Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230810

Water Pollutant Classification Method Based on Multi-Class Support Vector Machine

Hengming LIU[a1], Yuwen WANG[b2]
[a]*School of Marine Technology and Environment, Dalian Ocean University, Dalian, China*
[b]*School of Environment, Hohai University, Nanjing, China*

Abstract—Aiming at the problem of water pollution classification, a water pollutant classification method based on multi-classification support vector machine is proposed. By constructing and optimizing the coding matrix, a classification coding table and decoding table are formed, and SVM sub-classifiers are used for data classification. The classification experiment was carried out on the measured water quality data of the sewage treatment plant. The results show that, compared with other traditional classification methods, this method has a higher classification accuracy rate, greatly reduces the number of sub-classifiers required, and improves the classification efficiency.

Keywords—multi-class SVM; water pollutant; classification method

1. Introduction

Water pollution is a serious global problem that poses a major threat to human health, ecosystems and economic development. Accurate classification and timely identification of water pollutants are crucial for effective water resource management and pollution control. Traditional water pollutant classification methods mainly rely on chemical analysis and manual detection, which is time-consuming and laborious, and the accuracy is not high. In recent years, with the continuous development of machine learning technology, various algorithms have been widely used in water pollutant classification, such as neural network, decision tree, random forest, naive Bayesian, support vector machine, K nearest neighbor, etc. However, these methods have more or less limitations, such as complex classifier structure, low classification accuracy, long processing time, prone to overfitting, and the result cannot be guaranteed to be the global optimum. Therefore, we need a new method to solve these problems. In this paper, we propose a water pollution classification method based on multi-class support vector machines. Firstly, by reconstructing and optimizing the Error Correcting Output Code (ECOC), the classification coding matrix is constructed, and then the SVM the sub-classifier divides water pollutants into multiple categories, and finally the validity and accuracy of the method are verified by experiments, and compared with other classification methods.

[1] Corresponding Author: Hengming LIU, School of Marine Technology and Environment, Dalian Ocean University; e-mail: 963028347@qq.com
[2] Yuwen WANG, School of Environment, Hohai University; e-mail: 973331413@qq.com

2. Current Status of Domestic and Foreign Research

In recent years, with the development of various classification algorithms and the continuous improvement of data collection technology, water pollution classification research has gradually become popular and mature, and many remarkable results have been achieved. Abuzir[1] used J48, naive Bayesian and MLP models to analyze and compare the classification performance of each model in the case of different feature numbers in the data set for the classification of water pollutants, and proved that the naive Bayesian algorithm can be used in small Effectiveness in sample classification. Manaf[2] used the naive Bayesian algorithm as a classifier to collect and classify the temperature, pH value and turbidity of water quality, and obtained an accuracy rate greater than 96.89%. Koranga[3] analyzed and compared different binary classification and multi-classification algorithms for the water pollution classification problem of Nainital Lake, and found that the random forest algorithm is more suitable for regression prediction, and there are three kinds of stochastic gradient descent, random forest and support vector machine Algorithms are more efficient for data classification. Shakhari[4] proposed a classification method to classify water quality data, and compared it with two existing classification methods (C-4.5 and logistic regression), and the experimental results verified the effectiveness of the method. Ramadhani[5] used an improved K-nearest neighbor algorithm (MKNN) for water quality monitoring and classification in Riau province, Indonesia, with a classification accuracy of 85.1%. Grbčić[6] proposed a method for classifying pollutants in water supply networks based on random forest algorithm, and the proposed method has high accuracy in locating potential pollution sources. Muhammad[7] validated and compared two different kernels of SVM classifier for the increasing rate of water pollution in Malaysia and achieved an accuracy of 91.67% in classifying water quality pollutants. Khullar[8] proposed a deep learning based Bi-LSTM model (DLBL-WQA) for water pollutant identification with better accuracy than traditional methods such as random forest and artificial neural network for the deteriorating water quality in Yamuna River, India. Solanki[9] analyzed and compared the application of deep learning algorithms and other unsupervised learning algorithms for classification of river pollution problems near Nasik, Maharashtra, India. Ladjal[10] proposed a data fusion method using principal component analysis (PCA) combined with support vector machines, artificial neural networks and decision templates for water quality classification and monitoring in the Tilesdit dam area in Algeria. The research found that SVM, The integration of ANN and PCA achieves a classification accuracy of 98%. Yusri[11] combined support vector machine (SVM) and extreme gradient boosting (XGBoost) algorithm to propose a classification algorithm to predict water quality classification (WQC). The number of experimental samples was 2000, and the average classification accuracy rate reached 90%.

