Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 9, No. 1, 2024

Water Quality Image Recognition based on SVM

Xinpeng Qin

School of computer, Guangdong University of Petrochemical Technology, Maoming Guangdong, 525000, China

Abstract: It is proposed a method that combines digital image processing technology with machine learning algorithm. Firstly, python is used to extract the features of water sample image, the data were pre-processed, divided into training and test sets, and the SVM is used to construct water quality classification models. In the work of classifier evaluation, we need to combine the classifier model performance evaluation indicators: accuracy, Recall, F-value. According to the test results, the accuracy of SVM is more than 90%.

Keywords: Water Quality Image; Color Moment; SVM.

1. Introduction

For aquaculture, it is important to choose non-polluted waters. By observing the color of the water, experienced fishermen can make a preliminary determination of whether the water quality is suitable for fish growth. Water quality assessment is not only based on the current water quality, but also needs to consider the potential drift and release of harmful substances in order to predict the trend of water changes in the future, so as to take preventive measures in advance. At present, the on-line monitoring system for ammonia nitrogen, temperature, reduction potential, PH and dissolved oxygen has become an important means of water quality assessment. Based on the existing standards of aquaculture water quality and the effect of different water quality conditions on aquatic biological activities, the correlation and influence degree of water quality indexes on the overall evaluation of water quality can be deeply discussed. In the traditional aquaculture environment, the naked eye observation and the experience judgment always occupy the dominant position, although this kind of method has the certain practicability, but its observation result is often influenced by the subjectivity and the instability, easy to cause observational bias. Therefore, it is urgent to seek a more stable and objective Quantitative analysis method for its wide application. With the rapid development of computer vision technology, using machine vision technology to replace artificial vision has become the first choice in many areas of industrial production, which has also brought new technical support for aquaculture.

In recent years, with the progress of science and technology, water quality model research more mature, various models emerge one after another. Among them, WASP model, EFDC model, Quasar Model [1] and Mike Series model all provide strong support for water quality prediction and control in different degrees. In addition, the modeling method based on artificial neural network [2] also provides a new idea for water quality prediction.

Compared with foreign countries, the research on water quality model started a little later in China. However, since the early 1980s, scholars in our country have put forward a joint model of transport and transformation in the study of water environmental capacity of heavy metals in Xiang River, which marks the beginning of the study of water quality models in China.

At present, the methods of water quality prediction can be divided into five categories: mathematical statistics prediction, grey system theory prediction, neural network model prediction, water quality simulation model prediction and chaotic theory prediction [3]. These methods can be divided into explicit relation and implicit relation when establishing the mapping relation between water quality parameters and monitoring data [4]. The explicit relation is mainly realized by linear model, which is easy to operate, but its reliability needs to be improved [5], while the implicit relation such as artificial neural network is faced with the problems of network structure determination, over-fitting and under-fitting [6]. In order to solve these problems, the statistical learning theory [7] and its Support vector machine [8] methods have gradually been paid attention to and become one of the hot spots in the current research of water quality prediction.

In this thesis, we first use Python to Data pre-processing the water sample image, then convert the image information into data information [9], and select the parameters of color features, whether the color moment [10] is selected properly is evaluated. The second is to develop Support vector machine models for classifying and predicting water quality samples. This article includes the following sections:

(1) Image collection and data set formation: firstly, the image of water quality sample is collected and a part of it is selected as modeling data set.

(2) Data pre-processing: after the model data set and the incremental data set are obtained, the image is preprocessed. This step mainly includes image cutting and color moment feature extraction, aiming at extracting the key information related to water quality evaluation from the image.

(3) Classification model construction: using the above-mentioned information, a water quality classification model of the Support vector machine is constructed, which will be used as a basis for water quality assessment and will be used for classification and evaluation of real-time extracted water quality images.

2. Image Preprocessing

When using color features to recognize water quality image, the color information carried by color image is especially rich, which can describe the surface color features of water samples more accurately. Therefore, in order to extract the water color features accurately, the first task is to segment the water color image from the complex background. However,

when water samples are actually taken, they are usually taken together with the glass container containing the water. As the color of the glass container is significantly different from that of the water body, and the water body is often located in the center of the image, a method is needed to eliminate the interference of the glass container and background noise in order to extract the representative image of the central region of the water sample.

