Research Journal of Applied Sciences, Engineering and Technology 4(16): 2851-2856, 2012
ISSN: 2040-7467

Submitted: March 31, 2012 Accepted: May 12, 2012 Published: August 15, 2012

SVR-D1.2: A Prediction Model for Population Occurrence of Paddy Stem Borer

[1]Lichuan Gu, [2]Jinqin Zhong and [1]Youhua Zhang
[1]School of Information and Computer, Anhui Agriculture University, Hefei 230036, China
[2]School of International Business, Anhui University, Hefei 230031, China

Abstract: In this study, we analyse the SVR-based prediction method for selecting the optimal model framework based on kernel matrix. Moreover, SVR-D1.2 is proposed with the help of the kernel matrix's symmetry and positive definition and kernel alignment. Test results show that there exactly exists the non-line relation between the insect population occurrence and the meteorological factors and the new prediction model, SVR-D1.2, improved prediction accuracy compared with other methods.

Keywords: Feature selection, kernel alignment, rice paddy stem borer, support vector regression

INTRODUCTION

Paddy stem borer (Scirpophaga incertulas) is an important ricepest in tropical and sub-tropical Asia. This monophagous insect feeds only on paddy rice. Insect larvae bore into the plant and feed on leaf-sheath tissue, on tassel buds and on the stem. Damaged plants wither and their tassels die or become white and infertile leading to decreased grain production. Forecasting of the development of population occurrence with as much accuracy as possible and describing the population dynamics correctly enable best control of this pest and thus best damage minimizations.

Nowadays, there are three main ways to forecast pest occurrence. These predictions are based on experience, experiment and statistical method as the following.

- **Experience:** Many useful predictions are made by experts based on their practical experience gained over long periods of time. Experience prediction is used all round the world to predict the occurrence of pest outbreaks with short lead times and these predictions play an important part in guiding useful intervention in agricultural production. There is, however, limitations with experience prediction (Aaron *et al*., 2008; Chih and Chiu, 2002; Kim and Yoon, 2004).

First, the relationship between the occurrence of a pest and various outside factors is not a simple, linear one. This means that the simple correlations used in guiding experience prediction are hard to extrapolate to develop forecasts having high accuracy (Rosset and Neumann, 2003; Nath, 2003; Muller *et al*., 1999).

Secondly, experience prediction is a method able to make predictions with only quite short lead times, not with the longer lead times often required to effectively mitigate a pest population outbreak.

- **Experiment:** This is a prediction method based on a good knowledge of the pest's life table. Experiment prediction was applied in studies outbreak 'hot spots' in the This is a mathematical method based on probability and multiple-factor principles and is used to identify the key driving factors and to build a corresponding prediction model.
- **Statistical method:** Statistical prediction has recently become the main prediction method. Many statistical methods are used today, such as regression analysis, stepwise regression, stepwise discriminate analysis, principle component analysis, multiple correlation analysis, fuzzy mathematics and systematic analysis and so on.

However, statistical prediction can have problems. Some of these are that:

- The prediction effect is not steady so that the usefulness of the models often varies under real-life conditions. In practice this means that the fitting of equations is sometimes good but, at other times it poor.
- Some statistics predictions are based on general investigations but not on ecological theory. Few include considerations of pest biology or the laws of physics. In the absence of ecological theory, the occurrence of the pest cannot be accommodated by statistical theory.

With the rapid development of nonlinearity science, we can combine traditional mechanical theory, statistical theory, chaos theory and some new mathematical

Corresponding Author: Lichuan Gu, School of Information and Computer, Anhui Agriculture University, Hefei 230036, China

calculation technologies to develop a new approach for predicting pest occurrence. SVR (Support Vector Regression) is an appropriate and good choice for use in pest occurrence forecasting. We present a forecasting methodology combining dynamic feature selection with regression models where we propose Support Vector Regression for model construction. Our methodology, however, is independent of the particular regression model, i.e., any other regression approach can be used within the proposed framework (Ruiz and Lopez-de-Teruel, 2001; Bernhard and John, 2001; Manevitz and Yousef, 2001; Yi *et al*., 2004; Li and Luo, 2005).