Domestically, Guo[12] introduced the ANN method into the field of water pollution classification and control planning for the first time in China. Through the design of the Liapunov energy function structure, he proposed the Hopfield network model for comprehensive water quality evaluation and the BP model for water quality evaluation membership degree, which are used for water quality samples. Classification, which improves the accuracy and reliability of classification results. Liang[13] used the BP neural network improved by the L-M algorithm to train the water quality sample data in view of the uncertainty, high nonlinearity, complexity of the water quality model, and difficulty in solving the water pollution process. The accuracy of the water quality model was higher than that of the traditional method. got improved. Zhou[14] aimed at the

shortcomings of BP neural network that is easy to fall into the local optimal solution and slow convergence speed, using the genetic algorithm with global optimization characteristics to optimize its weight, and correcting the training performance of BP neural network through Bayesian regularization The function solves the overfitting problem of the water quality classification model. Taking the water quality monitoring data of an ecological monitoring area in Liaodong Bay as an example, Zhang[15] used the singular value decomposition of the matrix and the K-means clustering algorithm to classify the water quality monitoring data in the ecological monitoring area, avoiding the pre-set K-means algorithm. The lack of the number of fixed categories improves the classification speed. Jin[16] used the autoregressive model to estimate the water quality background signal, used the K-means clustering algorithm to obtain the response category center of the water quality parameters, and used the phase-based measurement method combined with the SVM multi-classification algorithm to classify and identify pollutants. Ma[17] established a semi-supervised remote sensing water quality inversion model, improved the standard PSO-based support vector machine regression model, and established a semi-supervised regression model (PSSRM) based on particle swarm algorithm support vector machine. Compared with The traditional method obtains higher regression accuracy. Zhou[18] used the Multivariate Bayesian Uncertainty Processor (MBUP) to probabilistically model the relationship between the predicted value of the deep learning artificial neural network (ANN) and the observed value of water quality. Significant improvements have been made in performance and accuracy. Ming[19] put forward the least squares support vector machine model of the water quality total organic carbon index through the classification modeling of the water quality organic matter index, and established the classification model AP-LSSVM based on the clustering algorithm. Compared with the whole data single model method with higher precision. Xin[20] proposed a water quality detection and classification model based on multimodal machine learning algorithms, and designed 15 mainstream water quality detection algorithms based on XGBoost, CatBoost, RF, Naive Bayes, LGBM, etc. The results showed that XGBoost, CatBoost, The LGBM model has a good processing effect on the monitoring of water quality indicators. Wang[21] took the Harbin section of the Songhua River as the research object, studied the identification mode of pollutants entering the river by using computational intelligence methods and other numerical simulation methods, and integrated the Bootstrap method and wavelet technology to establish an artificial neural network model for conventional water quality time series (BWNN), the classification accuracy is higher than ARIMA and other traditional ANN models.

