A NEW SHORT-TERM LOAD FORECASTING MODEL BASED ON RELEVANCE VECTOR MACHINE

Zhinong Wei
College of Energy and Electrical Engineering, Hohai University, 210098, China
wzn_nj@263.com

Xiaolu Li
Alstom Grid, China Technology Center, No.500, Jiangyue Road, Shanghai, 201114, China
lily2.li@alstom.com

Kwok W. Cheung
Alstom Grid, 10865 Willows Road NE, Redmond, WA 98052, USA
kwok.cheung@alstom.com

Jiang Wu
College of Energy and Electrical Engineering, Hohai University, 210098, China
Jiang_wu1989@126.com

Shuaidong Huang
College of Energy and Electrical Engineering, Hohai University, 210098, China
hsd_1436@163.com

ABSTRACT

Considering the limitation of traditional feature extracting only the algebraic features of samples to the neglect of the practical significance of the original problem, a short-term load forecasting model based on the relevance vector machine (RVM) is proposed. By using the nonnegative matrix factorization (NMF) algorithm, the dimension of input variables is reduced, a short-term load forecasting model based on the RVM is proposed. The input data is decomposed by using the NMF algorithm, where the nonnegative lower-dimension mapping matrix derived is taken as the input of RVM for training and predicting. Due to the nonnegative property of the lower-dimension matrix, it retains the practical significance of the original problem while eliminating the redundant data and reducing dimensions. Simulation results show that the dimensions of the input variables can be effectively reduced and the predicting accuracy can be greatly improved.

Index Terms—Short-term load forecasting, Nonnegative matrix factorization (NMF), Relevance vector machine (RVM)

1. INTRODUCTION

Due to the natural factors such as weather conditions, economic activities and the impact of human factors, changes in short-term load are usually seen as a random process[1]. In recent years, artificial neural network, support vector machine and other nonlinear predicting models have been used to fit the load curve, which have achieved a good forecasting performance[2,3]. However, when the input factors are excessive, it will make the predicting model complex and the training efficiency low. Therefore, choosing reasonable predicting model is one of the key issues to improve the predicting accuracy.

In [4], the predicting model is built based on constructing a linear combination of the original variables by principal component analysis (PCA) method, which can extract the main components from the original data set and reduce the dimensions so that the problem was simplified. In [5], noise reduction filtering algorithm based on singular value decomposition was used to eliminate noise and extract the main features. In fact, the aforementioned methods can merely analyze sample matrix from the algebraic point of view, thus the decomposition results will bring negative values. Non-negative matrix factorization (NMF) algorithm is a new matrix decomposition algorithm proposed by Lee and Seung in 1999[6], compared with other feature extraction algorithms, the decomposition of this algorithm is non-negative and has the advantages of simplicity realization, interpretable decomposition form and decomposition result, etc.

On the other hand, relevance vector machine (RVM) is a probability learning method based on Bayesian theory[7]. In [8], empirical mode was used to divide load into several components, then using RVM to establish predicting model for each component, as a result, the predicting effect has been improved obviously.

Based on the above discussions, this paper uses the NMF algorithm and extracts the main features of the input variables, then the non-negative matrix with lower dimensions is derived and is taken as the input of RVM, thus it can improve the predicting accuracy.

2. METHODOLOGY

2.1 Non-negative matrix factorization

Non-negative matrix factorization can be described as follows[9,10]: For a given nonnegative dataset expressed by $V_{n\times m}$, it can be decomposed as the product of two nonnegative matrices:

$$V_{n\times m} \approx (WH)_{n\times m} = \sum_{j=1}^{r} (W_{nj} H_{jm}) \tag{1}$$

where the r columns of W are called NMF bases and the columns of H are its combining coefficients. The dimensions of W is $n\times r$ and H is $r\times m$. In order to reach the goal of reducing the dimensionality, the rank r of the factorization is chosen such that $(n+m)r < nm$.

There are two methods to find an approximating factorization $V \approx WH$. The conventional method is the square of Euclidean distance between V and W:

$$\| V - M \|^2 = \sum_{ij} \left(V_{ij} – M_{ij} \right)^2 , \tag{2}$$

where V_{ij} and M_{ij} are the elements of V and M, respectively.

Another popular method is the Kullback-Leibler divergence between V and WH:

$$D(V \| M) = \sum_{ij} \left(V_{ij} \lg \frac{V_{ij}}{M_{ij}} \right) - V_{ij} + M \tag{3}$$

This formation is not well-defined if $V_{ij} = 0$ or $(WH)_{ij} = 0$. Hence, we do not consider this formation in this study. These cost functions that have not any auxiliary constraints to W or H can be regarded as standard NMF algorithms.

