Environ Earth Sci (2014) 71:475–482
DOI 10.1007/s12665-013-2455-3

ORIGINAL ARTICLE

Identification of algal blooms based on support vector machine classification in Haizhou Bay, East China Sea

Yong Xu · Changchun Cheng · Ying Zhang · Dong Zhang

Received: 24 August 2012 / Accepted: 25 March 2013 / Published online: 5 April 2013

Abstract Harmful algal blooms commonly known as red tides have been observed at increasing frequencies, which are causing serious economic and ecologic problems in Haizhou Bay off the eastern coast of China. It is important to study the inducing factors of red tides including a wide variety of environmental variables and the complex interactions between them. This study explores the possibility of predicting the occurrence of red tides using support vector machine (SVM) with environmental variables. Seventeen in situ environmental variables which are known to affect the occurrence of red tides were collected between May and October of 2004–2006. Seven characteristic factors were extracted from these variables via factorial analysis to reduce computation complexity. Three of them are related to nutrients, others are contributed by temperature, oxygen depletion, pH, hydrodynamics, and precipitation, respectively. The classification models based on SVM were constructed to identify the red tides samples using the seven factors as independent variables and radial basis function as the kernel function. The model with the combination parameters of $C = 10$, $\gamma = 0.7$, and $\zeta = 0.1$ has the highest accuracy of 92.06 %. It indicates that the model is highly valuable in predicting the occurrence of red tides by environmental variables in this region for its conservative threshold of surface algae concentration.

Y. Xu (✉) · C. Cheng
School of Urban and Resources Environment, Yancheng Teachers University, Room 301, No. 1 Building, Middle Daqing Road 29th, Yancheng 224000, China
e-mail: xuyyc@163.com

Y. Zhang · D. Zhang
College of Geography, Nanjing Normal University, Nanjing 210046, China

Keywords Red tides · Factor analysis · Support vector machine classification · Haizhou Bay

Introduction

Harmful algal blooms, commonly known as red tides, are a kind of natural environmental phenomenon (Hodgkiss and Ho 1997) that are causing serious economic and ecologic problems (Yang and Hodgkiss 2004). Red tides have been increasingly observed in coastal waters of Haizhou Bay, the Yellow Sea of China since the 1990s. Terrestrial sourced pollutants cannot be dispersed timely for the area is far away from the Kuroshio (Douglass et al. 2012), it prevents the water exchange between this bay and other areas which is feeble especially in summer. On the other hand, rapid economic development in the shore areas has accelerated the discharge of pollutants into the sea. According to marine pollution baseline investigation in Jiangsu province, the concentration level of dissolvable oxygen in the waters of Haizhou Bay decreased by 62.3 %, while the concentration level of nitrate and nitrite rose, respectively, 10.67- and 18-folds during 1991–1998. Since the beginning of this century red tides have taken place ever increasingly in frequencies, duration, and spatial extent.

Red tides are influenced by a wide variety of environmental variables and their complex interactions. Traditional algorithms are limited in identifying the samples with high-dimensional data. Recently, a relatively new algorithm, support vector machine (SVM), gained extensive application. It was developed from machine learning theory (Vapnik 1982) based on the structural risk minimization (SRM) principle. Burges (1998) reviewed the support vector classification machines and Smola and

Scholkopf (1998) gave a review on the support vector regression machines. It has not only higher accuracy but also a better generalization ability than traditional algorithms (Li et al. 2009); SVM-based approaches of classification achieved much higher (e.g., 92.0 %) overall accuracy than the maximum likelihood classifier (64.8 % overall accuracy) (Sanchez-Hernandez et al. 2007). After comparing SVM with the maximum likelihood classifier, neural network classifiers, and decision tree classifiers, Huang et al. (2002) found it competitive with the best available machine learning algorithms in classifying high-dimensional datasets, such as the environmental variables involved in algal blooms. Therefore, it has been widely used in data mining and knowledge discovery with small samples and complex variables. It aims to decrease uncertainty in the model structure and the fitness of data. SVM has found applications in classifying agricultural crops from multispectral satellite data (Foody and Mathur 2004), predicting toxic activity of chemical industry (Zhao et al. 2006), identifying protein functional class (Cai et al. 2003), predicting blood to brain partitioning behavior (Hassan et al. 2012), forecasting short-term wind speed (Zhou et al. 2011), identifying genes (Bao and Sun 2002), and diagnosing diseases (Zhao et al. 2004). As a means of data analysis, SVM is able to overcome the computational difficulty associated with data of a large dimension. Moreover, it is also able to avoid the local minima and slow convergence problems commonly associated with the neural network method of machine learning. In addition, its performance is independent of the actual distribution of samples.