3. Water Pollutant Classification Method Based on Multi-Class Support Vector Machine

To sum up, in the process of water pollutant classification, various classification algorithms have certain advantages, but the disadvantages mainly lie in the high complexity of the model, the phenomenon of over-fitting or under-fitting, and the poor interpretability of the model, the large number of samples required and the low degree of automation of parameter optimization. For example, the structure optimization problem, overfitting problem, and generalization ability problem of the ANN algorithm still need to be further improved. The selection of the number of hidden layers and hidden layer nodes in neural network algorithms still relies on experience and lacks a unified standard. The BP neural network is easy to fall into the local optimal solution and the

convergence speed is slow, and its structural optimization problems, over-fitting problems, and generalization ability problems still need to be improved. In addition, a large amount of water quality sample data is usually required in the process of model training, which requires the classification algorithm to be applicable to the classification of small sample data and to ensure a certain classification accuracy. Based on this, this paper proposes a water pollutant classification method based on multi-class support vector machine.

3.1. Support Vector Machine

Support Vector Machine (SVM) was first proposed by Vapnik[22] et al. as a supervised machine learning algorithm. It uses nonlinear transformation and structural risk minimization principle to convert the binary classification problem into a quadratic programming problem. It has unique advantages in sample, nonlinear, and multidimensional attribute classification, and it seeks the best compromise between the complexity of the algorithm and the ability to classify unknown samples. Compared with other algorithms such as genetic algorithm and artificial neural network, which are plagued by local minimum solutions, the support vector machine obtains the global optimal solution, has strong generalization ability, and has a good performance while ensuring classification accuracy. Stability[23], has been widely used in various fields such as classification, regression, recognition and prediction.

Suppose the sample data set to be classified is $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where n is the number of samples, $x_i \in R^P (i = 1,2, \ldots, n)$ is the i-th training sample, R^P is a P-th dimensional real vector, P is the dimension of the sample, and $y_i \in \{+1, -1\}$ is the category attribute of the training sample x_i. Then the linear discriminant function in P-dimensional space is:

$$g(x) = w \cdot x + b \tag{1}$$

In equation(1) $w \cdot x$ is the inner product (dot product) of w and x, w is a P-dimensional real vector, and b is a real number.

Then the classification function $f(x)$ is:

$$f(x) = sgn\{g(x)\} = sgn\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\} \tag{2}$$

Where sgn is a symbolic function; when the sample data set X is linearly separable, $K(x_i, x)$ is a linear function; when X is linearly inseparable, $K(x_i, x)$ is a kernel function; α_i^* and b^* is the optimal solution after optimization.

3.2. Multi-class SVM method

Since the support vector machine is a binary classifier, when it is applied to the field of multi-classification, the strategy is to divide the sample space by constructing multiple binary classifiers and then realize the classification of each category. Commonly used methods mainly include:

(1) The One-Against-All (OAA) method requires a total of k SVM sub-classifiers for multi-classification of k categories. This classification method is simple in principle and fast in classification speed. But its disadvantage is that there are classification overlapping regions and indivisible regions, and there is sample imbalance in the training

stage, which will affect the classification accuracy of SVM.

(2) The One-Against-One (OAO) method requires a total of $k(k-1)/2$ SVM sub-classifiers for multi-classification of k categories. This classification method uses only two class samples each time to construct a classifier, which overcomes the problem of imbalance between positive and negative samples, and its classification performance is better than that of the OAA method. Its disadvantage is that as the number of classification categories increases, the number of SVM sub-classifiers required increases by the order of magnitude, and when the number of classification categories is large, the amount of calculation will increase sharply.

(3) Decision Directed Acyclic Graph (DDAG) method, for the multi-classification of k categories, the DDAG method requires $k(k-1)/2$ classifiers in the training phase, and only $(k-1)$ classifiers in the classification output phase That's it. The classification method does not have classification overlapping areas and inseparable areas. Compared with the OAO method, the DDAG method has a higher classification speed. Its disadvantage is that there is a risk of error accumulation, especially for the selection of the first-level root node.

