Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 10 (2011) 451 – 457

Procedia Environmental Sciences

2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011)

# A Method of Water Quality Assessment Based on Biomonitoring and Multiclass Support Vector Machine

Yue Liao, Jianyu Xu*,Wenjing Wang

*Institute of Information Science and Technology, Ningbo University, 818 Fenghua Road, Ningbo 315211, P.R. China*

**Abstract**

Integrating biological monitoring method with computer vision technology, acute toxicity test was performed to study the toxic effects of $Cu^{2+}$ with different concentration on zebrafish (*Danio rerio*). The Behavioral response of school of fish in a tank was quantified and an early warning system was developed in the study. In the system, the real-time quantified data were saved to database and the Multiclass Support Vector Machine (SVM) was used to make comprehensive assessment according to behavioral difference of fish school in different toxicity environment. The prediction accuracies are satisfactory, which indicate that this approach is effectual for assessment of water quality.



*Keywords*: biological monitoring; water quality assessment; Support Vector Machine (SVM); Multi-class classification.

## 1. Introduction

Water resources protection is one of the important contents for man to realize sustainable development. During the past years, water pollution incidents have frequently happened and water security has received common attentions. How to assess water quality is becoming a study focus. At present, two main methods have been used to evaluate water quality: one is physical and chemical analysis, the other is biological monitoring method. The biological monitoring method is an effective way for water quality assessment by monitoring the change of health status, physiologic characteristics and behavioral responses of aquatic organisms individual or population. Compared with traditional method of physical

* Corresponding author. Tel.: +86-574-87600945.
*E-mail address*: xujianyu@nbu.edu.cn.


doi:10.1016/j.proenv.2011.09.074

and chemical analysis, this method has more advantages. It has the ability of long-time on line monitoring the water quality, which the physical and chemical analysis cannot. Besides, the response of aquatic organisms to water quality is more sensitive and reliable. For the mixed pollution, the method can also make a positive effect.

To give a method to analyze water quality parameters and extract useful information, there were some studies [1,2] using Artificial Neural Network (ANN) to evaluate the water quality directly, without taking advantage of the indicator organism in water quality monitoring. Compared with ANN model, an approach combining multiclass SVM with biomonitoring was proposed in this paper. It can overcome the defects of slow training speed, poor network generalization and low learning accuracy in traditional ANN method. Furthermore, it can get more effective and reliable parameters from indicator organism quickly by using computer vision technology. In order to verify the validity of the method, acute toxicity test was performed.

## 2. Acute toxicity test

### *2.1. Experimental system*

The system shown in Fig.1 is mainly composed of an experimental tank (72cm×38cm×12cm length × height × width), a CCD (Charge Couple Device) camera (Microvision MV-VE120SC) and a four-processor computer (Intel Core i5-740). The light source is a 40 W fluorescent bulbs mounted on the ceiling. The tank is made of acrylic with one transparent side and it is divided into three parts by two fish baffles. The fish baffles prevent fish passing and there are many uneven sizes of holes in the baffles to make water flow through with the same speed. Fish can swim freely in the middle part which is the concentrated part of the CCD camera. There is a filtration unit in the right part of tank composed of a suction inlet and an overflow outlet to filter fish waste and other residues. An 8-watt pump can draw the water from the right side to the left side to form a water circulation. To simulate the real condition, the fish survive in the flowing water and the toxicant is infused into the whole tank within the flowing water. Thermometer, water pump, heating rods and pH meter are placed in the right area to avoid any influence on fish behavior.

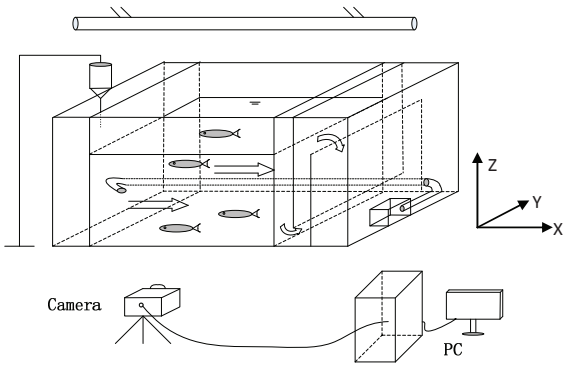


Fig. 1. System schematic diagram

### 2.2. The fish

Zebrafish (*Danio rerio*), of 2 to 3 cm length, weight of about 0.3 g and from either sex, were used. They were obtained from an aquafarm in Ningbo and all the fish were from the same batch. The new acquired zebrafish must be acclimated for 2 weeks in the laboratory to adapt the testing environment. During acclimation, pH of water was maintained at 7.2±0.2 and water temperature was about 23℃. A 12h of light (08:00-20:00) was provided for the zebrafish each day. And the used water had been sufficiently aerated with an air pump. In most of the time, the fish occupied the whole tank and swam actively and randomly in all directions.