With the assistance of SVM classification, this study attempts to ascertain the environmental settings conducive to the outbreak of red tides in the Haizhou Bay area in the middle of the Yellow Sea, East China using in situ measured environmental data, including hydrologic and meteorological data. The prediction of red tide outbreaks requires an understanding of the relationship between environmental variables and the occurrence of red tides. The prediction of red tide occurrence is challenging in that an outbreak stems from the complex interactions of physical, chemical, and ecological processes (Cai et al. 2002). Besides, noise in the readings of these environmental variables impedes an understanding of the exact environmental setting. Thus, it is difficult to build a reliable model for the prediction of red tide outbreaks whose mechanism of formation and dissipation is not fully comprehended at present (Gilbert et al. 2007). This study attempts to develop a SVM-based method for understanding the environmental settings under which red tides have occurred. Such knowledge is important to reliably predict outbreaks of red tides in the future.

Research methodology

Algal blooms result from a wide variety of environmental variables that interact with each other in a complex manner. The determination of a bloom and the environmental variables is ideally accomplished via SVM by selecting the most relevant environmental variables.

Selection of characteristic factors

Of all the variables affecting the occurrence of algal blooms, only a few most influential ones can be included in the identification model to keep the complexity of the identification manageable. This reduction is conducive to expediting computation efficiency on the one hand. It also simplifies the structure of the identification model on the other, and improves the reliability of the built model. The reduction of data dimensionality can be effectively accomplished via factorial analysis that retains most of the information on all variables. It aims to extract a few characteristic factors to replace the original environmental variables based on the covariance matrix or the correlation matrix of all samples without any prior knowledge about them (Yin and Bura 2006). The most commonly used factor model is the orthogonal one:

$$x = \mu + AX + \varepsilon \tag{1}$$

where x represents a p-dimensional vector of the raw samples; X stands for an m-dimensional vector of characteristic factors ($m < p$); ε is the special factor for x, and $\mathrm{E}(x) = \mu$. Since this is an orthogonal factor model, $D(\varepsilon) = diag\left(\delta_1^2 \ldots \delta_p^2\right) = D$; $A = \left(a_{ij}\right)_{p\times m}$ represents the loading matrix of the factors. Let $\sum$ be the covariance matrix of the samples, its characteristic values $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$, corresponding to characteristic vectors of l_i ($i = 1, 2, \ldots p$). If the last (p-m) characteristic values are very small, the loading matrix A can be solved via $\sum = AA' + D$ namely,

$$A = \left(\sqrt{\lambda_1} l_1, \ldots \sqrt{\lambda_m} l_m\right) = \left(a_{ij}\right)_{p\times m} \tag{2}$$

To comprehend the exact meaning of a given characteristic factor, it is essential to rotate transform matrix A. Assume Γ be an orthogonal matrix, then the factor model:

$$x = AX + \varepsilon \Leftrightarrow x = A\Gamma\Gamma' X + \varepsilon \tag{3}$$

is multiplied by A repeatedly until $A\Gamma$ has a definite meaning (Greet et al. 2003). Finally, each characteristic factor is expressed as a linear combination of the original variables through inversion. With the assistance of equation $\hat{X} = A' R^{-1} x$ (R stands for the correlation matrix of the samples), the scoring function for factor X_i is determined,

together with the factor loading. Subsequently, all analyses can be based on the selected factors that have a smaller dimension than the original dataset (Janneke et al. 2012).

Fundamentals of support vector machine classification

SVM classification attempts to construct an optimal classification hyperplane through solving a second-order equation for the training dataset $T = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$, $X_i \in R^m, Y_i \in \{-1, 1\}, i = 1 \ldots n$ under an inequality constraint (Vapnik 1999). This classification plane is expressed as $\langle X \cdot \omega \rangle + b = 0$, and must satisfy the following constraint:

$$Y_i(\langle X_i \cdot \omega \rangle + b) - 1 \geq 0 \tag{4}$$

From analytical geometry, it is known that interclass distance has a value of $2/\|\omega\|$. Thus, the classification problem is reduced to minimize the function $\Phi(\omega) = \|\omega\|^2/2$. This optimization problem can be solved by introducing the Lagrangian function.

$$L = \|\omega\|^2/2 - \sum_{i=1}^{n} \alpha_i Y_i(\langle X_i \cdot \omega \rangle + b) + \sum_{i=1}^{n} \alpha_i \tag{5}$$

where $\alpha_i > 0$ represents Lagrange multipliers. Both ω and b are solved through partially differentiating the above equation against them. Under the assumption that both ω and b equal 0, the above problem is transformed into the following "symmetrical" problem, namely under the constraint of $\sum_{i=1}^{n} \alpha_i Y_i = 0, \alpha_i \geq 0$, find the maximum value for

