Based on Randomly Executed Sequence and GA-SVC Applied on Air Quality Analysis

Fan Wei1, Honglin Sun2
1School of Computer Science, Xi’an Shiyou University, Xi’an Shanxi, 710065, China.
2School of Computer Science, Xi’an Shiyou University, Xi’an Shanxi, 710065, China.
Corresponding author’s Email: sunhonglin@ime.ac.cn

Abstract. Air quality is a key factor affecting people’s daily travel. In order to analyze and classify air quality, and to solve the problem of low classification accuracy, this paper mainly studied the SVC (multi-classification support vector machine) Algorithm based on the random sequence selection of optimized parameters through Genetic Algorithm. Aiming at the problems of insufficient classification accuracy in current multi-feature sequence analysis and unstable parameter selection in SVC algorithm, comprehensive classification evaluation of data is carried out by analyzing data features and correlation among features and integrating genetic algorithm into SVC to optimize parameter selection, so as to improve classification accuracy. The experimental results show that the classification accuracy has been improved by 5% on average with the current popular decision tree classification algorithm, unoptimized SVC algorithm and KNN algorithm.

1. Introduction

In recent years, with the deepening of global industrialization and modernization process, and with the increasing attention to air pollution, the air quality affects every aspect of human production and life, so for the evaluation of air quality in air pollution control research, people’s daily work travel is important.

With the research and application of sorting algorithms by a large number of researchers in recent years, many models for air quality classify and assess have emerged. At present, the main models include: Yang jinwei et al's prediction of air pollutant concentration based on gray markov model [1]: Shen jin et al's air quality prediction model based on clustering and multiple regression [2], the predicted result of the model has a correlation with the actual value of 0.76. Prediction and analysis of air quality based on BP neural network by Tian jingyi et al [3].

Although various algorithms have a lot of research results in the field of air quality prediction, there are still some shortcomings at present:

1) The accuracy of classification results is unstable or even low due to the influence of many factors.

2) The parameter selection of the model has coarse granularity, which leads to the problem that the optimization degree cannot reach a higher level.

In this paper, a method of air quality classification by random sequence and GA-SVC is proposed to solve the above problems. First, according to a certain area of the UK, daily air levels from 2009 to 2018 include O3, NO2, SO2, CO, PM2.5, PM10 six major factors affecting air quality [4].

Concentration data and the quantitative evaluation of local meteorological department daily air quality level (level 1-10, the greater the number the worse air quality) data, through data analysis and processing, the data into the matter has been through random sequence and genetic algorithm to optimize
test evaluation of air quality in the SVC training, and then classified forecast data in SVC after training evaluation to achieve the purpose of classification of air quality.

2. Data collection and preprocessing

Daily ozone, nitrogen dioxide, sulphur dioxide, carbon monoxide, PM2.5, PM10 concentrations and assessed air quality grades in a UK region in this study are derived from the website of the UK center for atmospheric information resources [5]. The data of January 1, 2009, solstice and January 1, 2018 were used to analyze the data and build the test model.

2.1 data preprocessing

2.1.1 missing value processing. There are missing values of various factors affecting air quality in the data to varying degrees. In this paper, the four-point adjacent point mean algorithm is used to fill in the missing values, namely.

\[
X_i = \frac{1}{4} (X_{i-2} + X_{i-1} + X_{i+1} + X_{i+2})
\]

Where, \(X_i\) is the missing value. If there are more than 5 missing values in a row, it is considered that the high probability of prediction will produce a huge deviation, and the data will be discarded.

2.1.2 feature correlation test. Pearson two-tail test was carried out according to covariance and cross product deviation between any two variables of the six substances by SPSS [6-7] statistical software, the results showed that the correlation between PM2.5 and PM10 reached 0.912, which was significant. The characteristic of PM10 was abandoned through observation and comparative analysis.

2.1.3 data filtering. According to the air quality index (the classification standard), according to the local actual situation and data analysis, the air quality index is considered to be normally distributed, and the mean and standard deviation are calculated to be 2.44 and 0.716, respectively.

\[
X = \{ \mu - 3\sigma, \mu + 3\sigma \}
\]

Filter outliers of data, and finally treat tuples with index values greater than and equal to 5 as outliers and eliminate them. After multi-layer data filtering, 3029 groups of valid data were finally selected.