### 2.3. Behavioral indices

In this study, some behavioral indices of fish school were quantified to represent water quality situation and pollution degree. As shown in fig.1, the camera faced the transparent side of the tank and images were acquired by the computer vision system. Israeli and Kimmel's methods [3] were used to calculate the coordinates of the center of gravity *CX*, *CZ* of fish school and the spatial standard deviations *SDX*, and *SDZ* in the directions *X*, and *Z*. In the coordinate system (*X*, *Y*, *Z*), *CX, CZ* show the mean location of school in the projection plane and *SDX, SDZ* measure the density of the fish school in two directions. For the activity level of fish school, the method of Xu [4] was adopted to calculate the average swimming speed *AV* of the school of fish.

### 2.4. Experimental preparation and procedures

Before the formal experiments, 24h half-lethal concentration ($LC_{50}$) of $Cu^{2+}$ for zebrafish in the environment was tested. According to the 24h-$LC_{50}$ (1.472mg/L) and water capacity of the tank, the solution of copper sulfate was made. During the experiments, water temperature was kept at 23±0.5℃ and the flow of water was controlled at about 5.5L/min. pH varied between 6.9 and 7.2 in the course. In the experiment, the first monitoring phase lasted about 20 min was a reference monitoring phase. It recorded the state of fish school in normal condition. Then, the second phase lasted 60 min and it recorded the behavioral responses of fish school after the injection of copper sulfate solution. The video sequences of two monitoring phase were sent to image processing module to get further information.

### 2.5. Image Processing

Image de-noising process with mean filter was firstly performed on the video sequences captured by the camera. Secondly, image enhancement was done to improve the contrast of object and background. The intensity image was converted to a binary image. Fragments of feed, bubbles which are smaller than structuring element were removed from the binary image and holes on the fish body were filled in by morphological closing and opening operations. Then, the connected components of the binary image were labeled. At last, the behavioral and location parameters were calculated and saved to database system.

## 3. Multi-class classification with SVMs

A support vector machine is an emerging machine learning technology that has already been used for water quality assessment in the field of environment [5, 6]. Traditional classifiers such as Naive Bayes, BP network and KNN may not work well in the case of limited samples due to the proneness of over-fitting, while SVM does [7]. It is based on the principle of the structural risk minimization in statistic

learning theory, which has good generalization ability. It performs the classification between two classes by finding a decision surface that is based on the most informative points of the training set [8]. Given a training samples $\{(x_1, y_1), (x_2, y_2),..., (x_n, y_n), x_i \in R^n, y_i \in \{-1,1\}\}$, where $x_i$ is the feature vector, $y_i$ is the label, SVM is used to find the optimal classification plane which makes the positive vectors lie on one side and the negative vectors on the other if the samples are linear separable .The plane which makes the margin between two categories maximal is the optimal classification. It can be constructed by solving an optimization problem:

$$\min_{w} \frac{\| w \|^2}{2} \qquad (1)$$
$$s.t\ y_i \cdot (w \cdot x_i + b) \geq 1, \forall i = 1,2,.....N$$

The dual problem is:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \qquad (2)$$
$$s.t \sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i \geq 0$$

If $\alpha_i^*$ is the optimal Lagrange multipliers, the optimal solution is

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \qquad (3)$$

and $b^*$ can be computed by any support vector and formula $y_i \cdot (w \cdot x_i + b) = 1$. So, the binary linear classification function can be written as

$$f(x) = \mathrm{sgn}(w^* \cdot x + b^*) = \mathrm{sgn}\{\sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x) + b^*\} \qquad (4)$$

In some practical applications, the linearly separable condition cannot always be satisfied in most of the case. In this condition, the input $x$ is mapped into a high-dimensional feature space $\phi(x)$ by nonlinear transformation (Fig.2) and constructs an optimal hyperplane to separate the two kinds of data points from two classes. In addition, a slack variable $\xi_i$ is introduced to handle misclassification errors. Therefore, the problem becomes:

$$\min_{w,\xi} \frac{\| w \|^2}{2} + C \sum_{i=1}^{N} \xi_i \qquad (5)$$
$$s.t\ y_i \cdot (w \cdot x_i + b) \geq 1 - \xi_i, \forall i = 1,2,.....N, \xi_i > 0$$