In order to achieve this goal, this paper uses an image cutting technology. First, the size of the original image is determined to be m × N, and then a 101 × 101pixel sub-image is truncated from the center of the image. This sub-image is captured from fix (m/2) -50 pixels to fix (m/2) + 50 pixels, both in length and width. The Fix (x) function here means to round off to zero, making sure you are capturing the center of the image.

By using this method, the interference of glass container and background noise can be removed effectively, and the central region of water sample image can be more clear and accurate. The water sample image after cutting not only retains the key information of the original image, but also is easier to extract and analyze the following features. In short, this step is to precisely cut the original water sample image (Figure 1) and save it as the cut water sample image (Figure 2), which lays a solid foundation for the subsequent water quality evaluation work.

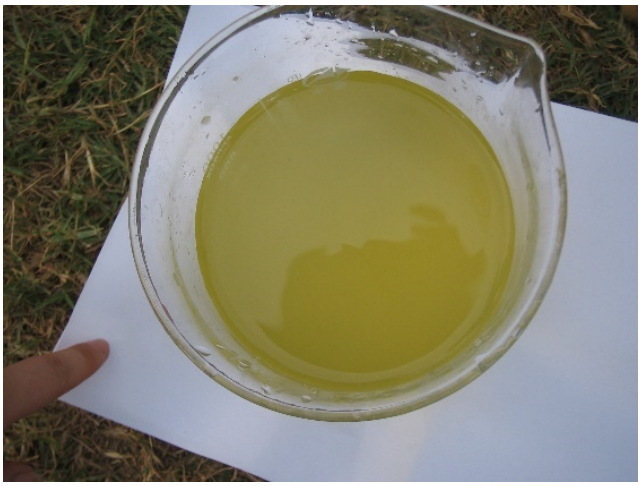

Figure 1. Original image of water quality

Figure 2. Truncated sub-image

In the research of image processing, recognition and classification, color, as the most intuitive visual feature, is very important for the description of image content. Effective expression and extraction of color features become the core challenge of these studies. In order to characterize color features, many methods are proposed, such as color histogram, color aggregation vector, color moment and color set. Because of its wide application, the color moment is chosen as the main method to study the color characteristics of water samples in this paper. Based on the principle of probability theory, every moment of any variable can describe its probability distribution. Similarly, the color layout in an image can also be expressed by its moments of order, which can be regarded as the probability distribution of color in the image. The color moments include the first-order moments (mean), second-order moments (variance) and third-order moments (skew) of the three color channels R, G and B.

In this paper, the feature of water sample image is extracted by extracting color matrix:

(1) first-order color moment: the average value of each pixel color channel color value, reflects the overall intensity of the image.

$$\mathrm{E_i} = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \tag{1}$$

It represents the first-order color moment of the i th color channel and p_{ij} represents the color value of the i th color channel of the j th pixel.

(2) second-order color moment: the variance calculation method is used to reflect the range of image color distribution.

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^2} \tag{2}$$

In the above expression, σ_i represents the second-order color moment of the i th color channel, and $\mathrm{E_i}$ represents the first-order color moment of the i th color channel.

(3) the third-order color moment is calculated by the slope, which reflects the symmetry of the image color distribution.

$$s_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^2} \tag{3}$$

In the above expression, s_i represents the second-order color moment of the i th color channel, and $\mathrm{E_i}$ represents the first-order color moment of the i th color channel.

A total of 197 water quality images were collected by digital camera from several aquaculture ponds and lakes in a city on the internet. To facilitate batch processing and analysis, all images are named in accordance with the unified naming rules, specifically “Category number. JPG.”.

In the field of image processing, image features usually include shape features, spatial relationship features, color features and texture features. Among them, the color feature shows strong robustness because of its insensitivity to the size and direction of the object. In view of the uniformity of the water color image in this paper, the color features of the water color image as the main basis for water quality classification are mainly concerned. In order to identify the images of different water quality categories accurately, expert knowledge was introduced and the water quality categories

were divided into five categories: light green, Gray blue, yellow brown, Brown and green. These different colors represent different water quality conditions. Through machine learning technology, the goal is to let the computer automatically identify and classify these images, so as to replace experts to judge water quality. The specific water quality classification criteria are shown in Table 1, detailing the color characteristics and possible actual water quality status for each type of water quality.

