POWER QUALITY INDEXES PREDICTION BASED ON CLUSTER ANALYSIS AND SUPPORT VECTOR MACHINE

Jie SONG
State Grid Shanghai EPRI – China
songjie@sh.sgcc.com.cn

Zhenzhen XIE
Shanghai university of electric power – China
xzz90s@126.com

Jian ZHOU
State Grid Shanghai EPRI – China
zhou.j@sh.sgcc.com.cn

Xiu YANG
Shanghai university of electric power – China
yangxiu721102@126.com

Aiqiang PAN
State Grid Shanghai EPRI – China
panaq@sh.sgcc.com.cn

ABSTRACT

In this paper, a methodology based on cluster analysis, feature selection and support vector machine is presented to predict power quality indexes such as harmonic distortion and voltage deviation. First, cluster analysis is made on power quality index data, and historical data sets to be used are selected according to the classification of the data to be predicted. Then, with the feature selection algorithm, the optimal factor combination relevant to the power quality index to be predicted is obtained from several factors, such as weather, active power, reactive power and others. Finally, the historical data and associated factor data are taken as the training sample data sets to establish the support vector machine prediction model. Through the analysis of the actual sample data during 2013 and 2014 in a certain area of Shanghai, China, it is proved that the average relative error of this method is greatly improved.

KEYWORDS

power quality, feature selection, cluster analysis, support vector machine, prediction.

INTRODUCTION

With the development of social economy, a large number of PQ problems have drawn the attention of power sectors and the users. For example, harmonic may cause the loss of power supply equipment, insulation aging and the interference to electrical equipment which will affect the safe operation [1-2]. And voltage deviation may have a greater harm as well, the quality of the industrial products will be affected when the voltage is too high or too low [3]. So more and more power quality information has been collected and analyzed [4]. The mining of PQ monitoring data in depth and the prediction of the trend of change in a scientific manner would become one of the necessary measures to ensure the safety and economic operation of power grid [5-6].

The generally utilized power quality index prediction methods are Time Series Algorithm, Neural Network Algorithm, etc [7]. Time Series Algorithm and Gray Model Algorithm are used to predict the PQ indexes in reference[8], but in this method, only the historical data is considered, and the dynamic effects on PQ by other factors is ignored, so the prediction accuracy is low. The method Time Series Algorithm and Neural Network Algorithm are used in reference[9], because only the correlation between active power and PQ indexes is considered, and the selected associated factors is not optimal, so the better prediction cannot be obtained.

In view of the above problems, a method based on support vector machine (SVM), feature selection and cluster analysis is proposed. SVM is proposed by Cortes and Vapnik, it is suitable for the prediction due to various advantages, such as high convergence rate [10]. In addition, associated factors selection is based on information theory and cluster analysis is based on Euclidean, which are conducive to the establishment of SVM model. So the prediction accuracy is greatly improved.

ASSOCIATED FACTORS SELECTION

Mutual information in information theory can describe the strength of the interdependence between the two random variables [11]. The mutual information between a PQ index and an associated factor characterize the degree of interdependence of them, and the greater value will make the factor more conducive to the classification. On the other hand, the mutual information between different associated factors represents the information redundancy between them [12]. Two factors with oversized information redundancy between each other should not be selected into the optimal associated factors combination at the same time.

Numerical sequences X and Y are respectively defined as class variable and feature variable. Then they will be discretized and converted into interval probability [13]. The mutual information between X and Y is:

$$I(X,Y) = -\sum_{i=1}^{N_i}[\frac{M_i}{M}\log\frac{M_i}{M}] - \{-\sum_{u=1}^{N_j}P(Y_u)\cdot \sum_{v=1}^{N_i}[\frac{M_{uv}}{M}\log\frac{M_{uv}}{M}]\} \quad (1)$$

In the equation, M is the sum of the quantities of all the elements in class variable X and feature variable Y. N_i is

the quantity of intervals which constitute class variable X. M_i is the quantity of the feature variable elements in the ith interval. N_j is the quantity of intervals which constitute feature variable Y. $P(Y_u)$ is the probability of feature variable Y falling into the uth interval. M_{uv} is the quantity of class variable elements in the vth interval when feature variable Y falls into the uth interval.