In this study, we review the basic knowledge of feature selection and support vector regression reviewed. In addition, a Support Vector Machine (SVM) regression model for prediction based on dynamic feature selection (SVR-D) is presented. Moreover, the experiment was done with SVR-D1.2 in the Paddy stem borer and the comparative analysis was made with the traditional methods. The results showed that the method has a considerable increase in the effectiveness and the degree of fitting with the existing methods.

LITERATURE REVIEW

Features selection: As for the complexity of data mining application, such as regression problem, the most important features are needed to be selected to construct a suitable model. Feature selection method research in this field is a important hotspot problem. Features selection has following benefits for model:

- Improve the accuracy of the model
- Reduce the computation time during constructing the model
- Facilitate data visualization and model understandability
- Reduce additional risks

At present, the features selection methods can be divided into three main types: filtering, packing and embedding.

Filtering method treats the features selection as a pre-process, which is independent of learning algorithm for model construction. An example for this is variable sort, which take advantage of each characteristic and correlation coefficient between interdependent variables to accomplish. Another filtering method is to select features based on linear model (which is corresponding to pre-process) and then construct the non-linear model with the selected features. Apparently, filtering mechanism is not dependent on the regression algorithm of application, so it is nothing to do with the selected features. It is the main disadvantage of the method. However, it is frequently adopted for its simpleness and understandability. In addition, the general filter methods do not consider the problem of multicollinearity.

Packing is a method defining a selection of subset. Its main idea is to assess the variables subset according to the variables' validity for given learning algorithms which is considered as a black box. The best features subset is determinated by the specific algorithm used for constructing the regression model (e.g., linear regression or nonlinear regression, neural network, SVR and etc.). The packing method needs a criterion to compare the results from different features subset (for example, criterion may be the minimum of average and absolute training error) and a search strategy with supervised process.

Forward selection and backward exclusion are widely used in search strategies. Forward selection strategy begins with a null feature set and the weights of their related prediction will be considered in iteration. While the backward exclusion starts at the available features from the selected features subset. The features of minimal reduction in pre-process will be deleted in each iteration. In order to acquire the best features subset, a stopping criterion is needed for these two strategies. Although it demands high computing capability, the advantages are the predictive model's output is considered and how the given features subset is processed.

Embedding method takes feature selections as a part of training process. Features selection will not be executed until construct prediction model. This mechanism usually contains the changes in the objective function of the learning algorithm, which is related with specific predictors.

SVM: The SVM is a robust classifier superior to many other classifiers in the scenarios of binary classifications. Converted into a vector space, a classification problem in SVM is to find a decision surface which separates the data points into two classes by solving the quadratic optimization problem. The SVMs combine two powerful ideas: maximum margin classifiers with low capacity and implicit feature space defined by kernel function. In a linearly separable space, the surface that separates the two classes is a hyper-plane that maximizes the margin-the distance between the parallel hyper-planes that separate the examples of the two classes. In a non-linearly separable space, SVMs use kernel functions to transfer the original data into a higher dimensional feature space where data points become linearly separable. Commonly used kernel functions include linear, sigmoid, polynomial and Gaussian kernels. Theoretically, a kernel function can be characterized as an estimating function of the data which reveal something interesting about the underlying distribution. Given a set of training data, there is no prior knowledge about the selection of kernel function. Empirical results show that SVMs turn out to be quite

sensitive to the representation and kernel in ways they are not well understood. In a non-linear separable space, SVMs use slack variables x_i for each pattern and a parameter to penalize the slack and result in soft margin hyperplanes.

The notion of SVM is briefly presented as follows:

Let X1, …, Xl be the training examples from χ and Φ:X→H be a kernel map which transforms the training examples into a RKHS space H. The training data are separated from the origin that is thought of as the second class by solving the following quadratic programming problem:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_{i=1}^{l} \xi_i - \rho \quad (1)$$

Subject to:

$$\left(w.\Phi(x_i)\right) > \rho - \xi_i \qquad i = 1,2,\ldots,l \qquad \xi_i \geq 0 \quad (2)$$

By solving the optimal objective function for w and r, the classification rule becomes:

$$f(x) = \text{sign}((w.\ \Phi(x)) - \rho) \quad (3)$$

In the above model, x_i (i = 1,…,l) are slack variables used in the non-linearly separable space and v is the parameter that restrict a fraction of normal examples outside the region. The commonly used kernels are linear kernel function, polynomial, Radial Basis Function (RBF), sigmoid kernels.

