Saudi Journal of Biological Sciences (2017) 24, 537–547

King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Support vector machine-based open crop model (SBOCM): Case of rice production in China

Su Ying-xue, Xu Huan, Yan Li-jiao *

College of Life Sciences, Zhejiang University, 310058 Hangzhou, Zhejiang Province, PR China

Received 26 October 2016; revised 5 January 2017; accepted 9 January 2017
Available online 30 January 2017

KEYWORDS

Crop model;
Crop simulation;
Scaling up;
Support vector machine;
SBOCM

Abstract Existing crop models produce unsatisfactory simulation results and are operationally complicated. The present study, however, demonstrated the unique advantages of statistical crop models for large-scale simulation. Using rice as the research crop, a support vector machine-based open crop model (SBOCM) was developed by integrating developmental stage and yield prediction models. Basic geographical information obtained by surface weather observation stations in China and the 1:1000000 soil database published by the Chinese Academy of Sciences were used. Based on the principle of scale compatibility of modeling data, an open reading frame was designed for the dynamic daily input of meteorological data and output of rice development and yield records. This was used to generate rice developmental stage and yield prediction models, which were integrated into the SBOCM system. The parameters, methods, error resources, and other factors were analyzed. Although not a crop physiology simulation model, the proposed SBOCM can be used for perennial simulation and one-year rice predictions within certain scale ranges. It is convenient for data acquisition, regionally applicable, parametrically simple, and effective for multi-scale factor integration. It has the potential for future integration with extensive social and economic factors to improve the prediction accuracy and practicability.

1. Introduction

Crop model is the general term used to identify a series of methods that use mathematical concepts to describe the process of crop growth. However, based on a combination of previous definitions (Curry, 1971; Edwards and Hamson, 1990; Gao, 2004; Sinclair and Seligman, 1996; Xiong, 2004) and the findings of the present study, a crop model is defined in this paper as a computer program that mathematically describes and models the rules of crop growth and can be used to quantitatively and dynamically explain the process of crop growth, development, yield, and reaction to environmental changes.

Crop models can be categorized as crop statistical models and crop simulation models (or crop growth models) based on the basic mathematical method of the modeling. Simulation models were generally considered to be better than statistical models because it facilitated the study of crop growth theory in a physiological sense through the enablement of experimen-

* Corresponding author.
E-mail address: tulipblue16@zju.edu.cn (L.-j. Yan).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.sjbs.2017.01.024
1319-562X © 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

tal comparison. However, with increasing dissatisfaction with the fit effect of the large-scale nonlinear problems associated with early statistical models, simulation models were more widely employed (Xie and James, 2002).

Famous crop models such as the CERES model (Charles-Edwards, 1986; Jones and Kiniry, 1986; Ritchie, 1972) series of America, SUCROS (Lin et al., 2003) and MACROS (Penning de Vries et al., 1989) models of Netherlands, and RSM model (Luo et al., 1990) of China are all crop simulation models. Through the efforts of several generations of experts, crop models and various agricultural production and decision systems that primarily utilize crop models have contributed to both crop development physiology studies and agricultural production, in which field there has been much achievement. Since the 1990s, a number of large-scale agricultural decision system software packages taking these crop models as kernels have been developed through application modules, human–machine interface optimization, integration of decision systems, and data normalization. Such developments include the Decision Support System for Agrotechnology Transfer (DSSAT) model of America, Agricultural Production Systems sIMulator (APSIM) model of Australia, and Crop Cultivational Simulation Optimization Decision-making System (CCSODS) of China (Gao, 2004). These models have been vigorously promoted and utilize to different degrees in various countries around the world.

Although simulation models are more widely employed, it is still difficult to determine whether they are actually presently better than statistical models (Dhungana et al., 2006). This is mainly because of the technological and practicality bottlenecks encountered by the former in the early 21st century when they were promoted as the main crop models. The technological bottleneck regarded how to implement simple operations and scaling up, while the practicality bottleneck was due to the grey system feature in agricultural extension. The weaknesses of crop simulation models gradually became apparent when they were put into practical use in the early 1980s. Meanwhile, statistical models had been found to be practical through several large-scale studies (Stewart and Dwyer, 1990). This led to the emergence of the American school of thought that used statistical models for simulation purposes as need arose. The currently popular American CERES model is a typical American school of thought, being a simulation model in a general sense, but with an integrated statistical estimation method (Swain et al., 2007).

While crop modeling was encountering its bottlenecks, nonlinear statistical theory, particular with regard to machine learning, was making a huge breakthrough in the 1990s. Since then, artificial intelligence has undergone comprehensive development and application through the use of computer iterative algorithms such as the support vector machines (SVMs) (Vapnik, 1998, 1999; Cortes and Vapnik, 1995). Owing to their good sparsity (Gunn, 1998), ability to fit small samples (Suykens, 2001), and global optimization (Xu et al., 2007), SVMs have outperformed other none-linear statistical models (Gualtieri and Cromp, 1998; Viaene et al., 2001; Van Gestel et al., 2001a,b; Xiong, 2009; Zhang, 2009). In recent years, SVMs have also been applied in agricultural production for purposes such as remote monitoring, moisture prediction, and plant disease and insect pest warning (Gill et al., 2006; Du et al., 2008; Kaundal et al., 2006; Trafalis et al., 2007; Yang et al., 2008; Yu et al., 2008).

Rice, which is China's main food crop, was considered in the present study. An SVM was incorporated into the developed crop model, which is here presented as SVM-based open crop model (SBOCM). The basic idea of this study was the use of basic geographic information obtained from surface weather observation stations in China (i.e., daily published meteorological data and the 1:1000000 soil database published by the Chinese Academy of Sciences [CAS] (Shi et al., 2002)) as input, and the rice development and yield records of all agricultural observation stations in China as output. A dynamic open reading frame was designed to dynamically input the daily meteorological data, and a scheduled developmental stage prediction was obtained by SVM classification (SVC), and yield prediction by SVM regression (SVR).