Information measurement function is defined as follows:

$$J(F_i) = I(C, F_i) - \beta \sum_{s \in S} I(s, F_i) \qquad (2)$$

In the equation, β is adjustment coefficient and C is numerical sequence of power quality index. All the associated factor sequences are taken as feature variable F_i, which constitute feature variable set F. All the selected feature variables are defined as s, which constitute selected feature variable subset S. The above numerical sequences of all the variables are discretized and converted into interval probability.

The steps of associated factors selection are as follows:

(1) Initialize the feature variable space. Input all the feature variables and assign F. Set set S to empty.

(2) Work out $I(F_i, C)$ of every F_i in set F.

(3) Delete F_i which leads to the maximum $I(F_i, C)$from set F and mark this operation as $F|\{F_i\}$. Then add F_i into set S and mark this operation as $\{F_i\} \rightarrow S$.

(4) Repeat the following steps until there are K feature variables in the set S.

(a) Work out $J(F_i)$ for every F_i in set F.

(b) Delete F_i which leads to the maximum $J(F_i)$ from set F and then add it into set S.

The mutual information $I(F_i,C)$ is used to measure the interdependence between associated factors and PQ indexes. And the mutual information between the candidate associated factors and the selected associated factors is used to measure the information redundancy. Finally, the latter mutual information is used to punish the former for selecting the optimal combination.

PREDICTION METHOD OF SUPPORT VECTOR MACHINE (SVM)

For training data set $\{x_i, y_i\}$, $x_i \in R^n$ is the input variable, while $y_i \in R$ is the variable for output. The data x of SVM are mapped to high-dimensional space through nonlinear mapping $\varphi$to perform the linear regression by using the following linear function.

In regression function $f(x)=\omega\varphi(x)+b$,ω is the weight vector while b is the deviation. ω and b can be solved with structure risk minimization principle equation(3):

$$\min\left\{\frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n} L\left(y_i, f\left(x_i\right)\right)\right\} \qquad (3)$$

In the equation, L(y, f(x)) is the loss function. Its expression is:

$$L\left(y, f\left(x\right)\right) = \begin{cases} 0, |y - f\left(x\right)| \le \varepsilon \\ |y - f\left(x\right)| - \varepsilon, \text{others} \end{cases} \qquad (4)$$

Introduce slack variable ξ_i and ξ_i^*. Then the question is converted into:

$$\min\left\{\frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)\right\} \qquad (5)$$

$$s.t.\begin{cases} y_i - f\left(x_i\right) \le \varepsilon + \xi_i \\ f\left(x_i\right) - y_i \le \varepsilon + \xi_i^* \quad i = 1, 2, \ldots n \\ \xi_i > 0, \xi_i^* > 0 \end{cases} \qquad (6)$$

In the equation, c is equilibrium coefficient while ξ_i and ξ_i^* are penalty functions.

Introduce Lagrange multiplier α and$\alpha*$.The results are as follows:

$$\omega = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varphi\left(x_i\right) \qquad (7)$$

$$\begin{aligned} f\left(x\right) &= \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)\left(\varphi\left(x_i\right)\varphi\left(x\right)\right) + b \\ &= \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K\left(x_i, x\right) + b \end{aligned} \qquad (6)$$

$K\left(x_i, x\right) = \varphi\left(x_i\right)\varphi\left(x\right)$ is the kernel function which satisfies the Mercer condition.

K-MEANS CLUSTER ANALYSIS

The cluster analysis of a power quality index is made, and the training data set is selected by the cluster results. The measurement is carried out by Euclidean distance.

First, define the Euclidean distance D_{ij} between two m-dimensions vectors X_i, X_j as:

$$D_{ij} = \sqrt{\sum_{l=1}^{m}(x_{il} - x_{jl})^2} \qquad (8)$$

The steps of cluster analysis are as follows:

1) Take K points from n m-dimensions data points X={X1, X2, ... , Xn} as initial cluster centers u_k (k=1, 2,…, K). Work out the Euclidean distance between X_i and every center.

