Research on weather classification pattern recognition based on support vector machine

Guo jia[1,a*],Li Teng[1],Cheng Rong[1] and Tan Lingfeng[2]

[1]STATE Grid Hebei Economic Research Institute, Shijiazhuang, China
[2]STATE Grid Economic Research Institute, Beijing, China

Abstract. weather is the most important factor affecting the photovoltaic power generation.In this paper,the irradiance data of a photovoltaic power station in crodora in 2020 are collected,and the daily out of ground irradiance and the measured irradiance curve of that day are compared and observed,then the weather of that year is classified by human work,and then the daily irradiance data records are counted for the relevant indicators,with the maximum third order Based on the attributes of difference value,discrete difference and normalized variance,it is unified with the classified weather type.Then,the SVM prediction model of weather category is established based on radial basis function,and the optimal model parameters are determined by cross validation,so that a large number of historical date weather categories can be classified and predicted.This is obviously different from the traditional prediction method based on linear statistical theory,and the results show that it has a good effect.

1 Introduction

In today's society,energy shortage and environmental pollution are increasingly prominent.People are actively looking for renewable clean energy as a substitute for fossil energy.Compared with the traditional energy,solar energy has been widely concerned because of its rich reserves,no transportation,clean and pollution-free.

The output power of photovoltaic power station mainly depends on weather factors such as irradiance[1-2].For the photovoltaic power prediction[3],the historical weather can be divided into several types first,so as to establish the corresponding historical model,and then the forecast weather types can be mapped to the above-mentioned different divisions,and then the relevant factors in the latest period of the same weather type can be extracted from the related database to achieve the power prediction of the photovoltaic power station.Such prediction accuracy ratio is not The prediction accuracy of classification is much higher.Therefore,the weather classification is the premise of power prediction.In essence,weather classification is a problem of pattern recognition,which includes two basic aspects: feature selection and classification.At present,the main methods of classification and prediction are as follows[4]:1.Bayesian classification;2.Back propagation neural network learning algorithm;3.K-nearest classification.

In reference[5],based on the theory of Bayesian inference learning,the paper studies the problem of rainfall prediction by using Naïve Bayes classifier,and proposes a simple Bayes algorithm,learn and classify raincall,which classifies each prediction factor and prediction target according to the meteorological classification standard.Literature[6] classifies the relevant data by using the competitive learning of neural network,and divides the historical data into several categories to find out the prediction categories of the same type as the prediction date.The corresponding BP algorithm is used for short-term load forecasting in the next 24 hours.In reference[7],aiming at a practical classification problem,a network is trained by using the idea of generalized radial basis function network and the classification prediction of test data set is realized.The algorithm uses k-value clustering algorithm to train the center of generalized radial basis function network,and uses singular value decomposition to calculate the weight of output layer.

However,the above methods are faced with some problems,such as the network parameters are difficult to be determined reasonably,easy to fall into local minimum,the network learning process is prone to concussion,and the generalization ability is not strong.Based on the support vector machine, it has a strong classification function and good generalization performance.It can obtain a strong generalization ability when there are few training samples,which depends on learning a large number of training samples.The neural network which can obtain generalization ability is incomparable,and is suitable for solving high-dimensional and nonlinear problems.Therefore,it has a very wide application prospect in classification prediction.

In this paper,we use the method based on historical data and support vector machine to build the weather classification prediction model.First,the classifier is constructed by analyzing or "learning" from the training set.The training set consists of database tuples

[a]Corresponding author: 523405800@qq.com

(represented by n-dimensional attribute vectors) and their corresponding class numbers.Assuming that each tuple belongs to a predefined class,the learning model can be provided in the form of classification rules,decision trees or mathematical formulas.Use the model to classify the future or unknown objects.First,evaluate the prediction accuracy of the model:for each test sample,compare the known class number with the learning model class prediction of the sample.The accuracy of the model on a given test set is the percentage of the correctly classified test samples.The test set should be independent of the training sample set,or "over fitting" will occur Condition.