2. Materials and methods

2.1. Support vector machine

The SVM-by-Steve Gunn v2.1 software in the MATLAB kit was adopted in our SVM program. The SVM software was presented by Vapnik in the middle 1990s (Cortes and Vapnik, 1995) and has been widely used for machine learning over the last 15 years (Vapnik, 1998, 1999). The theoretical basis of the SVM is the Structural Risk Minimization Principle in statistical learning theory (Vapnik, 1998). Kernel functions were used to convert the linear inseparability problem in low-dimension space into a linear partition problem in high-dimensional space. The optimal hyperplane was determined to separate the two groups of eigenvectors based on their respective longest distances from the interface.

For the given training set $T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (X \times Y)^l$, where $x_i \in X = R$ and $y_i \in Y = \{-1, 1\} (i = 1, 2, \ldots, l)$, a real-valued function $g(x)$ on $X = R^n$ was sort as the decision function $f(x)$.

$$f(x) = sgn(g(x)) \tag{1}$$

The value of y corresponding to x in any mode could be inferred by $f(x)$. In other words, a rule for dividing the points on R^n into two parts was sought.

The linearly separable training set $\min \frac{1}{2} w^2 + C \sum_{i=1}^{l} \xi_i$ was obtained under the following constraint in SVM:

$$y_i((w \cdot x_i) + b) + \xi_i \geqslant 1, \quad i = 1, \ldots, l \tag{2}$$

where ξ_i is the slack variable and C is the penalty coefficient, which should be set artificially in practice.

After obtaining the optimal solutions ω^* and b^*, the following separating hyperplane was constructed:

$$(w^* \cdot x) + b^* = 0 \tag{3}$$

The decision function was then obtained as

$$f(x) = sgn((w^* \cdot x) + b^*) \tag{4}$$

When the sample training set was non-linear and separable, kernel functions (sometimes denoted by K in SVM programs) were required in SVM to deal with the non-linear classification problem and build a mapping relationship between the input vectors and high-dimensional space vectors. Our study considered the non-linear inseparable problem, wherein, theoretically, the introduction of slack variables and kernel functions did not affect the solution of simple linear separable problems.

Hence, both slack variables and kernel functions were introduced into the SVM training. Linear, polynomial, and radical basis kernel functions were adopted in this study.

2.2. Open reading frame

Two types of SVM training samples were used: (1) SVC binary classification samples, which were used to investigate the occurrence time of certain developmental stages of rice; (2) SVR samples, which were used to investigate certain yield records of rice; a unit of which comprised a pair of input and output vectors, with each pair constituting a record. Modeling requires consistence of the unit structures within a sample. Five developmental stages of rice were considered in this study, namely, sowing, transplanting, tillering, heading, and milk, and the agricultural production time varied significantly among the stations. Thus, in the development of the training samples, we ensured as much as was possible that the sample rules included the principle of maintaining the biological significance of samples.

For the above reasons, the corresponding input vector of each record consisted of both static and dynamic variables. The static variables included basic information and soil information of each station that generated records. Hence, the records obtained from a particular station had the same static variables irrespective of the developmental stage. Meanwhile, the dynamic variables could vary with the developmental stage of a record or yield prediction target. An open reading frame was set to generate the dynamic variables.

The open reading frame was a fixed-length input window for the daily meteorological data. The length was fixed for a given sample so that the generated variables would be of a certain length. The frame could read the daily meteorological data for a certain period in accordance with the requirements for sample generation, and could generate input variables using one developmental record and yield record. This is done using the methods for generating positive samples in the developmental stages (Fig. 1) and for choosing two developmental stages for the generation of dynamic variables (Fig. 2).

2.3. Data preprocessing

Historical data obtained from two sets of observation systems of the China Meteorological Administration network database (http://www.cma.gov.cn/) were used in this study. Three types of rice plantings were considered, namely, middle-season, early, and late planting. After organization of the station information, soil information, and daily meteorological data, and screening them based on their biological significance, we chose a seven-day open reading frame. It was determined that the variables consisted of the initial input variables (Table 1), which added up to 53 dimensions. The samples were built in five developmental stages, namely, sowing, transplanting, tillering, heading, and milk, respectively. Principal component analysis (PCA) was used to screen the different factors, based on which all the classes of the samples used for the SVM training were built.

The samples for modeling the SVC developmental prediction were binary classified; i.e., the output of the samples for determining whether a developmental stage had occurred was labeled as "yes" or "no" The dynamic variables of the positive samples were the daily meteorological factors for seven days before the occurrence of a given developmental stage, while those of the negative samples were generated by off-season (150 days in advance) and 30 days in advance strategies, respectively. For convenient expression, the training samples consisting of the positive samples and off-season negative samples were identified as sample class 1, while the ones consisting of the positive samples and 30-days-in-advance negative samples were identified as sample class 2.

Thus, five developmental stage samples for each of the three planting types were generated, adding up to $3 \times 5 = 30$ different samples. These were respectively used to predict the five developmental stages of sowing, transplanting, tillering, heading, and milk, respectively. In the actual training, each set of samples was randomly divided into five parts of equal sizes. Five-fold cross-validation was then used for the model training and testing.

2.4. Building developmental module

By SVC, a developmental module is capable of organizing the data sets of the five developmental stages and separately modeling middle-season rice, early rice, and late rice. In the present study, by separately modeling the two classes of training samples and comparing the models, we finally obtained the best prediction model of the occurrence time of the sowing stage.

The specific process was as follows. Linear (linearly separable SVC without the use of kernel functions), polynomial, and radical basis kernel functions were chosen and used to conduct SVC training. Fivefold cross-validation was then used to model and test the prediction of the sowing stage for finding the optimal kernel functions. For polynomial and radical basis kernel functions, the optimal hyperparameters were determined via ergodic tests on the corresponding hyperparameters. The SVC penalty coefficient was subsequently further adjusted to improve the optimal model, which had been tested and found to be partly unsatisfactory. Through comparison of the two optimal models developed by the two different strategies, the more suitable model for sowing stage prediction was finally identified.

2.5. Building yield module

By SVR, a yield module is capable of yield prediction based on SVM analysis of the information obtained from a given station and the soil and daily meteorological data during different developmental stages of the particular type of rice. The module then outputs the record of the rice yield for the given year.

The SVR models developed in this study were respectively based on samples for the tillering stage, heading stage, tillering and heading stages, milk stage, and heading and milk stages. The optimal yield prediction models for the heading, tillering and milk stages were figured out after a comparison.