2) Mark X_i as the same class with u_k which make the Euclidean distance between X_i and u_k smallest. Then repeat 1) until all the elements of X has been classified. Finally update the cluster centers u_k.

3) Repeat 1) 2) until the number of iterations meets the requirement.

TOTAL PROCESS FLOW OF PREDICTION

(1) By the cluster analysis of PQ indexes, the PQ index data at the time point in the same class of the time point

to be predicted is selected as the historical data set.
(2) According to the associated factors selection method to select the optimal combination.
(3) Take the data set in step (1) and its corresponding associated factors data as the training data set, to train the SVM model and make prediction of PQ indexes.

CASE ANALYSIS

The data in one 35 kv substation of Shanghai, China from May 1, 2014 to May 1, 2013 is selected as a sample. The sample data also includes the corresponding weather, date type and other associated factors data. The voltage deviation data set of January 2014 and November 2013 is used as a test data set in the following test part.

Cluster analysis

According to Sum of the Squared Error (SSE) to assess the cluster quality with different quantities of clusters. The cluster quality is the highest when the cluster quantities of voltage deviation, frequency deviation, three-phase unbalance of voltage, total harmonic distortion of voltage are 5, 5, 6, 5 respectively, namely, they are divided into 5 classes, 5 classes, 6 classes and 5 classes in row. Figure 1 to figure 4 are the cluster diagrams of the above-mentioned PQ indexes from May 2013 to May 2014. The colors represent the classes, and the horizontal axis is the accumulated time of each data point.