2 Classification significance

Solar irradiance is the main factor affecting the power generation of photovoltaic power station.According to relevant data,if we know the basic information of longitude and latitude,date and time of a certain place,we can get the daily theoretical irradiance value[8] more accurately.The theoretical irradiance value outside the ground always presents a parabola like shape in the period with irradiance,which is independent of weather type and season,which also proves that it is only related to the location and date,but the actual power generation of solar power station is not directly related to the external irradiance.The main factor affecting the power generation is the measured irradiance,which is always less than or equal to (when the irradiance is 0,the two are equal) the external theoretical irradiance value,because the ground irradiance is the scattering and refraction of the external irradiance through the atmosphere,clouds,fog,snow reaching the ground.The following figure can be used as a special effect to see the representative weather of three types of situations according to the actual measured irradiance.Figure 1-Figure 3:no precipitation,low precipitation and high precipitation.Among them,the red is the theoretical irradiance outside the earth,the blue is the measured irradiance,and the sampling interval is 1 hour.From Figure 1-Figure 3,we can see that if the irradiance prediction is not classified,the theoretical results should be poor.

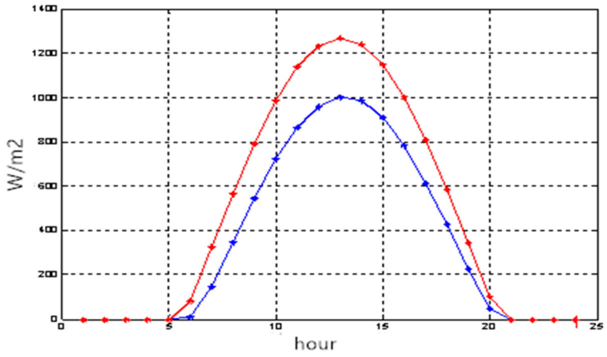

Fig1. no precipitation

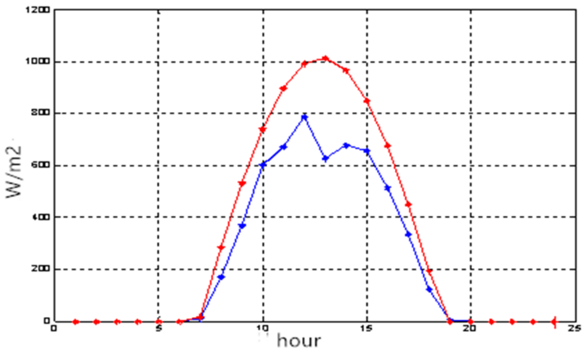

Fig2. low precipitation

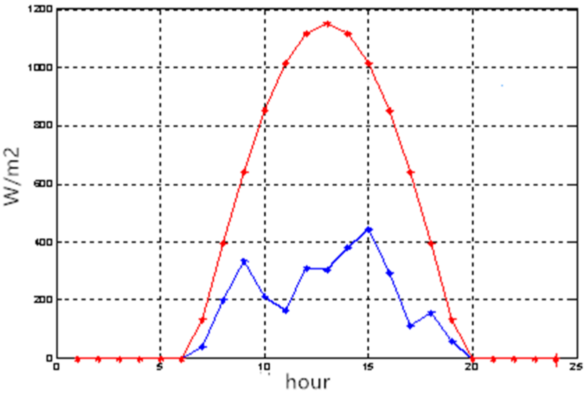

Fig3. high precipitation

3 Support vector machine[9-11]

Support vector machine (SVM) is a new machine learning algorithm proposed by Vapnik and others according to the principle of statistics.Its biggest feature is to improve the generalization ability of learning as much as possible according to the principle of structural risk minimization of Vapnik,that is,the small error obtained from the limited training samples can still guarantee the small error of independent test set.In addition,since SVM is a convex optimization problem,the local optimal solution must be the global optimal solution,which can prevent over learning.These characteristics are beyond other algorithms,such as neural network learning algorithm.For the classification problem,the SVM algorithm can be simplified to map the samples in the input space to a feature space through a certain nonlinear function relationship,so that the two types of samples(which can be extended to multiple types) can be linearly separable in this linear space.The real valuable application of SVM is to solve nonlinear problems and find the optimal linear hyperplane of samples in this feature space.Its discriminant function is

$$f\left(x\right)=sign\left(\sum_{i=1}^{k}\alpha_i y_i k\left(x,x_i\right)+b\right) \quad (1)$$

In order to determine the parameters of the optimal hyperplane partition, $k\left(x,x_i\right)$ is called kernel function.The kernel function $k\left(x,x_i\right)$ should be selected as a point product of the feature space,that is,the existence function.It has been proved that the kernel function can meet the above requirements as long as it satisfies Mercer condition.Common kernel functions are:

(1) Polynomial kernel function

$$k\left(x_i,x_j\right)=\left(x_i\times x_j+1\right)^d \tag{2}$$

(2) Radial basis kernel function

$$k\left(x_i,x_j\right)=\exp\left(-g\left\|x_i-x_j\right\|^2\right) \tag{3}$$

(3) Sigmoid kernel function

$$k\left(x_i,x_j\right)=\tanh\left[b\left(x_i\times x_j\right)+c\right] \tag{4}$$

In this paper,radial basis function is used as kernel function to establish the classification prediction model.Using the improved version of MATLAB forum of libsvm software package developed and designed by Professor Lin Zhiren of Taiwan University,this software package is convenient for improvement,modification and application in other operating systems.It involves relatively little parameter adjustment for SVM,and provides many default parameters.Using this software,we can solve the problems of C-SVM,v-SVM,ε -SVR and v-SVR, including many based on one-to-one algorithm Class pattern recognition.

At present,there are two main methods to construct SVM multi class classifier.One is the complete multi class SVM method,which is represented by the multi value classification algorithm proposed by Weston [12].By changing the original optimization problem of two class SVM,the multi class classification model is reconstructed and its objective function is optimized to achieve multi class classification at one time.The objective function of this kind of algorithm is very complex,the computational complexity is also very high,it is difficult to achieve,the classification accuracy is not dominant,and it is difficult to practical application.The other is the combination of multi class SVM method.Its basic idea is to realize the construction of multi class SVM by combining multiple two class SVM classifiers.There are five methods,including one to many(one to many) method,one to one method(pair classification,voting method),directed acyclic graph method,error correction coding method,and hierarchical method based on binary tree[13-15].According to SVM theory,the training time of combined multi class SVM depends on the number of training samples and has nothing to do with its feature dimension.The classification time depends on the number of two kinds of SVM needed for construction and the number of support vectors in two kinds of SVM.By comparing the advantages and disadvantages of the five methods,we can see that the number of two kinds of SVM constructed by one to many methods is moderate,the algorithm is simple and easy to implement,the classification results of the samples to be tested are independent of the order in which the two kinds of SVM are used,and the generalization upper bound is independent of the dimension of the feature space.The training time is short and can be used for large-scale data classification,so it is the most widely used.

4 Classification model based on support vector machine

The model in this paper is mainly based on cross validation.Cross validation is a statistical analysis method used to verify the performance of classifiers.The basic idea is to group the original data (data set) in a certain sense,one part as the training set,the other part as the validation set.First,the classifier is trained with the training set,and then the model is tested with the validation set to evaluate the performance of the classifier.The cross validation process is also the parameter selection process of the optimal model.At present,there is no good method for the optimization of parameters.Here,the continuous test parameters are used to build the model.SVM breaks through the traditional "inductive principle of empirical risk minimization"to evaluate and test the model,and puts forward "inductive principle of structural risk minimization",which takes into account both fitting error and generalization ability,so it not only considers the fitting accuracy of training samples,but also considers the generalization ability of the model[16].The parameters of the model are adjusted continuously.The two parameters are mainly adjusted,one is the penalty coefficient C,the other is the g in the radial basis function.The first step is to adjust the parameters with larger steps to determine the parameters of the optimal model,and the second step is to select a smaller cycle near the parameters of the optimal model to further optimize the model.The optimization contours of C and G are shown in Figure 4.

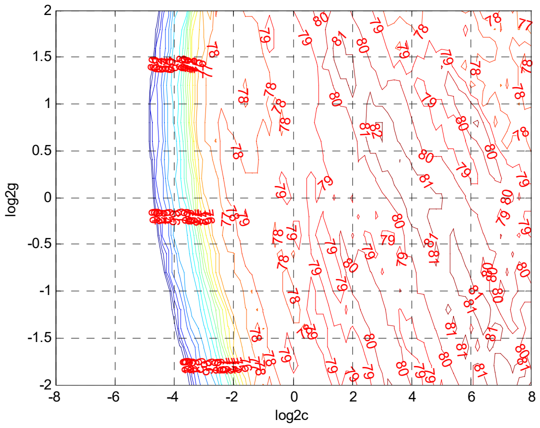

Fig4. C and g optimization contour map

5 Data and methods

5.1 data source and processing

The data in this paper are selected from the data of the American renewable energy laboratory.Colorado (srrl BMS) is located at 39.742°N 105.18°W and 1828.8 m above sea level[17].

the original data is processed according to the following five points:

1)Exclude abnormal data,such as maximum or minimum data;

2)For the missing data,linear interpolation is used to complete the data;

3)Remove the data that the exposure is greater than the theoretical value of the extraterrestrial radiation;

4)If the relative humidity data is greater than 100%,it shall be treated as 100;

5)If the calculated sunshine duration is greater than the theoretical sunshine duration of the day,replace it with the theoretical sunshine duration.