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j Y_i Y_j \langle X_i X_j \rangle \tag{6}$$

The Karush–Kuhn–Tucker theory stipulates that the solution to this problem must satisfy the constraint of $\alpha_i\{[\langle X_i \cdot \omega \rangle + b]Y_i - 1\} = 0$. Thus, the derived identification function takes the following form:

$$f(X) = \operatorname{sgn}\left\{ \sum_{i=1}^{n} Y_i \alpha \langle X_i \cdot X \rangle + b \right\} \tag{7}$$

Generally, α_i has a value of 0 for the large majority of samples. Those samples that have a non-zero α_i are virtually a support vector; b is determined from any of the support vector. Given that some samples cannot be correctly classified by the hyperplane, it is essential to introduce a slack variable ξ, so that the optimization problem is changed to find the minimum value for

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} \xi_i \, (C > 0) \tag{8}$$

under the constraint of $Y_i(\langle X_i \cdot \omega \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0$.

Those linearly inseparable samples must be projected to the characteristic domain of a higher dimension via a non-linear transformation function $\phi(x)$ in which the transformed samples are classified. Afterwards, they are projected back to the former non-linear space. Under Mercer's conditions on the kernels, the corresponding optimization problems are convex; hence, global optimal solutions can be readily computed. According to the density function theory, so long as a function $K(X_i, Y_j)$ meets the Mercer theorem, it corresponds to the inner product of a characteristic space. If classes are non-linearly separable in a low dimensional feature space, they can be linearly classified in a high-dimensional feature space via non-linear mapping. The conversion from a low dimensional feature space to a higher one is expressed as $\phi : R^n \rightarrow H$ (H represents the Hilbert space). Because the final identification function of SVM relies solely on the inner product $\langle \phi(X_i) \cdot \phi(X_j) \rangle$ (Dong et al. 2008), it is possible to overcome the dimensionality issue commonly associated with the computation of data of a high-dimensional characteristic space, if $K(X_i, X_j) = \langle \phi(X_i) \cdot \phi(X_j) \rangle$. Known as the kernel function, $K(X_i, Y_j)$ usually takes the form of a polynomial function, radial basis function (RBF), or sigmoid function (Hwang et al. 2012). The application of kernel function considerably simplifies the amount of computation, thanks to the reduction in the dimensionality of the parameters. It is not even necessary to know the exact form of the transformation function $\psi(x)$.

Data processing

Analyzed in this study are the environmental data of the Haizhou Bay near Lianyungang, Jiangsu Province of China, including hydrological and meteorological data obtained between 2004 and 2006. These data were collected at stations distributed in the open sea to the east of the Lian Island (Fig. 1). The climate data were recorded at a weather station located in the northwest of the Island. Of all the collected variables, seventeen variables are related closely to the outbreak and dissipation of red tides. Ten of them are about the physical and chemical properties of seawater, such as chemical oxygen demand (COD), dissolved oxygen (DO), nitrate, nitrite, ammoniacal salt, phosphate, silicate, salinity, water temperature, and pH. The remaining seven parameters are climatic, including air temperature, diurnal temperature range, air pressure, rainfall, relative humidity, wind speed, and wave height. Diurnal temperature range refers to the difference between the maximum and minimum temperature of a day. Wind speed is the average of all speed recordings in a day. Wave height is 1/10 wave height. After those observations that

