$(Z_{1,}Z_{2,}Z_{3,}Z_4)=\begin{cases} Z_1=1,2,3 \\ Z_2=1,2,3 \\ Z_3=1,2,3 \\ Z_4=1,2,3 \end{cases} \tag{1}$$

According to this method, the historical daily data would be recorded in this way and construct a data bank. It would become very easy to search the useful data.

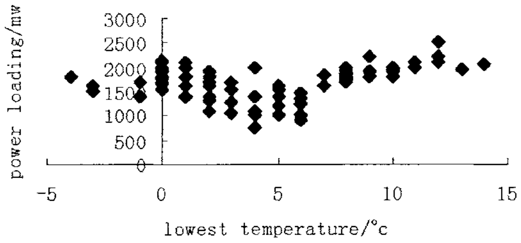

Figure 1. Diagram of Relation Scatter Points Between the Lowest Temperature and Power Loading

Select Daily Data Based on Gray Relation Analysis Method for Forecasting.

1. Gray relation analysis theory. Relation analysis is a method which analyzes the extent of relation among factors in gray system theory; the essence is to judge the extent of relation of the factors according to the relation of caves.

Detailed processes are as following:

(1) Construct sequence matrix. It needs to do relation taxis analysis when it has constructed data classification bank. Hypothesize represent the weather character of forecasting date, if the weather forecasting reports the highest temperature is 40°C, the average temperature is 25°C, the lowest temperature is 15°, the rainfall is 20mm, then= $T_0=(T_0(1),T_0(2),T_0(3),T_0(4))=(40,25,15,20)$. In this way, constructs comparing sequence with daily weather data of the obtained data bank, they are represented as $T_1,T_2,\cdots,T_n$

$$(T_0,T_1,T_2,\ldots T_n)=\begin{bmatrix} T_0(1) & T_1(1)\ldots T_n(1) & \\ | & | \quad | \quad | & \\ T_0(m) & T_1(m)\ldots T_n(m) & \end{bmatrix} \tag{2}$$

(2) Nondimension: Processing data with the method of initiation to eliminate dimension. The formulas are:

$$T_i^{'}(k)=\frac{T_i(k)}{T_i(l)}, i=0,1,2\ldots,n;\ k=1,2,\ldots,m \tag{3}$$

Nondimension matrix:

$$(T_0^{'},T_1^{'},T_2^{'},\ldots T_n^{'})=\begin{bmatrix} T_0^{'}(1) & T_1^{'}(1)\ldots T_n^{'}(1) & \\ | & | \quad | \quad | & \\ T_0^{'}(m) & T_1^{'}(m)\ldots T_n^{'}(m) & \end{bmatrix} \tag{4}$$

(3) Calculate relation coefficient:

$$\xi_{0i}(k)=\frac{\min\min|x_0(k)-x_i(k)|+\rho\max\max|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\max\max|x_0(k)-x_i(k)|} \tag{5}$$

In this formula, , $i=0,1,2\ldots,n; k=1,2,\ldots m$, ρ is resolution ratio, $\rho\in[0,1]$, Generally $\rho=0.5$, then, the relation coefficient matrix can be obtained:

$$\begin{bmatrix} \xi_{01}(1) & \ldots & \xi_{0n}(1) & \\ | & | & | & \\ \xi_{01}(m) & \ldots & \xi_{01}(m) & \end{bmatrix} \tag{6}$$

(4) Calculate associated degree:

$$r_{0i}=\frac{1}{m}\sum_{k=1}^{m}\xi_i(k), i=1,2,\cdots,n \tag{7}$$

2. Ascertain history loading sequence for forecasting.

In this article, the reference sequence is the meteorologic factor index vector of forecasting date; the comparing sequence is the meteorologic factor index vector of history dates which are similar with the forecasting date in weather T_t. Then calculate the associated degree between T_0 and T_t is r_t. Set a threshold value α , choose those dates whose associated degree $r_t \geq \alpha$. Then collate these loading dates orderly to be a new sequence.