5.2 detailed description of data attributes

Select the solar irradiance value of the test point every one hour in a 365-day period to obtain a 36524-dimensional matrix.Count the irradiance in each data record,repeatedly observe the attribute value distribution map,and then calculate some mathematical indicators related to irradiance,so as to obtain the maximum value of irradiance difference per hour and dispersion of irradiance in a day ,Normalized dispersion difference of irradiance,maximum third-order reciprocal difference of irradiance per hour and maximum absolute value of irradiance third-order reciprocal,normalized variance,actual irradiance curve and extraterrestrial irradiance curve The surrounding area ratio and other related attributes.Based on these indexes,a mathematical model of three types of weather is established,which can be unified with the physical model.That is to say, a certain weather type (belonging to the above three types of weather)is corresponding to a set of certain thresholds of attributes.Then only the corresponding attribute values of each day in the historical data need to be calculated to achieve the unification of physical model and mathematical model,so as to use the SVM model to classify the weather types of historical date.

5.3 data attribute visualization

In this way,the data contains 365 valid samples,which belong to three types of weather,with category labels of 1, 2 and 3.Each sample contains 9 relevant calculation indicators of irradiance mentioned in the previous section as category attributes.

The visualization of the data is shown in Figure 5 and Figure 6.

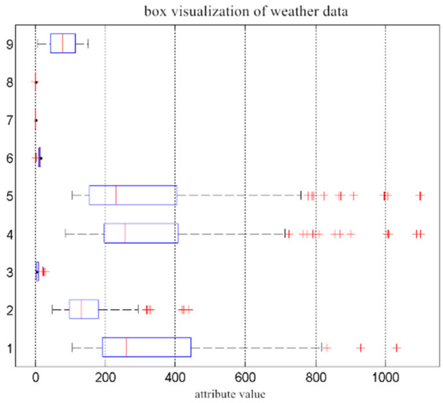

Fig5. Data Visualization Diagram

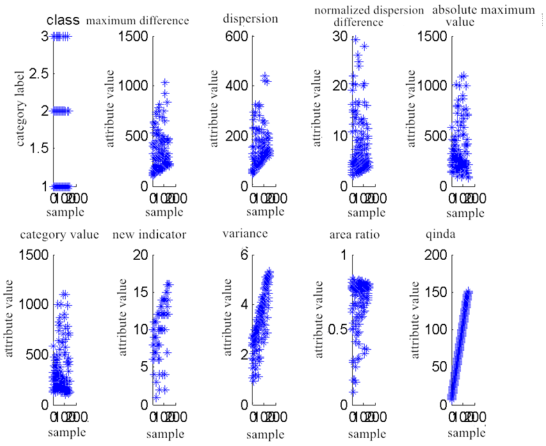

Fig6. Visualization diagram of data attributes

5.4 data normalization

In the process of classification,data need to be normalized.There are two ways to normalize factors:

(1) Maximum and minimum method

$x \to f(x) = (x - x_{\min})/(x_{\max} - x_{\min})$, $x_{\min}$ is the minimum value of each factor sequence,and $x_{\max}$ is themaximum value of each factor sequence.This normalization mode normalizes each feature component of the original data to a value on [0,1].

(2) Average method

$x \to f(x) = (x - x_{mean})/x_{\text{var}}$,this normalized mode Where,each characteristic component of the original data is normalized to a value on[- 1,1].

In this paper,the first normalization method is used to normalize the original data to the range of [0,1] by using mapminmax function in MATLAB.