The specific modeling and optimization process was as follows. The linear (linearly separable SVR without using kernel functions), polynomial, and radical basis kernel functions were used for SVR training, after which fivefold cross-validation was used for modeling and testing, respectively. For polynomial and radical basis kernel functions, the optimal hyperparameters were determined via ergodic tests on the corresponding hyperparameters. The SVR penalty coefficient was subsequently further adjusted to improve the optimal model, which had been tested and found to be partly unsatisfactory. Through comparison of the several optimal models

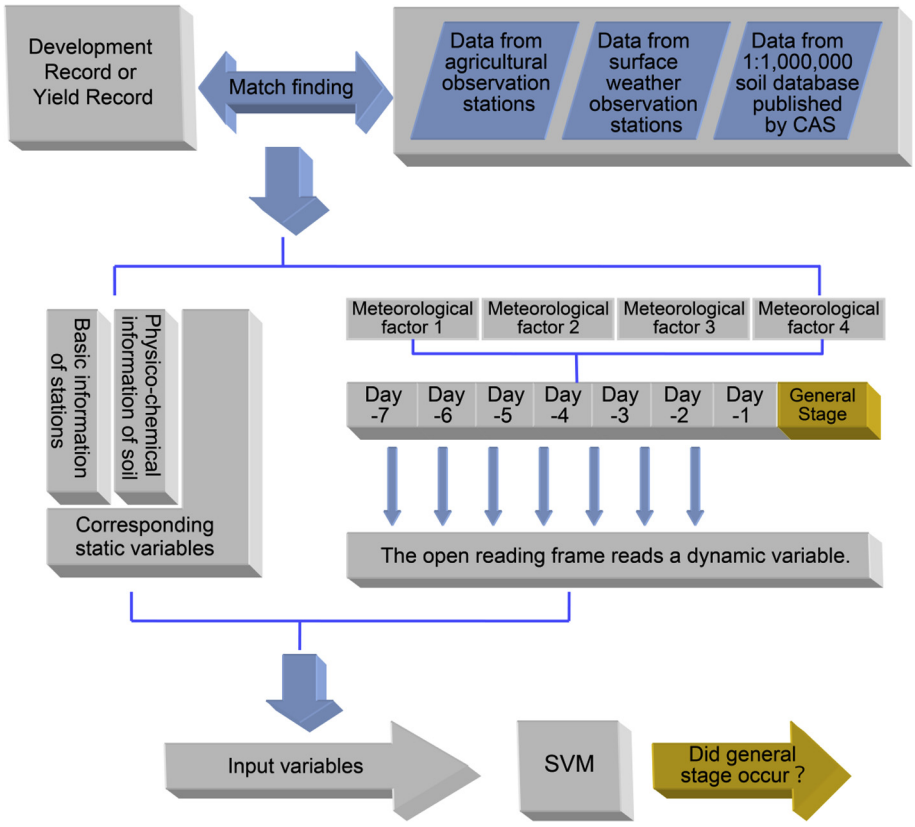

Figure 1 Building input vector based on a special development record: Show a flow chart about how to build an input vector based on a special development record in this study.

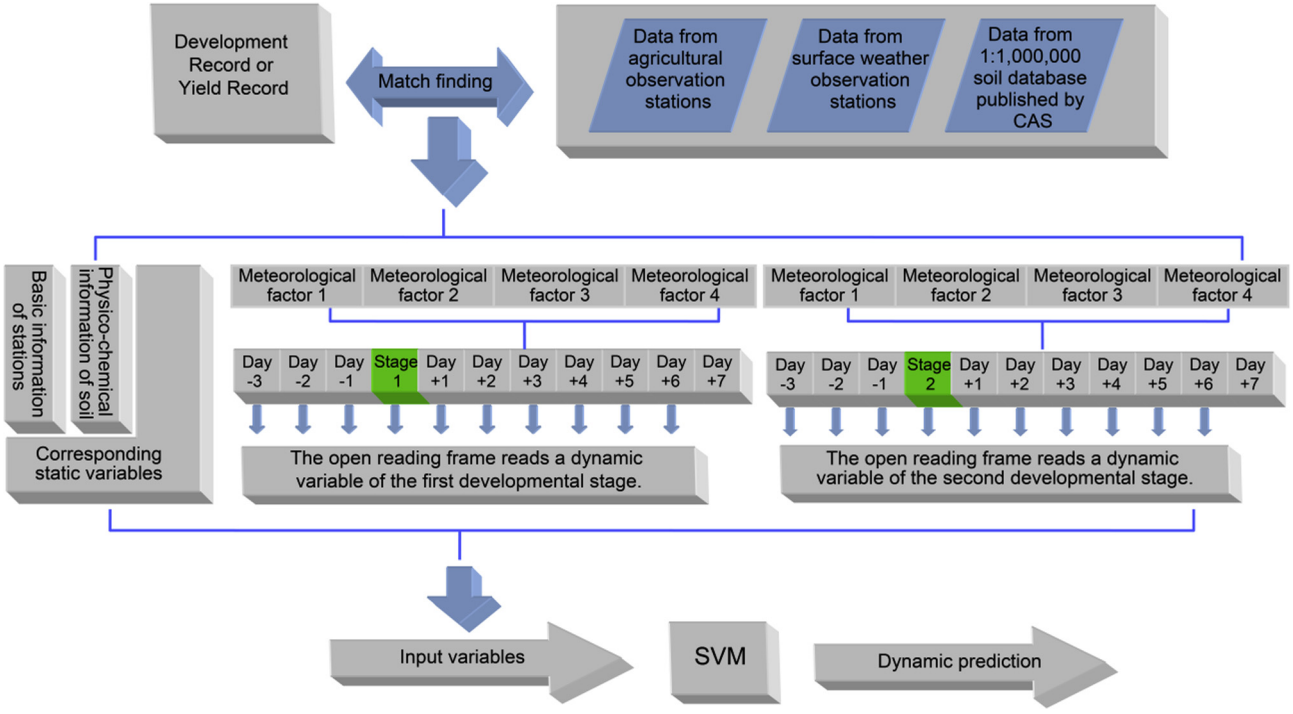

Figure 2 Building input vector based on a special yield record: show a flow chart about how to build an input vector based on a special yield record in this study.