3. SVM REGRESSION THEORY INTRODUCTION [6-8]

Suppose a set of data (x_i, y_i), $i = 1,2 \cdots n$, $x_i \in R^n$ are given as input , $y_i \in R$ are the corresponding output. SVM regression theory is to find a nonlinear map from input space to output space and map the data to a higher dimensional feature space through the map, then the following estimate function is used to make linear regression.

$$f(x) = [\omega \bullet \phi(x)] + b \quad \phi : R^m \to F, \ \omega \in F \tag{8}$$

b is the threshold value. The problem of the function approximate is equivalent with the minimizing the following problem.

$$R_{reg}[f] = R_{emp}[f] + \lambda \|\omega\|^2 = \sum_{i=1}^{s} C(e_i) + \lambda \|\omega\|^2 \tag{9}$$

$R_{reg}[f]$ is objective function and s is the number of the sample. $e(\bullet)$ is loss function and λ is adjusting constant meter. The following loss function can be gained concerning with the rarefaction character of the linear insensitive loss function ε .

$$|y - f(x)_\varepsilon| = \max\{0, |y - f(x) - \varepsilon|\} \tag{10}$$

Empirical risk function is:

$$R_{emp}^{\varepsilon}[f] = \frac{1}{n} \sum_{i=1}^{n} |y - f(x)|_\varepsilon \tag{11}$$

According to statistic theory, the regression function is determined by minimizing the following functions.

$$\min\left\{\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i^* + \xi_i)\right\} \tag{12}$$

$$y_i - (\omega \bullet \phi(x)) - b \leq \varepsilon + \xi_i^* \tag{13}$$

$$(\omega \bullet \phi(x)) + b - y_i \leq \varepsilon + \xi_i \tag{14}$$

$$\xi_i, \xi_i^* \geq 0 \tag{15}$$

C is used to equalize the complicated item of the model and the parameters of the item of training error. ξ_i^* and ξ_i are relaxation factors and ε is insensitive loss function. The problem can be converted into the dual problem:

$$\max[-\frac{1}{2}\sum_{i,j=1}^{n}(a_i^* - a_i)(a_j^* - a_j)K(X_i, X_j) + \sum_{i}^{l} a_i^*(y_i - \varepsilon) - \sum_{i=1}^{n} a_i(y_i - \varepsilon)] \tag{16}$$

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} a_i^* \tag{17}$$

$$0 \leq a_i^* \leq C \tag{18}$$

$$0 \leq a_i \leq C \tag{19}$$

Solve the problem, then the regression equation is:

$$f(x) = \sum_{i=1}^{n}(a_i - a_i^*)K(X_i, X) + b \tag{20}$$

4. SVM BASED ON DATA MINING

4.1 Process for Forecasting

1. Using the combined data mining technology which introduced above to input the history data and preprocess, get the training and testing sample bank which has the most similar weather character [9].

2. Using gray relation analysis method to construct the needed history loading sequence.

3. Parameters in SVM model should be initialized. $a_i a_i^*$ and *b*are assigned random values.

4. The objective functions like (12)-(15) are established by using training sample. Then previous functions are converted into their dual problems like (16)-(19), thus $a_i a_i^*$ and will be worked out. And then they are substituted into (20), by this way the predicted value of the subsequent time points will be worked out b [10].

5. APPLICATION AND ANALYSIS

Power loading data in Inner Mongolia region is used to prove the effectiveness of the model. Because the variation regulation of power loading in that area is influenced explicitly by the weather and the temperature of seasonal variation is very obvious, year temperature difference is very big, rainfall is relative centralized and agriculture loading occupies a big proportion, so it is very appropriate for this method. Comparing with the single SVM model and Neural Net Work which doesn't consider the weather factors, this new model has higher accuracy.

5.1 Sample Choosing

Some data are chosen from the data bank of Neimengu region. The power load data from 0:00 at 5/1/2002 to 12:00 at 6/5/2004 are as training sample and used to establish the single-variable time series. And the power load data from 13:00 at 6/5/2004 to 24:00 at 6/5/2005 as testing sample.

1. $A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$, Forecasting time-interval loading data of n days before the forecasting date;

2. $A=\{b_1,b_2,\cdots,b_n\}$, Loading data of m time-intervals before the forecasting date;

3. $C=\{c_1,c_2,\cdots,c_n\}$, There are s groups of weather forecasting data which contain the average temperature, the highest temperature, the lowest temperature and daily rainfall;

4. $D=\{d_1,d_2,\cdots,d_n\}$, The daily weather data of n days before the forecasting date. Meanwhile, every factor d_i contains s groups of data which is same as above.

Table 1. Comparison of Forecasting Error of 21:00 from June 5, 2005 to June 16, 2005

Date	Actual Load	DMSVM	SVM	BP
2005-6-5	628.73	2.85%	6.01%	6.98%
2005-6-6	626.88	1.69%	1.34%	1.45%
2005-6-7	639.29	1.61%	1.73%	2.95%
2005-6-8	670.06	3.84%	4.91%	4.32%
2005-6-9	716.74	1.60%	1.37%	1.70%
2005-6-10	657.63	3.63%	3.18%	4.03%
2005-6-11	772.16	2.55%	2.01%	2.27%
2005-6-12	655.07	2.29%	5.81%	5.69%
2005-6-13	644.87	1.93%	1.78%	2.27%
2005-6-14	628.18	1.36%	3.00%	4.56%
2005-6-15	651.19	2.01%	2.69%	3.23%
2005-6-16	610.57	3.31%	3.58%	3.81%