5.5 selection of training set and test set

Among the 365 samples in the selected database,there are 178 in the first category,108 in the second category and 79 in the third category,with category labels of 1, 2 and 3 respectively.In this paper,70% of the weather samples of each category are selected as training samples,and the remaining 30% of the corresponding data are used as test sets to test the accuracy of classification.Figure 7 is the classification result chart of the final forecast.It can be seen from Figure 7 that the accuracy of SVM model with 70% training samples is 95.4128%(104/109).There are 5 sample classification errors,3 belong to the second category,and the results are classified into the first category.The other two samples belong to the third category,and the results are divided into the second category.But the accuracy of the whole model is better,and it achieves the expected effect.

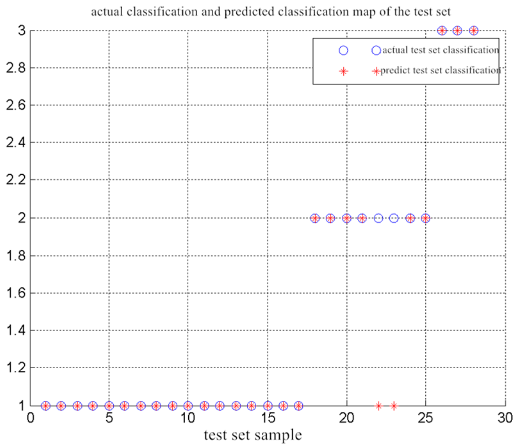

Fig7. actual classification and predictio classification diagram of test set

6 conclusion

The experimental results show that SVM algorithm can get quite good results in solving the classification problem.According to the final classification results,the weather type of the day to be predicted is corresponding to the classified weather type.The weather factors of the same type,such as irradiance in recent days,are extracted from the associated database to predict the weather factors of the day to be predicted.Then the above prediction values are taken as the input and the correlation data is matched,and the corresponding generation power value is retrieved as the power prediction value of the power generation unit.The sum of the predicted power of each generating unit is the predicted total power of the photovoltaic power station.

Because SVM has powerful classification function and good generalization performance,it has a very wide application prospect in classification and prediction.

References

1. TadrosMTY. Uses of sunshine duration to estimate the global solar radiation over eight meteorological stations in Egypt. Renewable Energy 2000, 21(2):231-46.
2. Yang Zhao, Yu Wenhong, Zhang Furen. Heat gain analysis of building solar radiation in winter [J]. Journal of solar energy, 2005, 26 (1): 108-l13
3. Almorox J, Hontoria C. Global solar radiation estimation using sunshine duration in Spain. Energy Conversion and Management 2004, 45(9–10):1529-35.
4. Fengguohe. Performance comparison of four classification methods [J]. Computer engineering and application, 2011, 47 (8): 25-26
5. Wu Xinling. Classification prediction based on Bayesian method [J]. Computer engineering and application, 2004, 33:195-197
6. Yao LiXiao, Yao Jinxiong, Li Baoqing, Wan Shixin. Short term power load forecasting based on neural network competition classification [J]. Taiyuan electric technology. 2008, 28 (10): 45-48
7. Zhang Li, Jiang Hao, Pu An Jian. Neural network classification prediction based on generalized radial basis function [J]. Computer technology and development, 2009, 41 (2): 105-109
8. Vapnik V. The nature of statistical learning theory[M].New York.Springer:1995.
9. Vapnik V. Statistical learning theory[M]. New York.Weily:1995.
10. Zhu Ming. Data mining [M]. China University of science and Technology Press, 2002
11. Weston J, Watkins C. Multiclass vector machines [c] / M. Verleysen: Proceedings of ESANN99, Brussels,1999: 41-83.
12. Xu Qihua, Shi Jun. aeroengine fault diagnosis based on support vector machine [J]. Journal of Aeronautical power, 2005, 20 (2): 298302
13. Hsu Chih-W ei,L in Chih-Jen. A Comparison on methods for multi-class support vector machines [J] . IEEE Transaction Neural Networks, 2002,13(2): 415-425.
14. Hsu Chih-W ei,L in Chih-Jen. A Comparison on methods formulti-class support vector machines [J] . IEEE Transaction Neural Networks,2002,13(2): 415-425.
15. Zheng Yongtao, Liu Yushu. Research on solving multi classification problems with support vector machine [J]. Computer engineering and application,2005, 23: 190-192.
16. Chen Yongyi support vector machine method and fuzzy system [J]. Fuzzy system and mathematics, 2005, 19（1）： 1-11 .
17. http://www.nrel.gov/midc/srrl_bms/