Table 1 Original components of input vector.

Static variables		Dynamic variables
Station information	Soil information	Daily information (running days)
Longitude (east); Latitude (north); Altitude (m)	Soil code, section thickness, soil composition entropy, organic matter, pH, total nitrogen, total phosphorus, total potassium	Daily air pressure, average daily temperature, average daily relative humidity, 24-h precipitation, daily wind speed, sunshine hours

Table 2 Uses of different developmental samples in modeling.

Sample name	Sample development strategy	Modeling purpose
Tillering 1	Eleven consecutive days in tillering stage	Yield prediction in tillering stage
Heading 1	Eleven consecutive days in heading stage	Yield prediction in heading stage
Heading 2	Eleven consecutive days in tillering stage + Eleven consecutive days in heading stage	Yield prediction in tillering stage
Milk 1	Eleven consecutive days in milk stage	Yield prediction in milk stage
Milk 2	Eleven consecutive days in heading stage + Eleven consecutive days in milk stage	Yield prediction in milk stage

developed by the different strategies, we finally identified the most suitable model for yield prediction.

3. Results

3.1. Development module

Two classes of samples, fivefold cross-validation, and different kernel functions and hyperparameters were used for separate training of all the developmental stages of the middle-season rice, early rice, and late rice. Based on the training performances for the same developmental stages, we chose the sample class 1 optimal kernel functions and hyperparameters for all the developmental stages (see Tables 2–4), for which the SVC penalty coefficient was always 1.

Because of the satisfactory F1 value of the sample class 2, the penalty coefficient C was further adjusted to improve all the models after the optimal kernel functions and their hyperparameters had been determined. We realized from the findings of previous studies (Cai et al., 2003; Gill et al., 2006; Gunn, 1998; Suykens et al., 2001; Van Gestel et al., 2001b,c) that the penalty coefficient generally increases at a rate of 10^2 and that an excessively large value of C significantly decreases the computation efficiency (Fig. 3). Hence, all the models were tested using $C = 1$, 10, 100, and 10000, respectively (Table 5). Through the adjustment of C, we chose the sample class 2 optimal kernel functions and hyperparameters for the training of all the developmental stages (Table 6).

3.2. Yield module

Five classes of samples, fivefold cross-validation, and different kernel functions and hyperparameters were used for the separate training and testing of the three classes of rice plantings. A penalty coefficient C of 1 was used to compute the root-mean-square error (RMSE) (kg/h m^2) and relative error (RE) (%) of each training. We then chose the most suitable sample, optimal kernel functions, and hyperparameters for all the developmental stages of the yield simulation (Table 7), for which the SVR penalty coefficient was always 1.

Because the yield simulation was not sufficiently accurate, we further adjusted the penalty coefficient C after the most suitable sample, optimal kernel functions, and hyperparameters had been determined (Fig. 4 and Table 8). This was done to improve the accuracy of all the models as we did with the development module. The most suitable prediction models for all the developmental stages and their respective performances (Table 9) were finally determined.

Table 3 Optimization of each developmental stage for sample class one (with unadjusted C).

Middle-season rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial
Hyperparameter	$D = 3$	$D = 1$	$D = 2$	$D = 1$	$D = 2$
F1	0.8282	0.9908	0.9968	1	0.9968
Early rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Radical basis	Polynomial	Polynomial	Polynomial
Hyperparameter	$D = 3$	$p = 1$	$D = 2$	$D = 2$	$D = 4$
F1	0.7957	0.9665	0.9792	0.9938	0.9876
Late rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Polynomial	Radical basis	Polynomial	Radical basis
Hyperparameter	$D = 3$	$D = 3$	$p = 0.75$	$D = 3$	$p = 0.75$
F1	0.8136	0.9778	0.9921	0.9924	0.9721

Table 4 Optimization of each developmental stage for sample class two (with unadjusted *C*).

Middle-season rice					
Development stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Polynomial	Radical basis	Polynomial
Hyperparameter	$D = 2$	–	$D = 2$	$p = 1.75$	$D = 2$
F1	0.8455	0.8255	0.8000	0.7221	0.6933
Early rice					
Development stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Polynomial	Radical basis	Polynomial
Hyperparameter	$D = 1$	–	$D = 2$	$p = 1.5$	$D = 2$
F1	0.8075	0.8066	0.7990	0.7036	0.6903
Late rice					
Development stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Radical basis	Radical basis	Radical basis
Hyperparameter	$D = 2$	–	$D = 2$	$p = 1.75$	$D = 2$
F1	0.8546	0.8353	0.8075	0.7436	0.7028

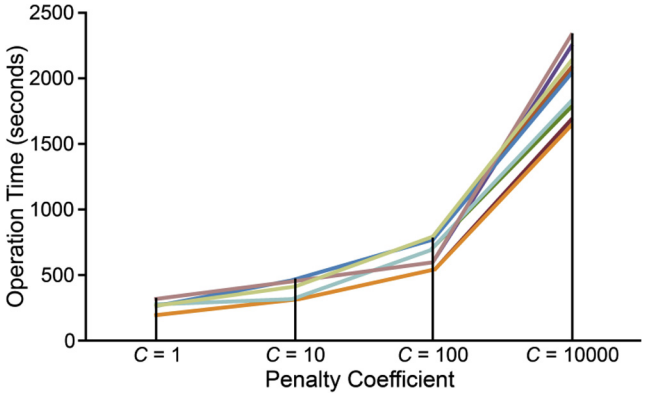

Figure 3 Relationship between the *C* value and efficiency (development module). All the models were tested using $C = 1$, 10, 100, and 10,000 in the development module. The operation time increased with the increase in the *C* value.

3.3. SVM-based open crop model

Based on the above results, an SVM-based open crop model (SBOCM) was designed as an application-focused crop model for regional rice development stage prediction and yield prediction. The emphasis was on simplicity of operation.

The framework of the entire SBOCM system was quite simple, comprising five parts, namely, the database module, data calling module, development prediction module, yield prediction module, and human–machine interface (Fig. 5).