In SVM, only part of the training samples or SV plays a key role in supervised learning, which implies that SV is sufficient in characterizing the whole training samples. Therefore, using SV rather than the whole training samples to train the classifier will greatly improve learning efficiency without decreasing classification precision.

METHODOLOGY

SVR based model framework SVR-D1.2: This section gives a predictive model framework based on SVR, which includes all key tasks like feature selection, model construction, model validation which are necessary for the model establishment.

SVR based Model Framework SVR-D1.2 is shown in Fig. 1. The framework include 6 general steps, which will be described as follows:

Step 1: Data division: At this stage, the data need to be divided into training subsets, validation subset and test subset. In which, training subset is used to construct the model, validation subset is used for model and feature selection, test data is fully independent subset for supplementary evaluate the level of error between the realistic model. These subsets can be redefined dynamically when it is validated.

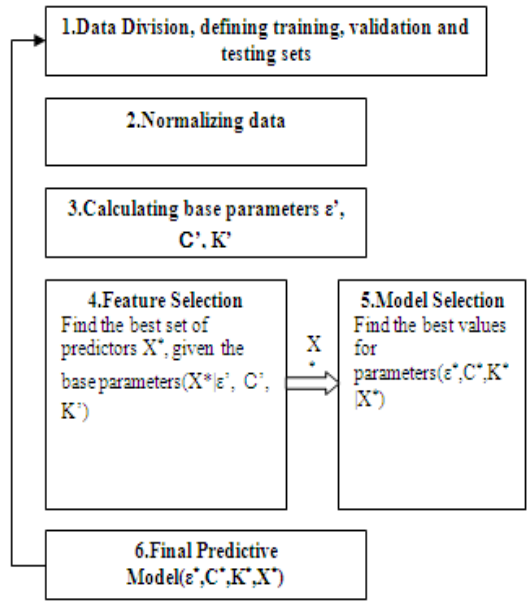

Fig. 1: SVR-D1.2 predictive model framework

Step 2: The basic parameters calculation: Mainly determined the ε′, C′ and the kernel parameters K'. In certain conditions, these parameters will play a role.

Step 3: Features selection: The features will be selected with the help of these parameters and packing method and the optimum predictor variable χ* set will be obtained with the forward choice strategy.

Step 4: Model construction: Obtain the optimum parametersϵ*, C*, K* by utilizing the predictor χ* and searching in the basic parameters space.

Step 5: Model selection: Determine the prediction model withϵ*, C* obtained in step 4, kernel function K* and predictor χ*.

Step 6: SVR-D1.2 model also defines the validating strategy which can help to validate the model as parameters updating dynamically.

Step 6: Obtain the predictive model

SVR based predictive modeling framework: Obtain the model parameters (ϵ*, C*, K*, X*), then iterating from step 1 until convergence.

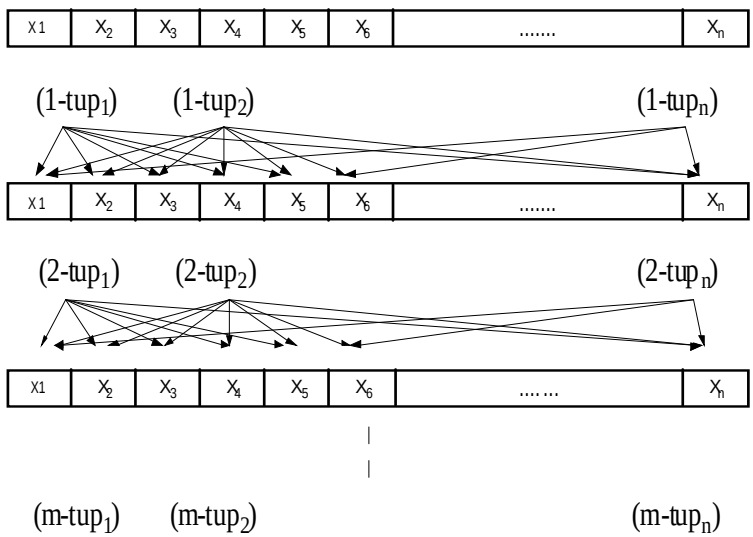

Fig. 2: Features selection

Choice of optimization model based on kernel adjustment: One of the key problems in SVR is the choice of model arguments ε′ and C′ and the kernel function using data for High-dimensional space. As shown in Fig. 2, the model in SVR-D1.0 is constructed via calculating the initial arguments and then carrying on a search with them.