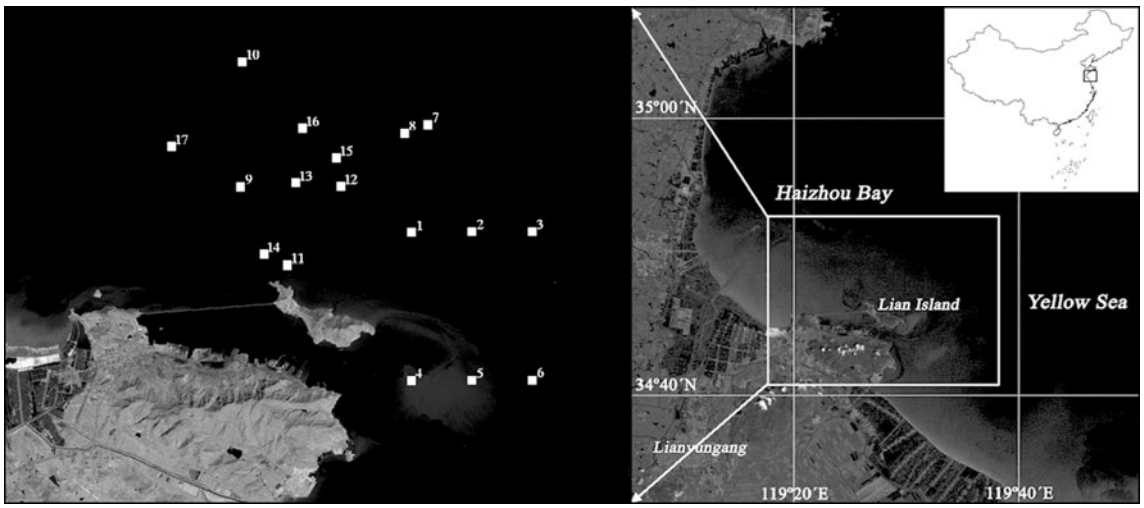

Fig. 1 Distribution of the sampling stations in the study area of Haizhou Bay and its geographic location

did not encompass the entire set of the 17 parameters were eliminated, 372 sets of samples were retained. Some of them were recorded during the outbreak of algal blooms, while others were recorded at normal algal levels. These samples were randomly divided into two groups, one for analysis and another for validation. The first group of 246 samples was used to construct the identification model, while the second (126 samples) was used to test the accuracy of the constructed model.

The first step of data processing is to standardize all sample values by subtracting the mean from the observed value and dividing the difference by the standard deviation of all samples for a given variable. Standardization was undertaken for all variables except algal density which was transformed logarithmically with a base of 10. These normalized data were analyzed using factorial analysis in SPSS. To reduce the dimensionality of the dataset, the most critical variables were selected from the 17 available variables to construct an orthogonal model. Its loading matrix A was solved through principal component analysis. This loading matrix was orthogonally rotated using the maximum variance (varimax) so that each of the retained factors has a definite physical meaning. Seven components with definite physical meaning were obtained by factor analysis. Finally, regression analysis was used to determine the derivation function for each of the characteristic factors and factor scores, and calculate the derivation vector X_i of the factor.

The logarithmically transformed algal density was compared to the criteria of red tide outbreaks (Table 1). The algal concentration level in the middle layer of seawater is one or two orders higher than the density in the surface layer prior to an algal bloom (Chen et al. 2006). In this study, it is assumed that a count of surface algae of 1×10^4 cells L^{-1} manifests an outbreak of algal bloom. This threshold is reduced to a very conservative value of 8×10^3 cells L^{-1} to ensure that the SVM model is highly sensitive to emerging red tides. Namely, a logarithmic value of 3.9 or higher for the surface algal density manifests an algal bloom. In this case, $Y_i = +1$; otherwise $Y_i = -1$. Thus, the training set $T = \{(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)\}$, $X_i \in R^m, Y_i \in \{-1, 1\}, i = 1 \ldots n$ was established.

Table 1 Criteria for identifying the outbreak of red tides (Chen et al. 2006)

Algal length (μm)	<10	10–29	30–99	100–299	300–1000
Algal density (cells L^{-1})	$>10^7$	$>10^6$	$>3 \times 10^5$	$>10^4$	$>3 \times 10^3$

The identification function was established using SVM. The inaccuracy of different kernel functions was estimated using WinSVW. This package makes use of the sequential minimal optimization algorithm that has a computation speed nearly 1,000 times faster than the traditional block algorithm (Sewell 2005). Through comparison among all error estimates, the radial basis function (RBF) with the smallest error was selected as the final kernel function (Tran et al. 2005). To determine the optimal parameterization, five values (0.5, 1, 10, 50, and 100) were tried for C; γ was set from 0.4–1 at a step width of 0.1, and $\xi = 0.1$, 0.01, 0.001. These parameter values were combined in all possible ways to identify the optimal setting that produces the smallest identification error, but still enables the constructed model to have the best applicability.

Results and discussion

Characteristic factors and their loading

After all samples were standardized they were tested against the Bartlett spherical score. They have a probability of 0, lower than the value at the 0.05 significance level. Therefore, it is permissible to perform factorial analysis on these samples. Under normal conditions, it is expected that the cumulative variance of all selected components should reach 70 % or more of the total variance in principal component analysis. The selected seven characteristic factors make up 78.81 % of the total variance (Fig. 2). They are able to capture the majority of information contained in the original samples. After the loading factor matrix A was rotated 11 times using the orthogonal rotation of the maximum variance (varimax), the following factor loading matrix was obtained Table 2.