The SBOCM had the following functions and features:

- Perennial simulation: Historical meteorological data was used as the basic input to directly simulate the time and yield of all the developmental stages in the various regions.
- One-year prediction: The real-time meteorological data of a particular year was used to dynamically follow and simulate all the developmental stages at all the stations, thus enabling real-time predictions of the developmental stages and yield.
- Regional prediction: The input of the SBOCM was regional data and the yield simulation of a particular place only required the integration of the corresponding meteorological data and soil and other information of the place. We could thus simulate an entire region without conversion between point and surface models. Previously, the application of the function required users to write MATLAB scripts by themselves and automatically and repeatedly input all point data into the SBOCM.
- Extendibility: The features of the input variables of the SVM imply the capability of the SBOCM for future absorption of more natural and social factor inputs. This is significant for the extension of the applicable functions of the model.

4. Discussion

4.1. Kernel function, hyperparameter, and penalty coefficient

The key issue in SVM modeling is the determination of the kernel functions, hyperparameters, and penalty coefficient. In the present study, the SVC of five developmental stage predictions of three rice planting types, namely, middle-season, early, and late rice planting, and the SVR of the yield predictions of three developmental points were separately determined.

It was observed from the final models that, as far as the kernel functions were concerned, the developmental stage predictions and yield prediction were all complicated nonlinear segmentation problems. Thus, the performances of the polynomial and radical basis kernel functions were better than those of linear functions. This was especially so for yield prediction, wherein the radical basis kernel function prediction accuracy increased with increasing variable dimensions.

For most problems, the hyperparameters were found to be within a rational range. For example, *D* for the polynomial kernel functions was generally between 2 and 4, while *p* for the radical basis kernel functions was between 1 and 1.5. In conformity with the empirical range of hyperparameters determined in most previous studies (Cai et al., 2003; Gunn, 1998;

Table 5 SVC results for scanning C value (F1).

	Kernel function	Hyperparameter	$C = 1$	$C = 10$	$C = 100$	$C = 10000$
Middle-season rice						
Sowing	Polynomial	$D = 2$	**0.8455**	0.8219	0.7936	0.7641
Transplanting	Linear	–	**0.8256**	0.8113	0.7654	0.7421
Tillering	Polynomial	$D = 2$	**0.8**	0.7959	0.7758	0.7364
Heading	Radical basis	$p = 1.75$	0.7221	0.7286	**0.7532**	0.6213
Milk	Polynomial	$D = 2$	**0.6933**	0.6919	0.6854	0.6534
Early rice						
Sowing	Polynomial	$D = 1$	0.8075	0.8013	0.7874	0.7544
Transplanting	Linear	–	**0.8066**	0.8062	0.7639	0.7217
Tillering	Polynomial	$D = 2$	**0.799**	0.7840	0.7710	0.7262
Heading	Radical basis	$p = 1.5$	0.7036	**0.7341**	0.6923	0.6741
Milk	Polynomial	$D = 2$	**0.6903**	0.6873	0.6660	0.6492
Late rice						
Sowing	Polynomial	$D = 2$	**0.8546**	0.8384	0.7921	0.7635
Transplanting	Linear	–	**0.8353**	0.8289	0.7536	0.7305
Tillering	Radical basis	$D = 2$	**0.8075**	0.7926	0.7706	0.7223
Heading	Radical basis	$p = 1.75$	0.7436	**0.7523**	0.7213	0.6921
Milk	Radical basis	$D = 2$	**0.7028**	0.6950	0.6722	0.6431

SVC, support vector machine classification.
The bold values means the combinations of kernel functions and parameters which performed best in the same developmental stages.

Table 6 Optimization of each developmental stage for sample class two (with unadjusted C).

Middle-season rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Polynomial	Radical basis	Polynomial
Hyperparameter	$D = 2$	–	$D = 2$	$p = 1.75$	$D = 2$
Penalty coefficient	$C = 1$	$C = 1$	$C = 1$	$C = 3$	$C = 1$
F1	0.8455	0.8255	0.8000	0.7532	0.6933
Early rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Polynomial	Radical basis	Polynomial
Hyperparameter	D = 1	–	D = 2	P = 1.5	D = 2
Penalty coefficient	C = 1	C = 1	C = 1	C = 2	C = 1
F1	0.8075	0.8066	0.7990	0.7341	0.6903
Late rice					
Developmental stage	Sowing	Transplanting	Tillering	Heading	Milk
Kernel function	Polynomial	Linear	Radical basis	Radical basis	Radical basis
Hyperparameter	$D = 2$	–	$D = 2$	$p = 1.75$	$D = 2$
Penalty coefficient	$C = 1$	$C = 1$	$C = 1$	$C = 2$	$C = 1$
F1	0.8546	0.8353	0.8075	0.7523	0.7028

Trafalis et al., 2007; Van Gestel et al., 2001c), excessively highly values of D and p were found no to be advantageous to nonlinear space mapping.

The contribution of the penalty coefficient to improving SVM segmentation was very limited. It led to a significant increase in the computation complexity (see Figs. 3 and 4). The actual crop modeling was a complicated nonlinear problem and it was very difficult to achieve optimal segmentation using a high-dimensional SVM. The determination of the optimal interface of the sample classes was thus often quite difficult. A higher penalty coefficient only increased this difficulty of the SVM identifying the optimal interface, which happened to be of no benefit to the present study (Gunn, 1998; Suykens et al., 2001).

4.2. Effect of negative samples

In the SVC training, the quality of the negative samples had greater effect on the results than the kernel functions and hyperparameters. The information used for the SVC learning was provided by both the negative and positive samples to ensure optimal interfacing. However, for the developmental SVC, the positive samples were determined by the dynamic

Table 7 Optimization of yield prediction at each stage.