Table 1. water quality categories

Water color	Water quality
Light Green	1
Gray-blue	2
Tawny	3
Tawny/turmeric	4
Green/Yellow-green/oil-green	5

It is used python language in programming implementation. The features of all photos were extracted through a for loop, resulting in 197 rows and 11 columns. Some of the data results are shown in Table 2.

Table 2. Color feature data

Class	No.	R_1	G_1	B_2	R_2	G_2	B_2	R_3	G_3	B_3
1	1	0.58276	0.543685	0.252426	0.014247	0.01619	0.041173	-0.01263	-0.01606	-0.0414
1	10	0.64187	0.570773	0.213572	0.015429	0.011138	0.013735	0.009572	-0.00399	-0.00203
1	11	0.603698	0.576798	0.282196	0.008728	0.007102	0.0123	-0.00466	-0.00234	-0.00955
1	12	0.589696	0.593718	0.252221	0.00799	0.005967	0.010698	0.003726	-0.00361	-0.00507
1	13	0.59113	0.592068	0.253413	0.007524	0.00653	0.0122	-0.00136	-0.00184	-0.00431
1	14	0.588769	0.569681	0.318891	0.00758	0.005039	0.008497	-0.0035	0.001897	-0.00572
1	15	0.588507	0.573246	0.288332	0.007371	0.005135	0.011889	0.003095	-0.00109	0.008591
1	16	0.61813	0.539698	0.274516	0.008626	0.006167	0.014641	-0.0025	0.003659	0.014072
1	17	0.618471	0.539224	0.273564	0.0089	0.005845	0.014155	0.003123	0.003313	0.014492
1	18	0.664273	0.576335	0.293741	0.011002	0.007727	0.019136	-0.00514	0.004122	0.012665

Nine color moment parameters are extracted from each sample and used as input vectors for three different algorithms. The specific values of these characteristic parameters have been listed in Table 3 in detail 36 different grades of water quality mean moment values.

Table 3. mean moments of color for different water quality categories

Mean	**R_1**	**G_1**	**B_2**	**R_2**	**G_2**	**B_2**	**R_3**	**G_3**	**B_3**
Grade 1	0.610204	0.574785	0.268092	0.009331651	0.006864815	0.012292184	-0.000988557	-0.001327807	-0.000304472
Grade 2	0.524593	0.53524	0.336173	0.010272269	0.008091553	0.012270426	0.002926114	0.002146608	0.004036228
Grade 3	0.540234	0.521514	0.199152	0.009416922	0.007167669	0.012705104	3.24976E-05	-0.001075195	0.000387175
Grade 4	0.450311	0.460043	0.200711	0.010082893	0.007277105	0.012888535	0.002815967	0.000520603	0.001396737

3. SVM Algorithm

Based on the statistical learning theory, Support vector machine support vector machine (SVM) has been developed as an excellent classification model in recent years, which opens up a new way for the study of water quality assessment. As a machine learning technique designed for classification problems, SVM not only has a clear structure, but also shows strong generalization ability in many data models. Especially when dealing with small sample data, SVM can effectively avoid the common “over-fitting” problem in the neural network, and prevent the model from falling into local optimum to ensure that the solution is global optimum. At present, SVM algorithm has been widely used in pattern recognition, time series analysis, cluster analysis and other fields

When dealing with multi-class classification problems, SVM provides many methods such as “One-to-one”, “One-to-remainder” and “Decision tree”. Among them, “Decision tree” method is the first choice in this paper because of its high classification accuracy, low learning training cost and reduce repeated training samples.

The core of SVM classification model is the principle of structural risk minimization, which aims at maximizing the distance of classification boundary to achieve accurate classification. When processing data, it is necessary to distinguish between linear separable and linear indivisible. For linearly separable data, SVM realizes classification by finding the optimal hyperplane which can divide the training set correctly and has the largest geometric interval, while for linearly indivisible data, then, a nonlinear mapping algorithm (kernel function) is used to map the samples from low-dimensional space to high-dimensional feature space, which is transformed into a linear separable problem and then solved by linear classification method.