By traditional gradient descent method, Lee and Seung proposed the multiplicative update rules for (2).

$$W_{ij} = W_{ij} \frac{\left(XH^T \right)_{ij}}{(WHH^T)_{ij}} \quad \forall i, j \tag{4}$$

$$H_{jk} \leftarrow H_{jk} \frac{\left(W^T H \right)_{jk}}{\left(W^T WH \right)_{jk}} \quad \forall j, k \tag{5}$$

2.2 Relevance vector machine

The output of RVM model can be defined as follows[11,12]:

$$y(x;\omega) = \sum_{i=1}^{N} \omega_i K(x, x_i) + \omega_0 \tag{6}$$

where N is the length of the data, $K(x, x_i)$ is a non-linear kernel function; ω_i is the model weights.

In RVM, the priori probability distribution for each model weight is given as:

$$p(\omega_i \mid \alpha)_i = (\frac{\alpha_i}{2\pi})^{\frac{1}{2}} \exp(-\frac{1}{2} \alpha_i w_i^2) \tag{7}$$

where α_i is hyper-parameter of the priori distribution of model weight ω_i.

Given a training sample set $\{x_i, t_i\}_{i=1:N}$, suppose the target value t_i is independent and the noise in data follows the Gaussian distribution with the variance σ^2, then the likelihood function of the training sample set can be represented by:

$$\begin{aligned} p(t \mid \omega, \sigma^2) &= \prod_{n=1}^{N} p(t_i \mid \omega, \sigma^2) \\ &= (2\pi\sigma^2)^{-N/2} \exp\{-\frac{\| t - \Phi\omega \|^2}{2\sigma^2}\} \end{aligned} \tag{8}$$

where $t = (t_1, \cdots, t_n)^T$, $\omega = (\omega_0, \omega_1, \cdots \omega_n)^T$, and Φ is the design matrix given by:

$$\Phi = \begin{bmatrix} 1 & K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_N) \\ 1 & K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(x_N, x_1) & K(x_N, x_2) & \ldots & K(x_N, x_N) \end{bmatrix}$$

Based on priori probabilities distribution and likelihood distribution, the posterior distribution over the weight form Bays rule can be written as:

$$\begin{aligned} p(\omega \mid t, \alpha, \sigma^2) &= \frac{p(t \mid \omega, \sigma^2) p(\omega \mid \alpha)}{p(t \mid \alpha, \sigma^2)} \\ &= (2\pi)^{-(N+1)/2|\Sigma|^{-1/2}} \exp\{-\frac{1}{2}(\omega - \mu)^T \ \Sigma^{-1}(\omega - \mu)\} \end{aligned} \tag{9}$$

where $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$

$\mu = \sigma^{-2} \sum \Phi^T t$

$A = diag(a_0, a_1, \cdots a_N)$.

The marginal likelihood distribution of the hyper-parameters can be obtained as:

$$p(t \mid \alpha, \sigma^2) = (2\pi)^{-\frac{N}{2}} |\Omega|^{-\frac{1}{2}} \exp\{-\frac{t^T \Omega^{-1} t}{2}\} \tag{10}$$

where $\Omega = \sigma^2 I + \Phi A^{-1} \Phi^T$

At last, the hyper parameter α and the variance σ^2 can be estimated by using the maximum likelihood method.

If the input value is x_i^*, then the corresponding output probability distribution obeys the Gaussian distribution, and the corresponding predictive value can be derived by:

$$y^* = \varphi(x_i^*)\mu \tag{11}$$

3. SHORT-TERM LOAD FORECASTING METHOD BASED ON RVM WITH NMF ALGORITHM

3.1 Data pre-processing

This paper selects the data collection and monitoring system measurement data acquisition (SCADA) as the original data source, the collected data often contains a lot of abnormal data. Therefore, it is necessary to detect and eliminate the dad data, in this paper, the transverse, longitudinal correlation method is used for pretreatment[11].

3.2 Sample selection

The characteristic factors selected in this paper include the following categories, as shown in table 1.