(4) The Error Correcting Output Code (ECOC) method performs binary segmentation on the sample data by constructing an encoding matrix with k rows and m columns ($\log_2 k < m < 2^{k-1}-1$), and divides the k categories The multi-classification problem is transformed into m binary classification problems for solution. Each SVM sub-classifier of the ECOC method is judged independently and plays a role. The result of a single classifier will not be transmitted to other classifiers, so there is no error accumulation, and there is no classification overlap and inseparability. In addition, the ECOC method has a certain error correction capability. Assuming that the minimum Hamming distance between each row of the coding table is d, at most $(d-1)/2$ classifier errors can be allowed[24].

3.3. The Method of this Article

Based on the application scenario of the SVM multi-classification method, this paper refers to the idea of literature[25], and according to the actual situation of water pollution classification, matrix coding is performed on the SVM multi-classification method, thus forming a coding table and a decoding table. The coding principles are shown in Table 1.

Table 1 The ECOC-SVM encoding(decoding)

Category	SVM_1	…	SVM_j	…	SVM_m
1	T_{11}	…	T_{1j}	…	T_{1m}
…	…	…	…	…	…
i	T_{i1}	…	T_{ij}	…	T_{im}
…	…	…	…	…	…
k	T_{k1}	…	T_{kj}	…	T_{km}

In the training phase, data encoding is first performed on the input sample data. The encoding here refers to encoding the category of the sample. The i-th category corresponds to the binary number composed of elements in the i-th row of the coding table, and the j-th column corresponds to the first the output of j-th SVM the sub-classifiers on each sample, the i-th sample belongs to the positive class or negative class on the j-th sub-SVM is determined by the value of T_{ij}. Each sample is sequentially sent

to m SVM sub-classifiers, and after all training samples are trained, the parameters of each SVM sub-classifier are obtained. In the classification stage, the water quality samples to be classified are sequentially sent to m SVM sub-classifiers for discrimination, and the m discrimination results form a binary number with a code length of m, and then compared with the binary numbers corresponding to the k rows of the coding table, the same or is the category corresponding to the row. If the classification output is different from each row, the row with the smallest distance[26] (Hamming distance) to it is selected as the output.

4. Experiment Analysis

The experimental environment is: computer with 64-bit Windows 10 operating system, memory 16GB, CPU Intel Core(TM) i5-8500, main frequency 3.0GHz. The calculation software is Matlab R2019a.

The experimental data uses the water quality data of the water inlet measured by a sewage treatment plant in Dalian. There are 876 sample data, of which 500 data are used for classification model training, and the other 376 data are used for classification testing. The water quality parameters use seven indicators: BOD_5, COD, SS, pH, TN, NH_3-N, and TP. In this paper, the classification test is carried out on the above data, and compared with other classification methods, the experimental results are shown in Table 2.

Table 2 Classification accuracy of different methods

Classification method	Classification accuracy
decision tree	85.23%
k-nearest neighbor	82.64%
artificial neural network	88.07%
naive bayes	80.98%
multi-class SVM	92.13%

5. Conclusion

From the experimental results, it can be seen that compared with other traditional methods, the method proposed in this paper has a higher classification accuracy and has the following advantages:

(1) The number of SVM sub-classifiers required by the classification algorithm is small, and the amount of calculation required is small, while maintaining a high classification accuracy.(2) The number of pollutant types to be classified is easy to adjust. It is only necessary to expand the ECOC code, recode the code, and only add a small number of SVM sub-classifiers to realize the classification of more pollutant types, or for the same type Pollutants are classified in a more granular degree of pollution.(3) The number of water quality samples required is small, and SVM has advantages in the classification of small sample data.

The deficiencies that exist are:

(1) The feature selection of pollutant samples mainly relies on experience. The next step is to study how to perform automatic sample feature selection.(2) The water quality sample data comes from a single sewage treatment plant. Whether the classification effect is applicable to other conditions and environments remains to be further verified.

References

[1] Abuzir S Y, Abuzir Y S. Machine learning for water quality classification[J]. Water Quality Research Journal, 2022, 57(3): 152-164.