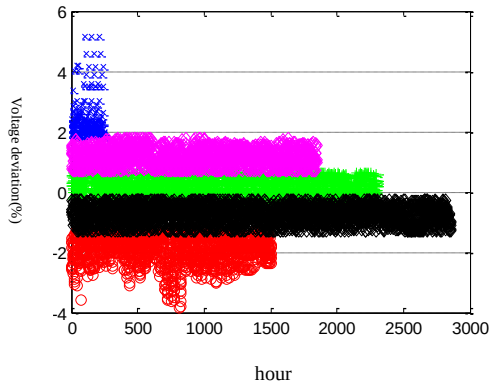

Fig 1 cluster analysis result of voltage deviation

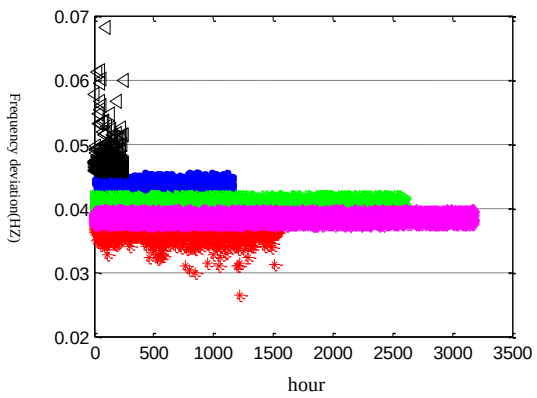

Fig 2 cluster analysis result of frequency deviation

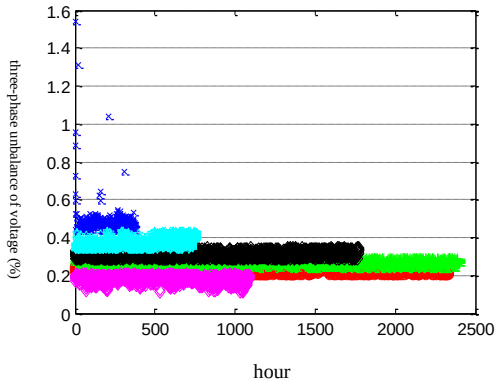

Fig 3 cluster analysis result of three-phase unbalance of voltage

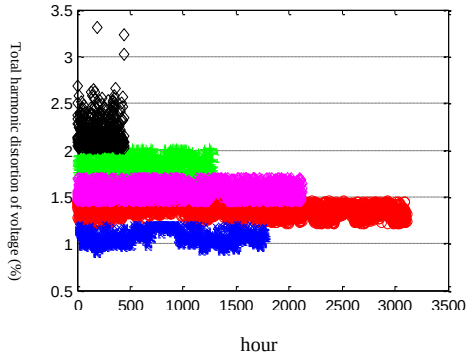

Fig 4 cluster analysis result of total harmonic distortion of voltage

For example, in the result of figure 1, the 5 clusters divide the voltage deviation in the numerical value into 5 grades, at the same time, the voltage deviation will be divided into 5 classes in accordance with the time and the date. The blue cluster, pink cluster, green cluster, black cluster and red cluster are nearly 2% to 4%, 0.8 to 2%, 0% to 0.8%, -0.7% to 0 and -4% to -0.7%. Besides, the time points of all the data points have been divided into 5 classes, which can't be seen in the figure.

Associated factors selection

The associated factors to be selected are weekend mark, holiday mark, time, temperature, air humidity, atmospheric pressure, rainfall, active power and reactive power. Take the adjustment coefficient β=0.7 to establish the function, the optimal combination can be seen in table 1.

Table 1 Combination of associated factors

Associated factor	result
weekend mark	×
time	√
temperature	√
holiday mark	√
air humidity	×
atmospheric pressure	×
rainfall	√
active power	√
reactive power	×

In Table 1, √ means that the said factor is selected into the combination, and × means that the factor is not selected.

Prediction results and analysis

Two methods are used respectively, that is, the method of this paper and the Time Series Algorithm [14] to make the prediction. Figure 5 is the prediction result of voltage deviation data by using Time Series Algorithm from January 1, 2014 to the second day in 24 h. The prediction is carried out based on the data of taking a point every 3 minutes. The blue part is the prediction curve and the red part is the true curve. Figure 6 shows that the prediction accuracy of the method is very low. Figure 7 shows the prediction of continuous 12 hours in January 6, 2014 by using the method of this paper. The prediction is based on the data of taking a point every hour. The diagram is drawn after converting voltage deviation to rms voltage.