Table 1 Selection of characteristic factors

Number	Characteristic factor	Input variable
1	The load at the same time the day before	L_1

2	The load one hour ago the day before	$L_{1,1}$
3	The load two hours ago the day before	$L_{1,2}$
4	The load at the same time two days ago	L_2
5	The load one hour ago two days ago	$L_{2,1}$
6	The load two hours ago two days ago	$L_{2,2}$
7	The load at the same time three days ago	L_3
8	The load one hour ago three hours ago	$L_{3,1}$
9	The load two hours ago three days ago	$L_{3,2}$
10	The load at the same time in the previous week	L_7
11	The highest temperature the day before	$T_{1\max}$
12	The lowest temperature the day before	$T_{1\min}$
13	The average temperature the day before	T_{1ave}
14	The average wind speed the day before	W_{1ave}
15	The average temperature the day before	M_{1ave}
16	The mean pressure the day before	P_{1ave}
17	The average rainfall the day before	R_{1ave}
18	The date type of the prediction day	$D(d)$

Considering the influence of the data type to load, a simple mapping transformation for each data type in the table 1 is proposed. Monday to Friday is mapped for [1, 1, 1, 1, 1] and Saturday, Sunday is mapped to [2, 2]. Thus a matrix of 18 input influencing factors is established.

3.3 NMF algorithm decomposition of the input sample

Due to the load data and meteorological data is not comparable in their respective dimensions, the data is mapped by using the following mapping formula:

$$y = [(y_{\max} - y_{\min})(x - x_{\min})/(x_{\max} - x_{\min})] + y_{\min} \quad (12)$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minmum values of the original data x respectively； $y_{\max}$ and $y_{\min}$ are the mapping interval of maximum and minimum value respectively.

This paper introduces the k-fold cross-training method to get the optimal values of dimension, typically value of k =10[12]. The specific process is as follows:

Step 1: Training samples are randomly divided into k mutually disjoint, equal to the number of a subset, which is $S_1, S_2, \cdots S_K$.

Step 2: The dimension r must satisfy the condition $(m \times n)r \le mn$, then the NMF method can be used to reduce the data dimension.

Step 3: For k wheel training and testing, which $i = 1,2 \cdots k$ for k iteration. Upon completion of k times after training, and it can get the average of the accumulated error $\sum_{i=1}^{k} MRE_i^r / k$, which is called fold cross-validation error.

Step 4: Repeat Step 1 and 2 to get r cross-validation errors, selecting the minimum corresponding dimension as the optimal dimension.

3.4 RVM Modeling

This paper chooses the combination of the typical local nuclear-RBF kernel with the global nuclear-polynomial kernel [13], the polynomial kernel used in the binomial is shown as follows type:

$$K(x, x_i) = \lambda G(x, x_i) + (1 - \lambda) P(x, x_i) \quad (13)$$

$$G(x, x_i) = \exp(-\frac{\| x - x_i \|^2}{\sigma^2}) \quad (14)$$

$$P(x, x_i) = (x \cdot x_i) = (x \cdot x_i + 1)^2 \quad (15)$$

where $G(x, x_i)$ is RBF kernel; $P(x, x_i)$ is binomial kernel function; λ is the weight of the kernel function; δ is the kernel width; λ and δ are the parameters to be optimized. The grid search method is employed to get the optimal value of λ and δ. In which λ =0.471, δ =1.2.

4. SIMULATION ANALYSIS

In order to verify the effect of the model's prediction, this paper uses the data of a city in East China from September 15 to September 29 (24 points one day) for a simulation test, and the original data is 18-dimensional. In order to compare the predicting effect of different models, the data is used as a training set to set up the predicting model, which can check the prediction effect. Since the NMF algorithm requires the original input matrix elements non-negative, so the samples are mapped to [0, 2]. The K-fold cross-validation method is used to get the optimal dimension. The training set is decomposed into 7-dimensional, then it is taken as the input of RVM for training to get the predicting model. Extracting 5 groups of samples from the sample after dimension reduction, as shown in table 2, the data is nonnegative and retains the actual meaning of the original sample.

Table 2 Samples of data decomposed by NMF

Sample	Characteristic variables after NMF dimensionality reduction						
	①	②	③	④	⑤	⑥	⑦
1	0.17	5.01	0.94	1.38	0	0.19	1.25
2	0.41	5.14	0.91	1.33	0	0.17	0.99
3	0.08	5.09	0.45	1.48	0	0.12	0.92
4	0.52	4.52	0.75	1.35	0	0.16	0.82
5	0.17	4.52	0.30	1.43	0	0.14	1.11

In order to make a comparison, this paper uses the RVM method and the RVM based on PCA(PCA-RVM) method to make a forecast. In PCA-RVM model, the PCA information contribution is 90% and the result of the train sets are reduced to 10-dimensional. As shown in table 3, 5 set of sample data extracted by PCA method appears a large number of negative data.