Experience rules are used to compute the initial value of ε′ and C′. There must exist an optimized kernel function to minimize the test error rate for the same SVM classification problem. So the selection of optimal kernel function is one of the key problems in constructing model where the difference among kernel functions can be measured. Because the relationship in kernel functions is equivalent to kernel matrix, kernel alignment is introduced to measure the difference among kernel functions. The definition of kernel alignment is as follows:

Define 1: Given sample set U = {x_1, ……, x_n}, P_1 and P_2 are for kernel function Φ (x_i) respectively. Its inner product can be expressed as (4):

$$\langle P1, P2\rangle F = \sum_{i,j=1}^{1} P_1(x_i, x_j) P_2(x_i, x_j) \tag{4}$$

Now, scalar ∏ (U, P_1, P_2), named kernel alignment, to calculate the error between P_1 and P_2.

$$\prod (U, P_1, P_2) = \frac{\langle P_1, P_2\rangle}{\sqrt{\langle P_1, P_2\rangle_F \langle P_2, P_2\rangle_F}} \tag{5}$$

From the point of geometry view, kernel alignment is the cosine value of the angle between the two vectors. On the basic concept of kernel alignment, we try to find the optimal kernel function, the algorithm is as follows:

Step 1: Construct a Lagrange equation with parameters and multiplier, as shown in (6):

$$L(\sigma, \alpha) = f(\sigma) + \alpha^T c(\sigma) = f(\sigma) + \sum_{i=1}^{m} ac(\sigma) \tag{6}$$

Step 2: Define function f (σ), as shown in (7):

$$f(\sigma) = \frac{(K, YY^1)_F}{l\sqrt{(K, K)_F}} \tag{7}$$

where, l is number of sample, K is kernel matrix, Y is set of the class identifier.

Step 3: Define the function c (σ), as shown in (8):

$$c(\sigma) = [(K, K) - 1(\leq)0; - K(\leq)0] \tag{8}$$

where, K is the kernel matrix.

Step 4: Construct sub-Quadratic Programming problem, as shown in (9), then reevaluating formula (6): min Δ1/2σ THΔσ+f(σ)TΔσ

$$\text{s.t. } \nabla\sigma ci(\sigma)\ T\ \Delta\sigma + ci(\sigma) \leq 0,\ i = 1, \ldots, m \tag{9}$$

where, H is the positive Hansen matri.

Step 5: Solving (9) and update H.

Step 6: Go to Step 4 until convergence and obtain the optimal σ.

Feature selection: The method of features selection is very important for prediction model. A suitable method can improve the accuracy of prediction, reduce the computation time, facilitate data visualization and model understandability, low additional risks.

Figure 2 shows the forward strategy and packing method adopted in this study.

Suppose the initial feature set $X = (x_1, x_2, \ldots\ldots x_n)$, define a predictor maximum m ($m \leq n$) according to the problem and the number of training data in prediction model. The iterate K will also be defined. The strategy search space will increase as K gets bigger. Once define two variables, each can be set as a separate predictor. When in the 2nd iteration, reserve the optimal predictor (1-tuple) and mix with the rest variables as the 2-tuple in the next iteration and so on, until obtain the optimal m-tuple. The predictor x* there is selected for SVR model.

Model validation: As a time series problem, the prediction can be affected by the related factors and lead to low the accuracy and even failure. A constructed model may work well in within certain period, but in the future it may result out an error in prediction because of the factors' change. In order to solve this problem, a validation module is included in SVR-D1.2. Here two sets are used: one for historical data, the other is for the latest data. If new data arrives, it will be merged in the training data set. When the model is validated, the ratio of training data to effective data in working set remains unchanged as the data transfer from training set to working set.

EXPERIMENT

In HuiZHou County, paddy stem borer usually goes through four generations a year. The 1st generation is from the last 10 days of March to the beginning of May when the pest feeds on the sprouts of early-transplanted rice seedlings. The actively growing shoots of damaged plantswither. The 2nd generation is usually from the first 10 days of June to about mid July. Mid-season, planted rice suffers most damage from this pest generation and plants appear to have white heads (tassels). The 3rd generation is from the beginning of August through to the last 10 days of September, late-growing plants also suffer damage from this pest. The last generation of paddy stem borer begins in October and continues over winter. This study applied the new method to analyze and predict the population occurrence of the 1st generation.