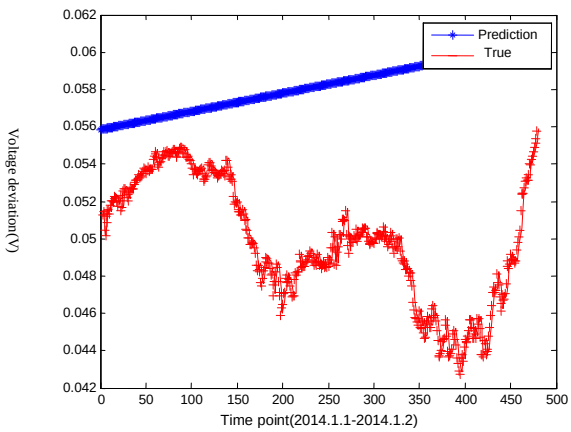

Fig 5 Prediction result of Time Series Algorithm

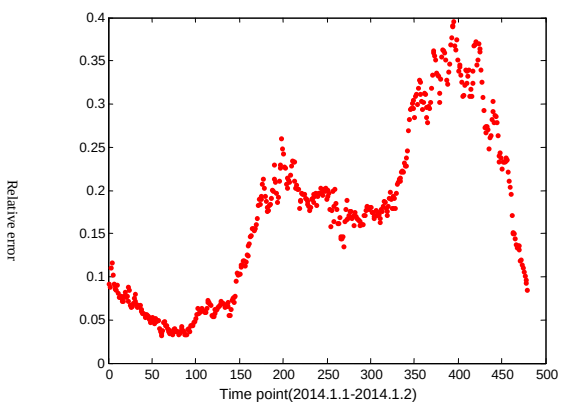

Fig 6 Relative error of Time Series Algorithm

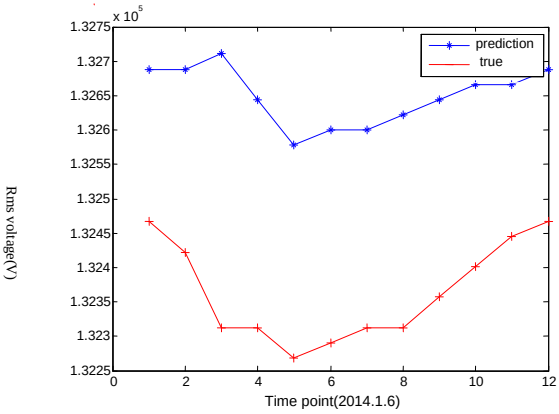

Fig 7 Prediction result of the method of this paper

Based on the results of 61 days in January 2014 and November 2013, the average relative error for the prediction of voltage deviation by using the method of this paper is 5.22%, and the average relative error of Time Series Algorithm is 20.01%.

CONCLUSIONS

In this paper, a new mathematical model using feature selection, cluster analysis and support vector machine is established to predict the power quality indexes. Implicit classifications of indexes have been found out by cluster analysis based on Euclidean distance. Then the optimal combination of the associated factors is selected according to the feature selection method from all kinds of factors such as weather, temperature, air humidity, atmospheric pressure and so on, which gives significant help to the SVM model. Finally, the SVM model is established, which can effectively control the prediction error and has a good practicability.

REFERENCES

[1] Basu M, Basu B, 2007, "Analysis of power quality (PQ) signals by continuous wavelet transform", Proc of power electronics specialists conference, 2614－2618.

[2] Zhuang Houshun, Meng Qinghai, Yuan Zhen, 2010 "preliminary analysis on power quanlity", Science and technology information, no. 35, 282.

[3] Han Hongke, Qi Linhai, 2010, "Application andresearch of multidi-mensional data analysis in power quality," Proc ofICC-DA.

[4] Wang Lin, Wang Qian, Guo Hanqiao, 2007, "An intelligent electronic device model for monitoring power quality based on IEC61850", Power System Technology, vol. S2, no. 31, 268-11.

[5] Pei Lin, Zheng Jianhui, Sun Yi, 2007, "Study on power quality comprehensive monitoring and management system", Power System Protection and Control, vol. 23, no. 36, 49-53.

[6] Wang Ling, Kang Jian, Zou Hongliang, et al,, 2011

"Construction and application of realtime power quality monitoring system", Power System Protection and Control, vol. 2, no. 39, 108-111.

[7] Su Weiwei, 2014, "Research of steady-state power quality forecasting and early warning based on data mining technology", NorthChina Electric Power University, Beijing, China, 6-11.

[8] Ding Junze, Liu Ping, Ou Yangsen, et al, 2015 "Mechanism of power quality forecast and early warning and their application", vol. 27, no. 10, 87-92.

[9] Su Weiwei, Ma Suxia, Qi Linhai,, 2014, "Predicting of Power Quality Steady Indexes Based on ARIMA and Neural Network", vol. 24, no. 3, 163-167.

[10] Zhang Lin, Liu Xianshan, Yin Hejun, 2004 "Application of support vector machines based on time suquence in power system load forecasting", vol. 19, no. 28, 38-41.

[11] Yao Xu, Wang Xiaodan, Zhang Yuxi, et al 2012 "Summary of feature selection algorithms", Control and Decision, vol. 2, no. 27, 161-192.

[12] Yang Dasheng, 2005, "Study of feature selection based on mutual information", Southeast University, Nanjing, China, 9-12.

[13] Yuan Yuan, Gu Jie, Huang Wei, et al, 2004 "Application of mutual information to power system medium and long-term load forecasting" , East China Electric Power, vol. 2, no. 37, 236-239.

[14] Zhang Li, 2008, "Study of time series forecasting algorithm and system realization based on the ARIMA model", Jiangsu University, Jiangsu, China, 80-83.