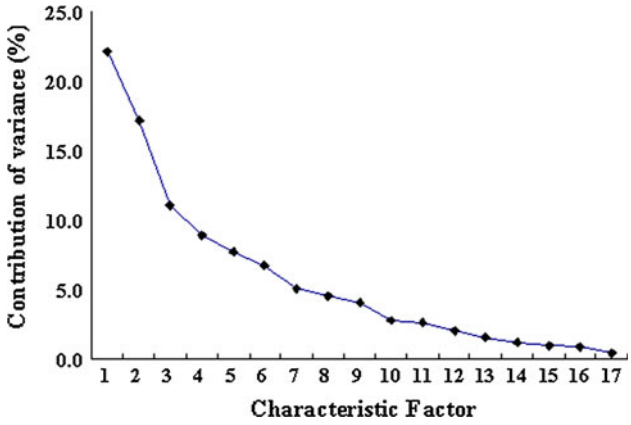

Fig. 2 The contribution of individual characteristic factors

In the Table 2, component X_1 is contributed chiefly by nitrate, nitrite, and silicate; X_2 reflects mainly water temperature and air temperature; component X_3 is contributed mainly by pH value, COD, and DO; component X_4 chiefly reflects wind speed and wave height; the loading of X_5 is high for phosphate and diurnal temperature range; the loading for rainfall and relative humidity is relatively high in component X_6; X_7 has a high loading for ammoniacal salt. Therefore, among the seven extracted characteristics factors, X_1, X_5, and X_7 are related primarily to seawater nutrient levels; X_2 is a reflector of temperature; X_3 is indicative of oxygen depletion and pH value; X_4 is related to hydrodynamics, and X_6 is related to rainfall. These identified factors are in agreement with our knowledge about the relationship between algal blooms and the environment.

Selection of support vector machine parameters and results

Selection of support vector machine parameters and results

The classification accuracy of SVM relies heavily on optimal parameterization of the SVM model. For instance, the support vector regression estimation accuracy depends on meta-parameters C (regularization) and the kernel

Table 2 Factor loading matrix after rotation by varimax

Parameter	X_1	X_2	X_3	X_4	X_5	X_6	X_7
pH	0.288	−0.097	0.760	−0.068	−0.184	0.006	−0.246
Salinity	−0.657	−0.321	−0.412	−0.130	0.100	0.046	−0.138
COD	0.105	0.182	0.833	−0.064	0.006	−0.109	0.096
DO	−0.021	−0.331	0.809	−0.172	0.097	0.079	−0.060
PO_4^{3-}	0.068	−0.043	0.039	0.234	0.707	−0.170	0.099
NO_2^-	0.832	0.019	−0.059	0.035	0.107	−0.241	0.111
NO_3^-	0.835	0.076	0.202	0.230	−0.030	0.118	0.141
NH_4^+	0.195	−0.015	−0.091	−0.007	−0.058	−0.001	0.916
SiO_3^{2-}	0.749	−0.073	0.080	−0.082	−0.206	0.175	−0.030
Water temperature	0.228	0.805	0.024	0.193	−0.354	−0.208	0.032
Air temperature	0.016	0.949	−0.059	0.029	−0.049	−0.034	−0.045
Diurnal temp. range	−0.248	−0.151	−0.109	−0.228	0.768	0.095	−0.208
Air pressure	0.165	−0.599	0.233	0.089	−0.299	−0.568	−0.125
Relative humidity	0.059	0.189	0.005	0.333	−0.414	0.550	0.270
Precipitation	0.049	−0.222	0.009	0.134	−0.065	0.751	−0.095
Wind speed	0.014	0.130	−0.138	0.896	0.070	0.089	0.035
Wave height	0.136	−0.011	−0.091	0.906	−0.043	0.110	−0.033