Planting type	Prediction point	Sample	Kernel function	Hyperparameter	RMSE (kg/h m^2)	RE (%)
Middle-season rice	Tillering stage	Tillering 1	Radical basis	$p = 1$	126.8	22.1
	Heading stage	Heading 2	Radical basis	$p = 1.25$	96.4	17.1
	Milk stage	Milk 2	Radical basis	$p = 1.5$	109.4	19.2
Early rice	Tillering stage	Tillering 1	Polynomial	$D = 4$	88.3	20.5
	Heading stage	Heading 2	Radical basis	$p = 1.25$	68.0	15.8
	Milk stage	Milk 2	Radical basis	$p = 1.25$	36.4	8.5
Late rice	Tillering stage	Tillering 1	Radical basis	$D = 4$	89.2	21.0
	Heading stage	Heading 2	Radical basis	$p = 1.25$	69.7	16.5
	Milk stage	Milk 2	Radical basis	$p = 1.5$	46.5	11.1

RMSE, root-mean-square error; RE, relative error.

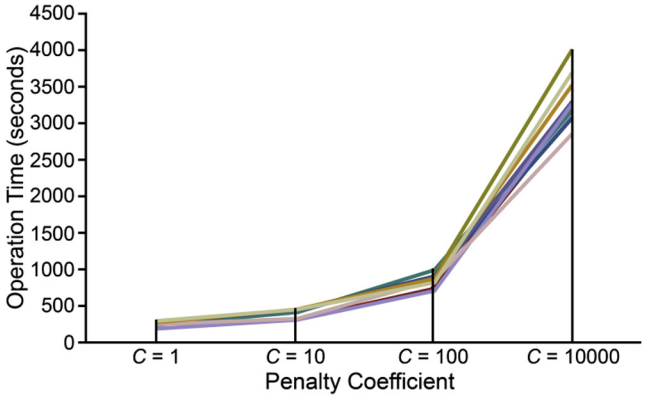

Figure 4 Relationship between the C value and efficiency (yield module). All the models were tested using $C = 1$, 10, 100, and 10,000 in the yield module. The operation time increased with the increase in the C value.

variables generated by the daily factors that were inputted to the open reading frame of "the occurrence day of a certain developmental stage." The use of the negative samples was more difficult. Theoretically, any open reading frame that does not correspond to the occurrence day of a certain developmental stage can generate negative samples. However, in practice, it is necessary to maintain proper "distance" between the positive and negative samples so that the SVC can achieve perfect classified learning.

The off-season and 30-day-in-advance samples were used in this study. The positive samples generated by the off-season strategy contributed to the improvement of the SVC sensitivity for developmental stage predictions, while low false positivity was observed in the 30-day strategy, as well as much greater SVC learning difficulty.

The question thus arises about whether the use of 30-day-in-advance samples is a proper alternative. Apparently not because the meteorological factors differed significantly for a three-month time difference. The SVC sensitivity would thus increase with the possible increase in false positivity. Actually, the 30-day-in-advance strategy still showed rather high false positivity because development is a complicated ecological, physiological, and biochemical process. The process not only includes metabolism and nutrition and water physiologies related to many enzyme systems, but also involves the cultivation environment (sunlight, temperature, water, fertilizer, soil, air, etc.) and the degree of coordination between the source, sink, and flow of the ecosystem. Furthermore, the thermo-sensitivities, photo-sensitivities, and basic vegetative growths of middle season rice, early rice, and late rice are not identical, being DNA-controlled. Concisely, genotype + environment = phenotype. Incidentally, there were phenotype differences (morphology, plant type, maturity, resistance, fertility, etc.) among the three rice and planting types. It would thus be difficult to improve the SVC accuracy by changing the sampling strategy, hence the need for new factors and methods.

Table 8 Results of SVR using scanning C value.

Planting type	Prediction point	Kernel function	Hyperparameter	$C = 1$	$C = 10$	$C = 100$	$C = 10000$
Middle-season rice	Tillering stage	Radical basis	$p = 1$	22.1	**21.1**	21.1	21.5
	Heading stage	Radical basis	$p = 1.25$	17.1	**16.4**	17.9	18.1
	Milk stage	Radical basis	$p = 1.5$	19.2	**18.3**	19.3	20.1
Early rice	Tillering stage	Polynomial	$D = 4$	20.5	19.6	**17.8**	18.1
	Heading stage	Radical basis	$p = 1.25$	**15.8**	16.6	18.1	18.1
	Milk stage	Radical basis	$p = 1.25$	**8.5**	8.9	8.9	8.9
Late rice	Tillering stage	Radical basis	$D = 4$	**21.0**	21.9	21.9	21.9
	Heading stage	Radical basis	$p = 1.25$	**16.5**	17.2	17.3	17.3
	Milk stage	Radical basis	$p = 1.5$	11.1	10.6	**10.2**	11.5

SVR, support vector machine regression.
The bold values means the combinations of kernel functions and parameters which performed best in the same developmental stages.

Table 9 Optimization of each developmental stage for yield prediction (with unadjusted *C*).

Planting type	Prediction point	Sample	Kernel function	Hyperparameter	Penalty coefficient	RE (%)
Middle-season rice	Tillering stage	Tillering 1	Radical basis	$p = 1$	10	21.1
	Heading stage	Heading 2	Radical basis	$p = 1.25$	10	16.4
	Milk stage	Milk 2	Radical basis	$p = 1.5$	10	18.3
Early rice	Tillering stage	Tillering 1	Polynomial	$D = 4$	100	17.8
	Heading stage	Heading 2	Radical basis	$p = 1.25$	1	15.8
	Milk stage	Milk 2	Radical basis	$p = 1.25$	1	8.5
Late rice	Tillering stage	Tillering 1	Radical basis	$D = 4$	1	21.0
	Heading stage	Heading 2	Radical basis	$p = 1.25$	1	16.5
	Milk stage	Milk 2	Radical basis	$p = 1.5$	100	10.2

RE, relative error.