RMSRE		2.67%	3.50%	3.93%

5.2 Error Analysis

Relative error and root-mean-square relative error are used as the final evaluating indicators:

$$e = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A(i)-F(i)}{A(i)}\right| \times 100\% \tag{21}$$

$$RMSRE = \sqrt{\frac{1}{N-n_{tr}}\sum_{t=n_{tr}}^{n}\left(\frac{x_t - y_t}{x_t}\right)} \tag{22}$$

5.3 SVM Forecasting

SVM is used to make prediction after the samples are normalized [11-13]. Libsvm toolbox is used to compute the results and radial basis function is chosen as the kernel function. The parameters are chosen as the following: C=79.31, ε=0.012, σ^2=4.28

Single SVM which hasn't been data mined, The parameters are chosen as the following: C=26.30, ε =0.006, σ^2 =1.60

BP algorithm is used to make prediction with sigmoid function. The parameters are chosen as the following: the node number of input layer is 11 and the node number of output layer is 1. The node number of interlayer is 8 according to the experience. The system error is 0.001 and the maximal interactive time is 5000.

A fixed point of twelve days data from 6/5/2005 to 6/16/2005 is forecasted separately. The results of three methods are shown in table 1.

6. CONCLUSIONS

The results show that SVM based on data mining has great effectiveness for short-term power load forecasting. And the conclusions are shown as the following:

1. It is necessary to preprocess the data which is influenced very much by uncertain factors for short-term power load forecasting. The past loading data can be classified to different group according to different weather character and choosing finite sample points which have same character to forecast. Then SVM prediction model is established to make prediction. The real load data prediction shows that the model is effective in short-term power load forecasting.

2. The main influential factors are considered adequately to forecast in this method, and it combines the fuzzy classifier and gray relation analysis to data-mine.

Through preprocessing it reduces training samples, picks up training speed and considers the weather factors.

3. Comparing with the single SVM and BP net-work, it can be proved that this method not only improves accuracy of short-term load forecasting, practicability of system, but also can be accomplished by software.

ACKNOWLEDGEMENT

The work is supported by the National Natural Science Foundation of China (No. 60602062) and supported by the scientific research foundation of ChiFeng College (No. ZRYB200608)

REFERENCES

1. D. Niu, S. Cao, and Y. Zhao, *Technology and Application of Power Load Forecasting* (China Power Press: Beijing, PK, 1998).
2. L.D. Chen and S. Toru, Data mining methods, applications, and tools, *Information system management*. Volume 17, Number 1, pp.65-70, (2000).
3. W. Zhang, T. Zeng, and H. Li, Parallel mining association rules based on grouping, *Computer engineering*. Volume 30, Number 22, pp.84-85, (2004).
4. Q. Li, L. Yang, X. Zhang, An effective apriori algorithm for association rules in data mining, *Computer application and software*. Volume 21, Number 12, pp.84-86, (2004).
5. N.V. Vladimir and X. Zhang, *Nature of Statistics Theory* (Tinghua University Press: Beijing, PK, 2000).
6. N. Deng and Y. Tian, *The New Approach in Data Mining-Support Vector Machines* (Science Press: Beijing, PK, 2004).
7. D. Tan and D. Tan, Small-Sample Machine Learning Theory-Statistical Learning Theory, *Journal of nanjing university of science and technology*. Volume 25, Number 1, pp.108-112, (2001).
8. Y. Li, T. Fang, and E. Yu, Study of support vector machines for short-term power load forecasting, in *Proc. of the CSEE*. Volume 25, Number 1, pp.55-59, (2003).
9. G. Xu and Y. Shi, Application of genetic algorithm in association rule mining, *Computer engineering*. Volume 28, Number 7, pp.122-124, (2002).
10. C. Liu and Y. Zhang, The data mining method based on the model of gm (1, 1) and the gray relation, *Changsha aeronautcal vocational and technical college journal*. Volume 5, Number 3, pp.60-62, (2005).
11. K. Li, C. Gao, and Y. Liu, Support vector machine based hierarchical clustering of spatial databases, *Journal of Beijing institute of technology*. Volume 22. Number 4, pp.485-488, (2002).
12. Z. Zi, S. Zhao, and G. Wang, Study of relationship between fuzzy logic system and support vector machine, *Computer engineering*. Volume 30, Number 21, pp.117-119, (2004).
13. W. Chen and T. Xu, The improving and realizing of association rule mining apriori algorithm, *Microcomputer development*. Volume 15, Number 8, pp.155-157, (2005).