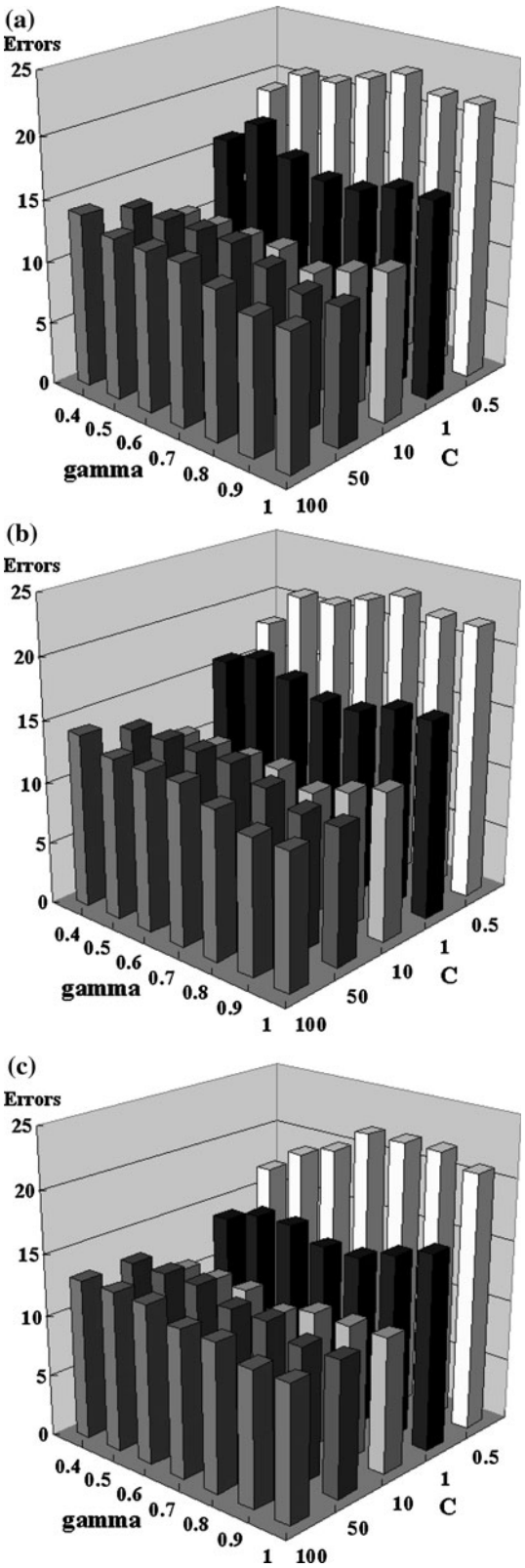

Fig. 3 Comparison of incorrectly identifying the validation samples by the model among three different ξ values. (**a**) $\xi = 0.001$, (**b**) $\xi = 0.01$, and (**c**) $\xi = 0.1$

parameters (Durbha et al. 2007). Often the meta-parameters are selected using prior knowledge and/or user expertise. Analysis of the results of classifying the validation samples (Fig. 3) shows that classification accuracy is the most sensitive to the change of C value. In the SVM classification, C is a penalty coefficient for an incorrect classification. The larger the C value, the heavier the penalty. Essentially, it plays a compromising role between the conflicting requirements for maintaining a high degree of model applicability and minimizing classification errors (Yao et al. 2005). Theoretically, the larger the C value, the higher the classification accuracy. Nevertheless, when C reaches 50 or higher, the model will overfit with the samples. When validated against the randomly selected samples, the classification accuracy of the model is lowered. A similar relationship also exists between the variation in the γ value and classification accuracy. Γ is a parameter defining the RBF width. It controls the radial range of influence of the function. In theory, the smaller the value, the more compact the optimization zone, and the higher classification accuracy. However, a too small γ value can result in inefficient computation, and hence a lower accuracy. Variation in the ξ value exerts the smallest impact on the classification accuracy, especially when its value lies between 0.01 and 0.001. In this case, the classification accuracy scarcely changes. In this study, all the combinations of the selected parameter values achieved a classification accuracy level over 80 %, defined as the ratio of correctly identifying samples to the total number of samples. In particular, three combinations of parameter values ($C = 10$, $\gamma = 0.8$, $\xi = 0.001$; $C = 10$, $\gamma = 0.8$, $\xi = 0.01$; $C = 10$, $\gamma = 0.7$, $\xi = 0.1$) achieved the highest accuracy of 92.06 %.

When the constructed model was applied to identifying the occurrence of red tides in the retained data, the errors of labeling red tide samples as normal (e.g., non-occurrence) far exceed the errors in the opposite direction. For instance, the accuracy of correctly labeling red tide samples has an average accuracy of 73.1 %, much lower than the mean accuracy of 94.7 % in labeling non-red tide samples. The discrepancy is caused by the lower proportion of red tide samples in the training dataset. The use of the same penalty coefficient C to all samples means that a relatively heavier penalty is imposed on minority classes. Theoretically, the penalty level should be set in accordance with the proportion of different classes in the training dataset. However, the proportion of red tide samples in the training dataset is affected by sampling frequency, observation duration, and the stability of environment. Therefore, it is impossible to set the penalty coefficient based on the observation data available due to their proportion. Analysis

of errors in identifying the samples reveals that a large majority of them are clustering around the threshold. This model still has a high degree of identification accuracy owing to the conservative threshold of red tide occurrence adopted.