Table 3 Samples of data decomposed by PCA

Sample	Characteristic variables						
	①	②	③	④	⑤	⑥	⑦
1	-1.58	1.07	0.27	-0.27	0.27	0.74	0.32
2	-1.66	1.07	0.26	-0.31	0.29	0.71	0.33
3	-1.53	1.06	0.26	-0.34	0.34	0.73	0.29
4	-1.51	1.09	0.26	-0.30	0.31	0.65	0.36
5	-1.41	1.10	0.26	-0.29	0.32	0.68	0.25

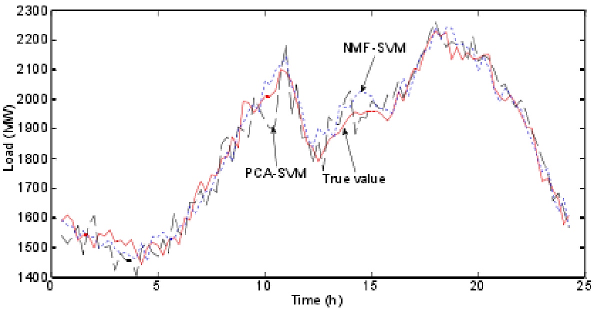
The errors of three predicting methods are shown in Table 4.

Table 4 Comparison of forecasting effect among three models

Prediction method	Variable dimension	The average relative error	Training time
RVM	18	3.55%	80s
PCA-RVM	7	2.70%	43s
NMF-RVM	7	1.92%	50s

Forecasting curves of load by PCA-RVM and NMF-RVM are given in Fig.1.

Fig.1 Forecasting curves of load by PCA-RVM and NMF-RVM



It can be clearly seen that compared with RVM method, the number of input variables is reduced based on NMF-RVM, the average relative error and training time are reduced. Besides, compared with PCA-RVM, the proposed method improves the accuracy of prediction obviously, the effect of dimension reduction is slightly better, but the training time becomes longer, this is mainly due to the use k-fold cross-training method to find the optimal number of dimensions and repeating use of RVM.

5. CONCLUSIONS

NMF algorithm theory and computational methods are still evolving. Further work of this paper is to find an effective way to reduce the time consumed by the algorithm in finding the optimal dimension. At the same time, the initialization of the algorithm, the convergence rate problems are worth further exploration and study.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants 51277052, 51107032 and 61104045.

REFERENCES

[1] P. Ju, J. K. Wang, X. Y. Zhao, 2001, "Ninety-six points short-term load forecasting theory and applications", *Automation of Electric Power Systems*, vol. 25, no. 22, 32-36.

[2] Z. N. Wei, D. Wang, G. Q. Sun, et al, 2005, "A novel method of short time load probability forecasting based on cascaded neural network", *Transactions of China Electrotechnical Society*, vol. 20, no. 1, 95-98.

[3] Y. Geng, X. H. Han, L. Han, 2005, "Short-term load forecasting based on least squares support vector machines", *Power System Technology*, vol. 20, no. 1, 72-76.

[4] B. Y. Liu, R. G. Yang, 2008, "Short-term load forecasting model based on LS-SVM with PCA". *Electric Power Automation Equipment*, vol. 28, no. 11, 13-17.

[5] Y. You, W. X. Sheng, S. A. Wang, 2008, "The study and application of the electric power system short-term load forecasting using a new model", *Proceedings of the CSEE*, vol. 28, no. 11, 15-18.

[6] D. Lee, H. Seung, 1999, "Learning the parts of objects by non-negative matrix factorization", *Nature*, vol.401, 788-791.

[7] M. E. Tipping, 2001, "Sparse bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, no. 3, 211- 244.

[8] Z. G. Sun, W. X. Zhai, W. L. Li et al, 2011, " Short-term load forecasting based on EMD and RVM", *Proceedings of the CSU-EPSA*, vol. 23, no. 1 , 92-97.

[9] W. J. Wang, J. Y. Zhang, 2009, "Fast algorithm for nonnegative matrix factorization", *Computer Engineering and Applications*, vol. 45, no. 25, 1-5

[10] P. Chen, H. J. Gao, H. Y. Shao, 2007, "Image recognition using SVM-weighted non-negative matrix factorization", *Third International Conference on Natural Computation*, vol.1, 577-581.

[11] X. M. Yang, 2001, "Daliy load forecasting based on rough sets and relevance vector machine", *2011 International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT)*, vol.6, 2946-2949.

[12] Y. H. Yang, Y. H. Liu, X. Z. Liu, S. K. Qin, 2011, "Billet temperature soft sensor model of reheating furnace based on RVM method", *2011 Chinese Control and Decision Conference (CCDC)*, 4003-4006.

[13] Q. Duan, J. G. Zhao, Y. Ma, 2010, "Relevance vector machine based on particle swarm optimization of compounding kernels in electricity load forecasting", *Electric Machines and Control*, vol.14, no. 6, 33-38.