Suppose we randomly select a given training sample set from an unknown cumulative distribution function, $\{(x_i, y_i), i = 1, 2, \cdots, l\}, x_i \in R^d, y_i \in R$, in order for the classification hyperplane to correctly classify all samples and to maximize the classification interval, the constraint conditions are:

$$\left.\begin{array}{ll} w \cdot x_i + b \geq +1 & for \quad y_i = +1 \\ w \cdot x_i + b \leq +1 & for \quad y_i = -1 \end{array}\right\} \Leftrightarrow y_i (w \cdot x_i + b) - 1 \geq 0 \quad (4)$$

x_i is the input value; w is an adjustable weight function; b is a threshold value; $w \cdot x_i$ represents a vector w and x_i of the inner product.

When the sample classification data is linearly separable, the classification interval is:

$$\min \frac{w \cdot x_i + b}{\|w\|} - \max \frac{w \cdot x_j + b}{\|w\|} = \frac{2}{\|w\|} \tag{5}$$

The problem of solving optimal hyperplanes can be transformed into the problem of constrained optimality, which can be solved by establishing a dual form of Lagrangian:

$$w(\alpha) = \sum_{i=1}^{1} a_i - \frac{1}{2} \sum_{i,j=1}^{1} a_i a_j y_i y_j (x_i \cdot x_j) \tag{6}$$

The constraints are:

$$\alpha_i \left\{ \left[\left(w \cdot x_j \right) + b \right] y_i - 1 \right\} = 0 \tag{7}$$

Further, the constraints may be expressed as follows:

$$\begin{cases} y_i \left[\left(w \cdot x_i \right) + b \right] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}, i = 1, 2, \cdots, l \tag{8}$$

The objective function is:

$$L(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \tag{9}$$

ξ is a relaxation variable; C Is a constant, and $C > 0$. It is mainly used to balance the complexity of the machine and the number of non-separable points and to control the penalty degree of the wrong sample.

In the process of constructing SVM classification model, a series of key steps are followed, these steps include the design of classifier, the selection of training and verification samples, the selection of kernel functions and parameters, the construction of training model and the verification of model. The reasonable choice of kernel function and parameters is the key to the success of the model. Given the distributed nature of the data, it is important to choose the appropriate kernel function, because different kernel functions are suitable for different scenarios. The widely accepted radial basis function (RBF) is used as kernel function in this study. In order to determine the optimal combination of penalty coefficient c and parameter γ of RBF kernel function in SVM classification model, cross-validation method is used, and four-fold cross-validation (V = 4) is chosen to solve the optimal value of C and γ. This method ensures the accurate selection of model parameters and improves the classification performance of the model.

4. Experimental Results

Classification model evaluation is an important method to measure the performance of classification model in machine learning. When there are many kinds of algorithms and model parameters, it can help to select the most suitable model for water quality classification by using different evaluation indexes, such as accuracy, accuracy, recall, F_1 score, etc. Recall ratio: the ratio of correctly classified positive cases to actual positive cases; precision ratio can predict the result, which is the ratio of correctly classified positive cases to actual positive cases; Values are harmonic mean for accuracy and recall.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

$$F_1 = \frac{2 * PRE * REC}{PRE + PEC} \tag{12}$$

TP represents positive samples predicted by the model to be positive, i.e., the number of levels correctly identified; FP represents negative samples predicted by the model to be positive, i.e., the number incorrectly identified to be of other levels; FN is predicted by the model as a negative positive sample, that is, incorrectly identified as a scaleless number.

The data set of 197 pictures is divided into training set and test set, 70% of which is training set and 30% of which is test set. Test results of the water quality image data in the SVM are shown in Table 4 below.

Table 4. SVM model classification

Water quality grade	Accuracy/%	Recall rate/%	F 1 score/%
1	100	91	95
2	95	100	97
3	90	90	90
4	90	90	90

As can be seen from Table 4, the Support vector machine model classification results show that the accuracy rate is above 90%, that is, the probability of correct prediction is 90% for all predicted samples, and the accuracy rate is above 90% for all predicted samples, the recall rate is more than 90%, indicating the proportion of positive samples predicted, which can be correctly predicted. The F 1 value is a comprehensive evaluation index.