Analysis of model applicability

If excessive attention is devoted to classification accuracy, then the model will overfit all samples in light of noisy data. To maintain a high degree of applicability, it is essential to optimize the model by leaving one variable out at a time, commonly known as the Leave-One-Out (*LOO*) method (Tran et al. 2005). It has been proven that *LOO* is the best unbiased estimate of the identification errors. The smaller the *LOO* value, the wider the applicability of the model. In this method, one sample is removed from the training dataset, and the remaining samples are used to construct the identification model. This constructed model is then used to classify the removed sample. A value of 0 is reserved for a correct classification, while a value of 1 is assigned for an incorrect classification. Let $f_i(X_i)$ be the classification criterion after the removal of the ith sample, and $P(f_i(X_i), Y_i)$ be the classification results, then the *LOO* error is calculated as:

$$LOO = \frac{1}{n}\sum_{i=1}^{n} P(f_i(X_i), Y_i) \quad (9)$$

The *LOO* value rises initially but decreases afterwards with the increase in the C value (Table 3). The *LOO* value also increases with ξ. The change in the γ value exerts an indefinite pattern of influence on the *LOO* value. Because the model has been constructed by overfitting the samples, it is imperative to judge its applicability by the *LOO* value.

All classifications are achieved at an accuracy level over 90 % at $C = 10$ after both classification accuracy and model applicability have been taken into consideration. Therefore, the combination that is associated with the lowest *LOO* value was selected among the 21 potential combinations of parameter values at $C = 10$. The exact values are $C = 10$, $\gamma = 0.7$, and $\zeta = 0.1$. These values were used to construct the identification model. Checked against the validation samples, this model achieved an overall accuracy of 92.06 %. The accuracy is lowered to 80.56 % when the model is applied to identifying red tide samples. However, the accuracy for labeling non-red tide samples is as high as 96.67 %. This combination of the parameter values is considered optimal in that it not only achieves the highest identification accuracy, but also maintains a high degree of applicability that is much more superior to the models based on other combinations. This accuracy level is highly comparable to 91.89 % reported by Pal (2006) in classifying 15 forest-based features, and 89.4 % in detecting forest cover change from high resolution IKONOS images (Huang et al. 2008).

Table 3 The *LOO* value in different combinations of parameters

γ	ξ	C				
		0.5	1	10	50	100
0.4	0.1	0.191	0.166	0.150	0.177	0.206
	0.01	0.207	0.188	0.155	0.181	0.220
	0.001	0.215	0.188	0.156	0.182	0.220
0.5	0.1	0.210	0.174	0.148	0.186	0.196
	0.01	0.233	0.195	0.153	0.191	0.205
	0.001	0.233	0.203	0.153	0.191	0.206
0.6	0.1	0.221	0.174	0.141	0.186	0.186
	0.01	0.237	0.187	0.148	0.195	0.195
	0.001	0.237	0.188	0.148	0.196	0.196
0.7	0.1	0.240	0.171	0.133	0.174	0.174
	0.01	0.248	0.180	0.143	0.187	0.187
	0.001	0.249	0.180	0.144	0.188	0.188
0.8	0.1	0.242	0.169	0.140	0.166	0.166
	0.01	0.259	0.180	0.136	0.174	0.174
	0.001	0.260	0.180	0.136	0.174	0.174
0.9	0.1	0.244	0.180	0.141	0.155	0.155
	0.01	0.253	0.189	0.144	0.161	0.161
	0.001	0.254	0.190	0.144	0.162	0.162
1.0	0.1	0.238	0.187	0.142	0.155	0.155
	0.01	0.255	0.189	0.153	0.159	0.159
	0.001	0.256	0.189	0.153	0.159	0.159

Conclusion

This study has demonstrated that the most important variables for predicting algal blooms are numbered seven from a total of 17 available ones. Three of them are related to the nutrients of seawater. The other factors are related to temperature, oxygen depletion, pH value, hydrodynamics, and rainfall. This reduction in data dimensionality closely matches what is known about the relationship between algal blooms and the environmental settings. Such a reduction simplifies the complexity of identifying the occurrence of algal blooms and accelerates the speed of computation.

The best model is constructed with the parameter values of $C = 10$, $\gamma = 0.7$, and $\zeta = 0.1$. However, whether this setting is always optimally depends on the proportion of red tide samples in the dataset. Successfully overcoming the dimensionality problem, SVM is able to label all samples at an overall accuracy of 92.06 %. The accuracy for identifying the samples of red tides is lowered to

80.56 %. The labeling of non-algal bloom samples is achieved at a much higher accuracy of 96.67 %. Besides, this learning method has a strong applicability.

The main problem with the use of the SVM method is the acquisition of accurate proportion of red tide samples. Their low proportion in the entire dataset is blamed for the lower accuracy in identifying algal bloom samples. In addition, the selection of the model parameter values is experimental to a large degree. In future, more research should be directed at developing guidance in selecting the model parameters directly from the values of the samples themselves.

Acknowledgments This research was financially supported by the National Science Foundation Project of China (No. 41071083) and the Natural Science Foundation of Jiangsu Higher Education Institutions (No. 11KJB170010). The authors are also indebted to Dr. Jay Gao (The University of Auckland) for his valuable suggestions on this research.