The dual problem becomes:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (6)$$
$$s.t \sum_{i=1}^{N} \alpha_i y_i = 0, C \geq \alpha_i \geq 0$$

Where $K(x_i, x_j)=\Phi(x_i)^T \Phi(x_j)$ is a kernel function under Mercer's condition. There are some commonly used kernel functions:

Linear: $K(x, x_i) = x^T x_i$

Polynomial: $K(x, x_i) = (\gamma x^T x_i + r)^d$, $\gamma > 0$

RBF: $K(x, x_i) = exp(-\gamma \| x - x_i \|^2)$, $\gamma > 0$
Sigmoid: $K(x, x_i) = tanh(\gamma x^T x_i + r)$
Finally, the classification function under non-linear case is

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*\} \quad (7)$$

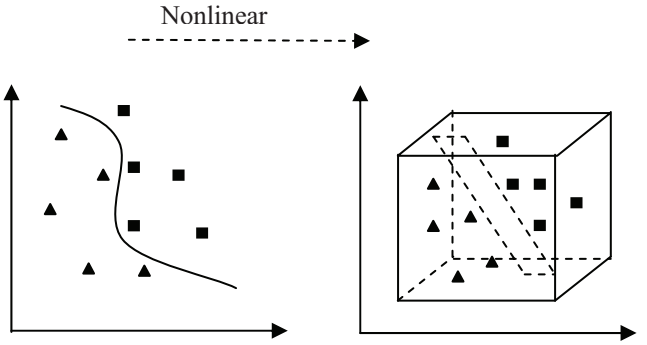


Fig. 2. Nonlinear Mapping

The SVM is originally designed for binary classification, but it can be conveniently used in multi-class classification problems by combining a series of binary classifiers. The main combination methods are one-against-one, one-against-rest and the directed acyclic graph SVM (DAG-SVM) [9]. The basic idea of one-against-one method is to choose different two classes and construct all possible binary sub-classifiers. On the other hand, the one-against-rest method compares a given class with all the others put together. In DAG-SVM, it divides all samples into two broad categories classes and goes on division in each category until there are only two classes left. In the study of Hsu and Lin [10], these methods have been compared. The one-against-one and DAG methods perform well. Because of their higher classification accuracy and faster classification speed, they are more suitable for practical use. However, the construction of DAG consumes more time and the structure of DAG is precarious. In our study, there are only three classes need to be classified. So, it is quite suitable to use one-against-one method, without a lot of classifiers.

## 4. Results and discussion

### *4.1. The effect and toxicity of copper ions*

Acute toxicity tests of 24h-LC50 of $Cu^{2+}$ and a double dose of that were performed individually and the behavioral responses of fish school were recorded. From the experimental observation, it was found that in different toxicity environment the fish had different acute responses. The higher concentration caused a stronger reaction of fish school. According to the conclusion, the multiclass support vector machine could be used in multi-level water quality assessment, taking fish behavioral indices as feature parameters.

### *4.2. Training and testing*

The data sets used in training and testing are shown in table 1. The sample in each data set is composed by normal data points and abnormal data points in two different concentrations. Training set

and testing set were allocated equally in a random manner. After training, the model was used to predict the testing set. The accuracy we got was the highest accuracy with best $c$ and best $\gamma$ in each sample. The best $c$ and best $\gamma$ are penalty parameter and kernel parameter which were acquired by parameter optimization. The prediction results are shown in table 2.

Table 1. General information about the data sets

| Data sets | Total number | Normal | Abnormal Ⅰ | Abnormal Ⅱ |
|---|---|---|---|---|
| 1 | 3073 | 1008 | 1049 | 1016 |
| 2 | 2770 | 800 | 1000 | 970 |
| 3 | 3202 | 1184 | 1006 | 1012 |
| 4 | 3206 | 1136 | 1010 | 1060 |
| 5 | 2632 | 912 | 804 | 916 |