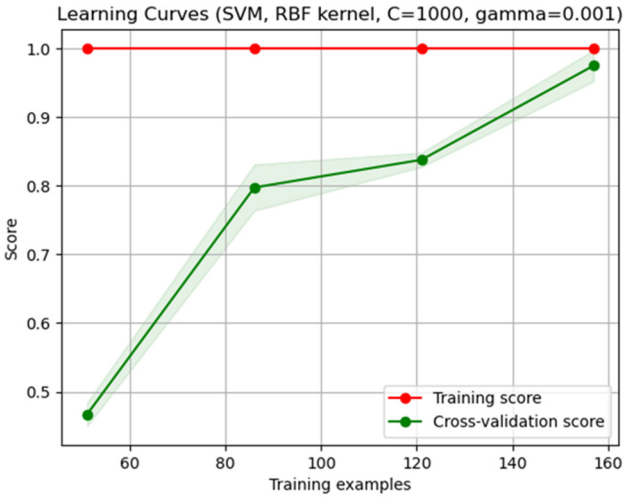

Figure 3. The learning curve of SVM

Furthermore, in order to test the performance of SVM algorithm, the water quality level is used as a label, and the Support vector machine model is trained to ensure that sufficient data are available for training and testing at each

compromise. Therefore, the average accuracy of 40% cross-validation of training set data can reach 98%. The learning curve was used to visualize the model fit (Figure 3).

It can be found that the accuracy of the model to the training data has been stable at 100%, and with the continuous input of the training data, its prediction accuracy for cross-validation has been increasing, finally, the convergence state of the training fitting degree is reached. As a result, the Support vector machine model can fit the water quality data well and there is no problem of underfitting or overfitting.

5. Conclusion

In this paper, we design a computer vision-based image analysis method of water quality, combined with the actual water quality samples, the image processing algorithm and data processing are programmed in Python language. The image cutting technology is used to cut the image, and the image processing algorithm is used to extract the color moments of water samples and get the data. The experimental results show that the accuracy of the Support vector machine model can reach more than 90%, and it is effective to use machine learning technology to evaluate the water quality of water color images. Through this technology can avoid the subjective observation error due to eye observation, and save a lot of manpower and material resources, simple and practical easy to promote the application.

References

[1] Wang Haiyan, Li Jianhui, Yang Fenglei. Review of Support vector machine theory and algorithms[J]. Computer should research, 2014,31(5): 1281-1286.

[2] Han Ding, Wu Pei, Zhang Qiang, etc. Feature extraction and image recognition of typical grassland pasture based on color moments[J]. Agricultural Engineering, 2016,32(23): 168-175.

[3] Rita Wong. A Feedforward neural network predictive classifier[J]. World of Digital Communications, 2021, (09): 67-69 + 81.

[4] Peric Lee, Lau chi-fong, Cheung Yuk-fuk. Study on water quality evaluation method based on machine learning [J]. Infocomm, 2019, (01): 97-98.

[5] du Yanqi. Application of Python image processing technology in water quality evaluation [J]. Innovation in science and technology, 2020, (31): 189-190.

[6] Wang Qiwei. Image histogram features and their applications [D]. University of Science and Technology of China, 2014.

[7] Zhou Yan, Liu, Zhou Can. Surface water quality assessment based on decision tree Support vector machine[J]. Environmental Science and technology, 2021,34(05): 57-61 + 66.

[8] Wang Tianwang, Jie liming,etc. Research on tobacco color classification method based on improved robust multi-classification SVM [J]. Electromechanical information, 2021, (05): 55-60.

[9] Zeng Chuanhua, Chen Hong, etc. Bamboo color grading method based on SVM and color moments[J]. Hubei agricultural science, 2010,49(02): 455 -457.

[10] he Heng, song die, Wang Yi Xu. Construction of water quality assessment model for Huzhou and key feature recognition based on machine learning[J]. Water Resources Development and management, 2023,9(01): 57-64.