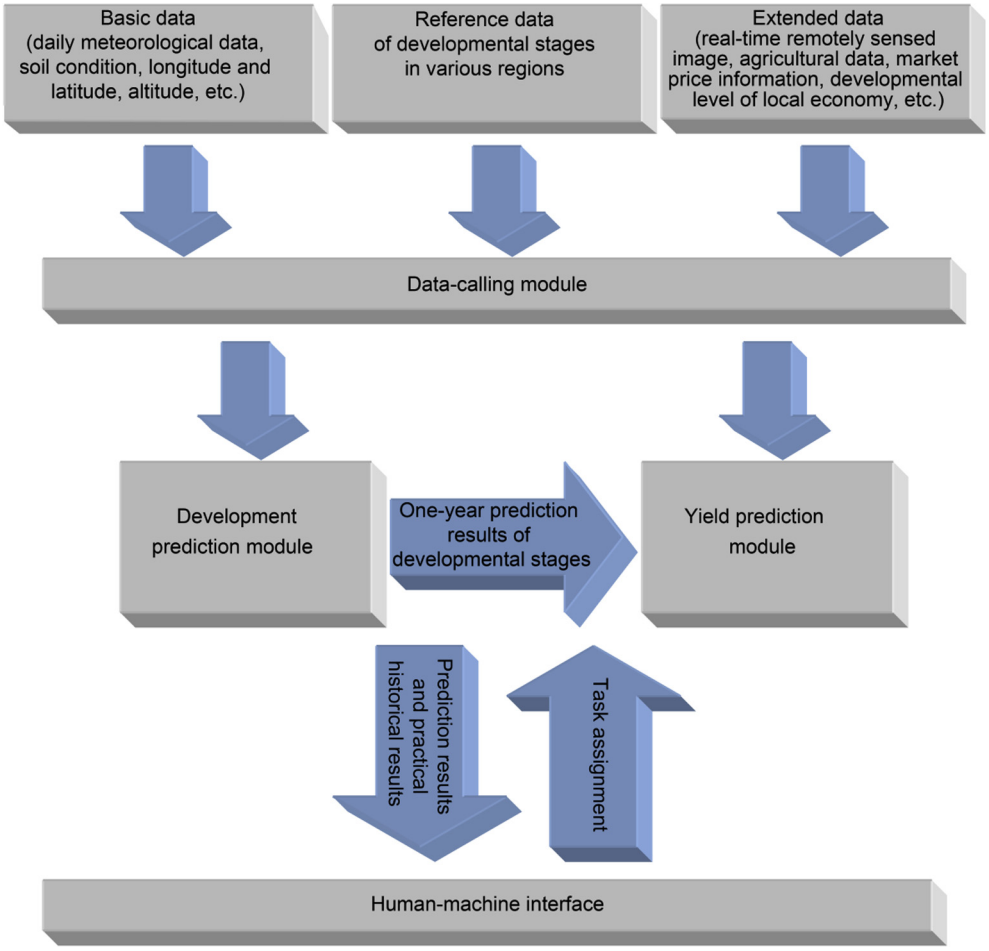

Figure 5 An SVM-based open crop model (SBOCM) was designed as an application-focused crop model for regional rice development stage prediction and yield prediction. The emphasis was on simplicity of operation. The framework of the entire SBOCM system comprised five parts, namely, the database module, data calling module, development prediction module, yield prediction module, and human–machine interface.

4.3. Error source analysis

4.3.1. Mechanism problems

Classic crop models are based on decades of research on the internal growth and developmental mechanisms of crops and are highly accurate for field application. As a typical machine learning method, SVM is outstanding for nonlinear fitting. It is characterized by a simple framework and explicit input and output, although insufficient attention is given to the internal physiology of the investigated object. SVMs were used in this study to explore the feasibility of applying machine learning to recent crop modeling, and the potential of an SVM open framework. The target was to achieve a prediction accuracy

comparable to that of a simple SVM within a short time. This was found to be apparently impossible.

4.3.2. Sample defects

This study mostly employed historical data obtained from the China Meteorological Administration, generated from stations with uncontrolled data quality. After preprocessing, the data still contained defects such as artificial errors, limited factors, excessive effects of macroscopic soil factors, too short open reading frame, insufficient field experiment data for referencing and correction, features that could not be explained by meteorological factors, and limited sample size. These defects could not be fully offset by statistical quality control and thus affected the SVM training results.

4.3.3. Regional differences

The yield prediction results contained only small errors for China's south-eastern coastal areas, but large errors for the western inland and north-eastern areas. There were two reasons for this: (1) there were more sample records for the south-eastern coastal areas, and this improved the corresponding learning. (2) The effects of other regional factors apart from longitude, latitude, and altitude, which were considered during the learning process; possibly including relative humidity (eliminated after PCA) and light angle.

4.3.4. Effects of water and fertilizer management

There is a common problem of current crop modeling, wherein modeling under certain production conditions is inapplicable to a complicated regional simulation. Hence, to simplify regional simulation, production management was excluded from the factors considered in the present study, and this reduced the accuracy of the model. Generally, paddy field rice is more affected by human water and fertilizer management than by natural and meteorological conditions. The soil-related factors too were more affected by human water and fertilizer management than by natural and meteorological conditions. The lack of this type of data reduced the accuracy of the study results.

It is certain that the irrigation and water conservation conditions in present China are undesirable. Rainfall is the main source of water supply and this was considered in the regional simulation in this study. In future adjustments of the proposed models, it would be necessary to consider the possibility of combining them for real filed production, and to examine the relationship between natural rainfall and human water and fertilizer management. The water and fertilizer management factors should also be simplified on the regional scale and organically integrated with the SBOCM for enhanced performance.

5. Conclusions

The SVM machine learning method was used to develop an SBOCM with simplified data acquisition, suitable for regional simulation, and that can be effectively integrated with multiple scale factors for early-stage theoretical investigations. The model input side is open to future integration of additional natural and social factors to improve the practicability and prediction accuracy. The samples used in this study were built through quality control of mass data. Dimensional reduction was done by factor analysis methods such as PCA and the models were evaluated by fivefold cross-validation. The objective of the SVM modeling was to determine the optimal kernel functions, hyperparameters, and penalty coefficient to enable separate investigations of three types of rice plantings and the several developmental stages. We found that the penalty coefficient made limited contribution to model optimization and therefore first determined the optimal kernel functions and hyperparameters, and then optimized the models by adjustment of the coefficient. The search efficiency was thusly improved fourfold.

The SVM modeling method proposed in this paper basically utilizes scale-independent factors and has an open input framework, which facilitates integration with large-scale data for scaling up. Because agricultural production involves both natural and socio-economic inputs, factors such as grain price, fertilizer price, seed price, labor cost, location, traffic conditions, governmental support, real status of agriculture, and scientific and cultural innovations may be further integrated into the proposed model to enable more robust simulation.