Table 2. Accuracy of training and testing

| No. | Training accuracy (%) | Testing accuracy (%) | Best $c$ | Best $\gamma$ | Accuracy of Class Ⅰ (%) | Accuracy of Class Ⅱ (%) | Accuracy of Class Ⅲ (%) |
|---|---|---|---|---|---|---|---|
| 1 | 86.7925 (1334/1537) | 82.0313 (1260/1536) | 89.0012 | 14.7790 | 87.500 (441/504) | 83.2061 (436/524) | 75.3937 (383/508) |
| 2 | 89.025 (1233/1385) | 86.1372 (1193/1385) | 0.6787 | 23.5854 | 89.250 (357/400) | 87.000 (435/500) | 82.6804 (401/485) |
| 3 | 90.3186 (1446/1601) | 82.2611 (1317/1601) | 4.7285 | 83.9747 | 85.6419 (507/592) | 86.8787 (437/503) | 73.7154 (373/506) |
| 4 | 85.0281 (1363/1603) | 80.3493 (1288/1603) | 27.4383 | 6.0910 | 86.7958 (493/568) | 77.2277 (390/505) | 76.4151 (405/530) |
| 5 | 92.0973 (1212/1316) | 89.8936 (1183/1316) | 15.0898 | 11.3998 | 98.9035 (451/456) | 92.2886 (371/402) | 78.8210 (361/458) |

It can be seen from table 2, the multiclass SVM method provides a high accurate rate in the case. So this method can be used to train the blended data which we have already got from different concentration experiments and use the model obtained to predict possible water quality changes with different degrees.

In this study, the RBF kernel function was chose to use in SVM because of its better ability to deal with the nonlinear relationship between label set and attribute set. Table 3 is a comparison of accuracies with different Kernel functions. The sample comes from the same data set. RBF Kernel function shows higher classification accuracy and faster classification speed.

Table 3. The testing accuracy comparison using different Kernel functions

| Kernel type | Testing accuracy (%) | Time(s) | Total support vector |
|---|---|---|---|
| Linear | 76.0289 (1053/1385) | 0.102791 | 780 |
| Polynomial | 79.3502 (1099/1385) | 0.409717 | 600 |
| RBF | 86.1372 (1193/1385) | 0.252151 | 847 |
| sigmoid | 51.4079 (712/1385) | 0.340593 | 737 |

## 5. Conclusion

In this study, the behavioral parameters of fish school were extracted by using computer vision technology and multi-classification method based on SVM was used to evaluate these parameters, revealing water quality indirectly. As an important model organism, zebrafish is widely used in scientific experiments by researchers. $Cu^{2+}$ pollution is common heavy metal pollution in some waters and the zebrafish shows obvious avoidance response in it. So, instead of traditional water quality assessment methods, we can use the biological monitoring method to give an early warning, which raises the efficiency and reduces the cost of human resources. In our work, the experimental data were processed and the classification accuracy of each experiment was shown in the paper. For three different degree of toxicity, the SVM classifier performs well. According to the different classification results, different alarm signals can be sent out and different emergency measures can be taken to control the pollution and protect the life and property of the people. In our further work, to test the behavioral responses of fish in different environment, many other toxicity reagents will be used in the experiments. Evaluating the joint toxic effects is also a key to study in further.

## Acknowledgements

We appreciate the financial support provided by grant No. 2010A610005 from Ningbo Natural Science Foundation of the People's Republic of China.

## References

[1] Palani S, Liong SY, Tkalich P. An ANN application for water quality forecasting. *Marine Pollution Bulletin* 2008; 56:1586-1597.

[2] Singh KP, Basant A, Malik A, Jain G. Artificial neural network modeling of the river water quality—A case study. *Ecological Modelling* 2009; 220(6):888-895.

[3] Israeli D, Kimmel E. Monitoring the behavior of hypoxiastressed *Carassius auratus* using computer vision. *Aquacultural engineering* 1996; 15, 423-440.

[4] Xu JY, Liu Y, Cui SR, Miao XW. Behavioral responses of tilapia (Oreochromis niloticus) to acute fluctuations in dissolved oxygen levels as monitored by computer vision. *Aquacultural engineering* 2006; 35: 207-217.

[5] Li ZZ, Xie YB, Yi JJ. The Eutrophication Evaluation of Dongting Lake Based on Support Vector Machine. Proceedings of the International Workshop on Information Security and Application, Qingdao, China. 2009, 384-387.

[6] Liu JP, Chang MQ, Ma XY. Groundwater Quality Assessment Based on Support Vector Machine. HAIHE River Basin Research and Planning Approach-Proceedings of 2009 International Symposium of HAIHE Basin Integrated Water and Environment Management, Beijing, China. 2009, 173-178.

[7] Qian HM, Mao YB, Xiang WB, Wang ZQ. Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters* 2010; 31: 100-111.

[8] Vapnik V. *Statistical learning theory*. New York, NY: Wiley; 1998.

[9] Platt JC, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. Proceedings of Neural Information Processing Systems, 1999, 547-553.

[10] Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Network. 2002, 13(2), 415-425.