References

Bao L, Sun ZR (2002) Identifying genes related to drug anticancer mechanisms using support vector machine. Fed Eur Biochem Soc Lett 521:109–114

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Cai HJ, Tang XX, Zhang PY, Yang Z (2002) The effect of initial cell density on the interspecific competition between three species of red tide microalgae. Acta Ecologica Sinica 22:1635–1639

Cai CZ, Wang WL, Sun LZ, Chen YZ (2003) Protein function classification via support vector machine approach. Math Biosci 185:111–122

Chen HL, Lu SH, Zhang CS, Zhu DD (2006) A survey on the red tide of Prorocentrum donghaiense in East China Sea. Ecol Sci 25:226–230

Dong XW, Liu YJ, Yan JY, Jiang CY, Chen J, Liu T, Hu YZ (2008) Identification of SVM-based classification model, synthesis and evaluation of prenylated flavonoids as vasorelaxant agents. Bioorg Med Chem 16:8151–8160

Douglass EM, Jayne SR, Bryan FO, Peacock S, Maltrud M (2012) Kuroshio pathways in a climatologically forced model. J Oceanogr 68:625–639

Durbha SS, King RL, Younan NH (2007) Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sens Environ 107:348–361

Foody GM, Mathur A (2004) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens Environ 93:107–117

Gilbert CSL, Li WK, Kenneth MYL, Joseph HWL, Jayawardena AW (2007) Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter. Ecol Model 200:130–138

Greet P, Peter JR, Peter F, Christophe C (2003) Robust factor analysis. J Multivar Anal 84:145–172

Hassan G, Zahra D, William EAJ (2012) Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. Eur J Pharm Sci 47:421–429

Hodgkiss IJ, Ho KC (1997) Are changes in N:P ratios in coastal waters the key to increased red tide blooms? Hydrobiologia 352:141–147

Huang C, Davis LS, Townshend JRG (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens 23:725–749

Huang CQ, Song K, Kim S, Townshend JRG, Davis P, Masek JG, Goward SN (2008) Use of a dark object concept and support vector machines to automate forest cover change analysis. Remote Sens Environ 112:970–985

Hwang SH, Ham DH, Kim JH (2012) Forecasting performance of LS-SVM for nonlinear hydrological time series. KSCE J Civ Eng 16:870–882

Janneke I, Georg S, Kai H, Bernhard D, Elisabeth D, Stephan O, Bernd W, Frank L (2012) Environmental conditions in the Donggi Cona lake catchment, NE Tibetan Plateau, based on factor analysis of geochemical data. J Asian Earth Sci 44:176–188

Li HD, Liang YZ, Xu QS (2009) Support vector machines and its applications in chemistry. Chemometr Intell Lab Syst 95:188–198

Pal M (2006) Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. Int J Remote Sens 27:2877–2894

Sanchez-Hernandez C, Boyd DS, Foody GM (2007) Mapping specific habitats from remotely sensed imagery: support vector machine and support vector data description based classification of coastal saltmarsh habitats. Ecol Inform 2:83–88

Sewell M (2005) A windows implementation of a support vector machine. http://winsvm.martinsewell.com/. Accessed 28 Jan 2013

Smola AJ, Scholkopf B (1998) On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. Algorithmica 22:211–231

Tran QA, Li X, Duan HX (2005) Efficient performance estimate for one-class support vector machine. Pattern Recogn Lett 26:1174–1182

Vapnik VN (1982) Estimation of Dependencies Based on Empirical Data. Springer, Berlin

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999

Yang ZB, Hodgkiss IJ (2004) Hong Kong's worst "red tide"—causative factors reflected in a phytoplankton study at Port Shelter station in 1998. Harmful Algae 3:149–161

Yao XJ, Panaye A, Doucet JP, Chen HF, Zhang RS, Fan BT, Liu MC, Hu ZD (2005) Comparative classification study of toxicity mechanisms using support vector machines and radial basis function neural networks. Anal Chim Acta 535:259–273

Yin XR, Bura E (2006) Moment-based dimension reduction for multivariate response regression. J Stat Plan Inference 136:3675–3688

Zhao CY, Zhang RS, Liu HX, Xue CX, Zhao SG, Zhou XF, Liu MC, Fan BT (2004) Diagnosing anorexia based on partial least squares, back propagation neural network, and support vector machines. J Chem Inf Comput Sci 44:2040–2046

Zhao CY, Zhang HX, Zhang XY, Liu MC, Hua ZD, Fan BT (2006) Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology 217:105–119

Zhou JY, Shi J, Li G (2011) Fine tuning support vector machines for short-term wind speed forecasting. Energy Convers Manag 52:1990–1998