[2] Manaf K, Kaffah F M, Mulyana E, et al. Implementation of Naïve Bayes algorithm in IoT-based water cleanliness monitoring system[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2021, 1098(4): 042007.

[3] Koranga M, Pant P, Kumar T, et al. Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand[J]. Materials Today: Proceedings, 2022: 1706-1712.

[4] Shakhari S, Banerjee I. A multi-class classification system for continuous water quality monitoring[J]. Heliyon, 2019, 5(5): e01822.

[5] Ramadhani D, Afdal M, Rahmawita M. The Classification Status of River Water Quality in Riau Province Using Modified K-Nearest Neighbor Algorithm with STORET Modeling and Water Pollution Index[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1783(1): 012020.

[6] Grbčić L, Lučin I, Kranjčević L, et al. Water supply network pollution source identification by random forest algorithm[J]. Journal of Hydroinformatics, 2020, 22(6): 1521-1535.

[7] Muhammad Z, Jailani N A J, Leh N A M, et al. Classification of Drinking Water Quality using Support Vector Machine (SVM) Algorithm[C]//2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE). IEEE, 2022: 75-80.

[8] Khullar S, Singh N. Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation[J]. Environmental Science and Pollution Research, 2022, 29(9): 12875-12889.

[9] Solanki A, Agrawal H, Khare K. Predictive analysis of water quality parameters using deep learning[J]. International Journal of Computer Applications, 2015, 125(9): 0975-8887.

[10] Ladjal M, Bouamar M, Brik Y, et al. A decision fusion method based on classification models for water quality monitoring[J]. Environmental Science and Pollution Research, 2023, 30(9): 22532-22549.

[11] Yusri H I H, Hassan S L M, Halim I S A, et al. Water Quality Classification Using SVM And XGBoost Method[C]//2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC). IEEE, 2022: 231-236.

[12] Guo Jinsong. Research on water quality evaluation and water quality simulation based on artificial neural network (ANN) [D]. Chongqing University,2002.

[13] Liang Nan. Water quality prediction based on artificial neural network and MATLAB implementation[D]. Changan University,2007.

[14] Zhou Rongrong. Research on Total Water Pollution Control in Jiaozhou Bay Nearshore Waters Based on ANN and Genetic Algorithm[D]. Ocean University of China,2009.

[15] Zhang Limei, Wanli. Mathematics in Practice and Theory,2014,44(20):140-147.

[16] Jin Yu. Research on the classification method of characteristic pollutants of water supply pipe network based on conventional water quality parameters[D]. Zhejiang University,2017.

[17] Ma Lei. Semi-supervised learning-based quantitative remote sensing of water quality in the Wei River[D]. Shaanxi Normal University,2010.

[18] Zhou Y. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques[J]. Journal of Hydrology, 2020, 589: 125164.

[19] Ming Qian. Study on the application of classification modeling method in spectral water quality analysis[D]. Zhejiang University,2013.

[20] Xin L, Mou T. Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification[J]. Wireless Communications and Mobile Computing, 2022.

[21] Wang Wei. Research and application of digital simulation of river water quality management based on computational intelligence method[D]. Harbin Institute of Technology,2013.

[22] Vapnik V.N.Statistical Learning Theory[M].Wiley-Interscience,New York:1998.

[23] Xu Peng, Liu Qiong, Lin Sen. Research on internet traffic classification based on support vector machine[J]. Computer Research and Development,2009,46(3):407-414.

[24] Liu Xiaofeng, Zhang Xueying. Application of support vector machine based on error correction output coding in speech recognition[J]. Journal of Taiyuan University of Technology,2006,42(1):34-37.

[25] Wang Hongzhi, Liu Zhen, Li Donghui. A network traffic prediction method based on multi-class support vector machine[J]. Science and Technology Review,2014,32(17):60-63.

[26] Wang Hongzhi. Research on Network Traffic Classification Based on PCA Feature Selection and Optimized ECOC[D]. Dalian Jiaotong University,2014.