3. Randomly executed sequence GA-SVC algorithm model

3.1 Genetic algorithm (GA)

Genetic Algorithm (GA) is derived from the research on computer simulation Algorithm of biological ecosystem [8]. It is an optimization algorithm based on natural genetic law and natural selection principle of survival of the fittest, drawing on Darwin's theory of evolution and Mendel's theory of heredity. Its essence is a high-precision, parallel and global optimization algorithm, which was first proposed by professor J.Holland in his book "adaptability of natural combination artificial intelligence system" in 1975.

Genetic algorithm is mainly used to seek the optimal solution of optimization problem step by step through binary coding, decoding, solving fitness, replication, crossover, mutation and other steps. In this study, it is mainly used for the optimization of SVC parameters. Figure 1 is the main flow of genetic algorithm:
3.2 Multi-classification support vector machine (SVC) for randomly executed sequences

3.2.1 Standard SVM. The standard SVM [9] According to the distance from the point to the straight line in geometry, it corresponds to the high-dimensional hyperplane to solve the dichotomy problem:

\[ d = \frac{|w^T x + b|}{||w||} \]  

(3)

Where \( ||w|| \) is the 2-norm of the weight vector of each dimension, \( b \) is a constant, and the objective function is:

\[ \arg \max_{w,b} \left\{ \frac{1}{||w||} \min_n \left[ y_i \left( w^T x_i + b \right) \right] \right\} \]  

(4)

Where \( y_i \) is the distance from the boundary, which is also the classification label of the exponential data points. Meanwhile, formula (4) can be converted into:

\[ \arg \min_{w,b} \frac{1}{2} ||w||^2 \]  

(5)

subject to \( y_i \left( w^T x_i + b \right) \geq 1, \forall i \)

Then, the constrained objective function of formula (5) is converted into the unconstrained objective function by introducing Lagrange method to solve:

\[ L(w,b,\alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \alpha_i \left( y_i \left( w^T x_i + b \right) - 1 \right) \]  

(6)

Where, \( \alpha_i \) is Lagrange multiplier, and is always greater than or equal to 0. Then, relaxation variable \( \xi_i \) is introduced to allow some data to be on the wrong side of the segmentation boundary to reflect the degree of data outliers, and penalty function \( C \) to reflect the importance of outlier data points, namely, regularization:

\[ \arg \min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m} \xi_i \]  

(7)

subject to \( y_i \left( w^T x_i + b \right) \geq 1 - \xi_i, \forall i, 0 \leq \xi_i \)

The corresponding transformation of Lagrange function is:

\[ L(w,b,\xi,\alpha, r) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \left[ y_i \left( w^T x_i + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{m} r_i \xi_i \]  

(8)

At the same time, the gaussian kernel function is introduced by mapping data into high dimensional space for classification [10] That is:

\[ k(\alpha,b) = \exp \left( -\frac{d(\alpha,b)^2}{2 \cdot \sigma^2} \right) \]  

(9)
And let the constant $\gamma = \frac{1}{2\cdot\sigma^2}$ be one of the undetermined parameters.

3.2.2 Model implementation. Since standard SVM can only carry out data binary classification, n classes are given in SVC, and each two classes are used as a classifier, so there are $\frac{n(n-1)}{2}$ classifiers. Each classifier is used to determine the same group of data. In the end, a class with the most determined times is used as the classification label of the final data, namely:

$$\text{class}_i = \arg \max \{\text{class}_{i1}, \text{class}_{i2}, ..., \text{class}_{in}\}$$ (10)

Where $\text{class}_i$ is the final classification label of item data.

For the difference in the allocation of training set and test set, there will be a big impact on the results. Therefore, we introduce the random shuffling of data sequence and the combination of cross validation [11] to find the most suitable assignment as training set and test set.

Against the standard SVM as punishment function C and gaussian kernel function (Radial Basis Function) and constant gamma problem determination, we use genetic algorithm to set C and gamma value within the scope of binary encoding, decoding through cross compiling selection operations such as parameter values gradually close to the optimal solution, and introduces numerical in SVC is the final classification accuracy and merit to the next iteration optimization.