Acknowledgements

This work was supported by a Grant from the National High Technology Research and Development Program of China (863 Program) (No. 2007AA10Z220). The authors wish to thank Ms. Jiang Hong for the help in inproofreading this paper.

References

Cai, C.Z., Wang, W.L., Chen, Y.Z., 2003. Support vector machine classification of physical and biological datasets. Int. J. Mod. Phys. C 14, 575–585.

Charles-Edwards, D.A., 1986. Modelling Plant Growth and Development. Academic Press, London.

Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Curry, R.B., 1971. Dynamic simulation of plant growth I. Dev. Model. 14, 946–959.

Dhungana, P., Eskridge, K.M., Weiss, A., Baenziger, P.S., 2006. Designing crop technology for a future climate: an example using response surface methodology and the CERES-wheat model. Agric. Syst. 87, 63–79.

Du, H.Y., Wang, J., Hu, Z.D., Yao, X.J., Zhang, X.Y., 2008. Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. J. Agric. Food Chem. 56, 10785–10792.

Edwards, D., Hamson, M., 1990. In: Guide to Mathematical Modeling, vol. 2. CRC Press Inc., Florida.

Gao, L.Z., 2004. Foundation of Agricultural Modeling. Pegasus Book Co., Ltd, Hong Kong.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42, 1033–1046.

Gualtieri, J.A., Cromp, R.F., 1998. Support vector machines for hyperspectral remote sensing classification. In: 27th AIPR Workshop on: Advances in Computer-Assisted Recognition, vol. 3584. SPIE-International Society of Optical Engineering, Washington, D. C., pp. 221–232.

Gunn, S.R., 1998. Support Vector Machines for Classification and Regression: Technical Report. University of Southampton, Southampton.

Jones, C.A., Kiniry, J.R., 1986. CERES-Maize: A Simulation Model of Maize Growth and Development. A&M University Press, Texas.

Kaundal, R., Kapoor, A.S., Raghava, G.P.S., 2006. Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC Bioinform. 7, 485.

Lin, Z.H., Mo, X.G., Xiang, Y.Q., 2003. Research advances on crop growth models. Acta Agron. Sin. 29, 750–758.

Luo, S.M., Zhen, H., Chen, C.H., Yang, W., 1990. Study on application of computer simulation in rice high yield cultivation. Guangdong Agric. Sci. 3, 14–17.

Penning de Vries, F.W.T., Jansen, D.M., ten Berge, H.F.M., Bakema, A., 1989. Simulation of Ecophysiological Processes of Growth in Several Annual Crops. Simulation Monographs. PUDOC, Wageningen.

Ritchie, J.T., 1972. Model for predicting evaporation from a row crop with incomplete cover. Water Resour. Res. 8, 1204–1213.

Shi, X.Z., Yu, D.S., Pan, X.Z., 2002. 1:1000000 Soil Database of China. Institute of Soil Science, Chinese Academy of Sciences, Nanjing.

Sinclair, T.R., Seligman, N.G., 1996. Crop modeling: From infancy to maturity. Agron. J. 88, 698–704.

Stewart, D.W., Dwyer, L.M., 1990. A model of spring wheat (*Triticum aestivum*) for large area yield estimations on the Canadian Prairies. Can. J. Plant Sci. 70, 19–32.

Suykens, J.A.K., 2001. Nonlinear modelling and support vector machines. In: 18th IEEE Instrumentation and Measurement Technology Conference, Budapest.

Suykens, J.A.K., De Vandewalle, J., Moor, B., 2001. Optimal control by least squares support vector machines. Neural Network 14, 23–35.

Swain, D.K., Heranth, S., Saha, S., Dash, R.N., 2007. CERES-rice model: calibration, evaluation and application for solar radiation stress assessment on rice production. J. Agrometeorol. 9, 138–148.

Trafalis, T.B., Adrianto, I. Richman, M.B., 2007. Active learning with support vector machines for tornado prediction. In: 7th International Conference on Computational Science, Beijing.

Van Gestel, T., Suykens, J.A.K., Baestaens, D.E., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., Vandewalle, J., 2001a. Financial time series prediction using least squares support vector machines within the evidence framework. IEEE Trans. Neural Network 12, 809–821.

Van Gestel, T., Suykens, J.A.K., De Brabanter, J., De Boor, B., Vandewalle, J., 2001. Least squares support vector machine regression for discriminant analysis. In: International Joint Conference on Neural Networks, Washington, D.C.

Van Gestel, T., Suykens, J.A.K., De Moor, B., Vandewalle, J., 2001. Automatic relevance determination for least squares support vector machine regression. In: International Joint Conference on Neural Networks (IJCNN 01), Washington, D.C., pp. 2416–2421.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Network 10, 988–999.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, J.A.K., Van den Poel, D., Vanthienen, J., De Moor, B., Dedene, G., 2001. Knowledge discovery in a direct marketing case using least squares support vector machines. Int. J. Intell. Syst. 16, 1023–1036.

Xie, Y., James, R.K., 2002. A review on the development of crop modeling and its application. Acta Agron. Sin. 28, 190–195.

Xiong, W., 2004. Modeling of Chinese Main Crops Based on Future Climate Change. China Agricultural University, Beijing.

Xiong, W.W., 2009. The Study of Face Recognition Method Based on Mixture Kernel Function Support Vector Machine. Wuhan University of Science and Technology, Wuhan.

Xu, X.M., Mao, Y.F., Xiong, J.N., Zhou, F.L., 2007. Classification performance comparison between RVM and SVM[C]. In: International Workshop on Anti-counterfeiting, Security, and Identification, Xiamen, IEEE, pp. 208–211.

Yang, X.H., Huang, J.F., Wang, X.Z., Wang, F.M., 2008. The estimation model of rice leaf area index using hyperspectral data based on support vector machine. Spectrosc. Spectral Anal. 28, 1837–1841.

Yu, H.Y., Lin, H.J., Xu, H.R., Ying, Y.B., Li, B.B., Pan, X.X., 2008. Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy. J. Agric. Food Chem. 56, 307–313.

Zhang, X.M., 2009. The Study of Audio Classification Based on Wavelet and Support Vector Machine. Yanshan University, Beijing.