In order to prevent overfitting, we added a correction module. When the new fitness is only slightly higher than the original fitness by less than 0.01, we selected the model with a smaller value.

4. Experimental results and analysis

Under the environment of the Windows operating system, Matlab scientific computing software was used to write the model execution program for the algorithm flow shown in the previous chapter, we randomly selected 3000 respectively comparing experiment data items and 3:1 in training set and testing set of data, SVC select Gaussian kernel function (RBF), the GA population size to choose 100, mutation probability is 0.9, the chromosome coding length select 20, C value range is (0.01,100.0), The gamma range is (0.01,10.0), 10 groups of randomly executed sequence loop optimization were used respectively, and the classification accuracy result of the test set and the optimal fitness result of the corresponding maximum evolutionary generations are 100.

The results of the 3000 data set are shown in figure 2:

![Figure 2. The results of 3000 data. (a)optimum and average fitness. (b)real and predict value.](image)

Compared with unoptimized SVC, KNN and Bagged Tree [12] comparison of algorithm accuracy is shown in table 1:

| Method | Number of Train & Accuracy |
|--------|---------------------------|
| SVC    | 400 | 85.3% | 1200 | 83.7% | 3000 | 78.4% |
5. Summary

Aiming at AQI classification problem, this paper analyzes and deduces the principle and operation process of traditional SVC classification method, and proposes a classification method by selecting the optimal random execution sequence and using genetic algorithm to find the optimal SVC parameters.

At first, this paper collected for 10 years in a British daily average of six major substances affecting the quality of the air content of the original data cleaning, using Pearson correlation variable theory and the characteristics of gaussian filter outliers. Then training set by the genetic algorithm and random execution sequence training iterative search makes SVC classifier with the highest accuracy parameter values and generate the optimal parameters of the model. The final test set to be included in the training good SVC model, this model to some extent, improved the traditional classification method constant parameters is difficult to determine. At the same time, the experimental results show that its classification accuracy on different training quantity sets is 5% higher than that of not optimized SVC and other common classification methods.

Reference

[1] Yang Jinwei, Sun Baolei. Prediction of air pollutant concentration in Pingdingshan city based on gray markov model [J]. Journal of Mathematics in Practice and Theory.2014,44(2):64-70.
[2] Shen jin, Zhong Liuju, He Fang-fang, et al. Development of air quality prediction model based on clustering and multiple regression[J]. Environmental science & technology, V.38(02):63-66.
[3] Tian Jingyi, Fan Zexuan, Sun ihua. Prediction and analysis of air quality based on BP neural network [J]. Journal of University of Science and Technology Liaoning.2015(2).
[4] Yuan G, Yang W. Evaluating China’s Air Pollution Control Policy with Extended AQI Indicator System: Example of the Beijing-Tianjin-Hebei Region[J]. Sustainability, 2019, 11.
[5] UK AIR Information Resource. (2019) https://ukair.defra.gov.uk/data/data_selector_service?Q=1157144#mids
[6] Zhang lu, Quan ting, Liu qing. Using SPSS for relevant analysis[J]. Modern business trade industry,2018,39(3):190-191.
[7] Mu Y, Liu X, Wang L. A Pearson’s correlation coefficient based decision tree and its parallel implementation[J]. Information Sciences, 2018, 435:40-58.
[8] Zeng D, Wang S, Shen Y, et al. A GA-based feature selection and parameter optimization for support tucker machine[J]. Procedia Computer Science, 2017, 111:17-23.
[9] Subasi A. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders[J]. Computers in Biology & Medicine, 2013, 43(5):576-586.
[10] Lee S, Daniels K. GAUSSIAN KERNEL WIDTH GENERATOR FOR SUPPORT VECTOR CLUSTERING[C]// International Conference on Clustering, 2014.
[11] Anonymous. Application of support vector machine based on grid search and cross validation in implicit random scheduling of cascade hydropower system [J]. Electric Power Automation Equipment, 2014, 34(3):125-131.
[12] Aburomman A A, Reaz M B I. A novel SVM-KNN-PSO ensemble method for intrusion detection system[J]. Applied Soft Computing, 2016, 38(C):360-372.