An experiment was conducted by applying the proposed SVR-D1.2 method for forecasting population occurrence of paddy stem borer. By selecting the initial features related with the outbreak of wheat scab, the SVR-D1.2 was applied to 2 time series data: CDCR data and BDCR data. As a result, different parameters sets are obtained for each series data, which can describe its own model. According to the description of packing strategy in 3.3, we also determine the selected features set for each series. In our experiment, it can be found that using SVR-D1.2 method with a periodic index and a binary variable of a month (Some 2 weeks, some 3 weeks) can acquire the most closely linked features.

Table 1: Ratio of average absolute error

	Testing data set		
	ARMAX	BP-ANN	SVM-D1.2
CDCR	10.68	12.23	11.9
BDCR	13.45	12.76	12.54
Avg. error	12.1	11.73	11.3

Table 2: results comparison for several methods

	Testing data set			
Model type	Accuracy	Recall	Overlay	Improved coefficiency
ARMAX	0.8669	0.869	0.4562	6.206
BP-ANN	0.8702	0.802	0.5251	6.357
SVM-D1.2	0.873	0.903	0.68	7.586

To verify the effectiveness of the system, we take 1980-1983 as the target years and test the prediction every ten days (a xun). The average of the correlation coefficient will be calculated using year as unit. And the correlation of the predictive value is analyzed statistically with the BP-ANN's and the ARMAX's.

Table 1, 2 shows the Cluster Disease Carrier Rate (CDCR) correlation coefficient, the Branch Disease Carrier Rate (BDCR) correlation coefficient and the Means Absolute Proportion Error (MAPE), respectively.

Table 1 show the results for three types of prediction methods. It is easy to discover that the average level of error acquired by SVR-D1.0 is better than the other twos. Table 1 and 2 show the accuracy error measures of Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), obtained over the test set, by using the 3 forecasting methods for predicting one period ahead sales for the five products.

When predict pest occurrence, the first task is to buildup a database of the relevant factors affecting pest occurrence. Then SVR-D1.2 systems should be allowed to learn this pest information from the database. After learning, the system can become an "expert" able to study and to gain knowledge from the database. Lastly, the system can judge and speculate according to the information gained to forecast future pest occurrence.

The advantage of SVR-D1.2 application research in pest forecasting is that the pest occurrence prediction is essentially a Basic Input Output System (BIOS). The data

conversion relations in the system include numeric fitness, fuzzy conversation and logical speculation. All of the above can be expressed by SVR-D1.2. Thus SVR-D1.2 can be used widely in pest forecasting. However, SVR-D1.2 is unfamiliar to most entomology researchers so its use is not widespread in entomological research.

CONCLUSION

SVM is a kind of general learning algorithm based on statistical learning theory, which can solve the problems of non-linear, high-dimensional and local minima for its solid theoretical basis. The proposed predictive model SVR-D1.2 combines the dynamic feature selection with SVR optimization model. It has good performance in prediction, the most relevant features selection and validation for the model itself. And it can be adjusted to improve its prediction performance as the change of related factors. However, because of the research in the field was started for a short time, some problems like optimal selection in kernel function and parameters, the weights for samples can be worthy of continued exploration and research.

We have applied the proposed methodology using SVR as well as neural networks as regression models to a sales forecasting problem and compared its performance to a standard ARMAX approach. Comparing the respective results shows that our methodology performs slightly better and additionally provides a selection of the most important features. This last point increases the comprehension of the phenomenon we are studying and could be useful to provide a better understanding of the regression model. Major advantages of the proposed methodology are expected when dealing with dynamic phenomena, where the performance of a forecasting model could be significantly improved by performing model updating.

Future study has to be done for predicting nonseasonal time series and for selecting the most appropriate parameters of the kernel function based on theoretical approaches.

ACKNOWLEDGMENT

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues. And also thank HeFei University of technology to provide part hardware. This study is supported in part by the National Natural Science Foundation of china (Grant No. 30800663 and 70871033).

REFERENCES

Aaron, C., M. Cristopher and M.E.J. Newman, 2008. Hierarchical structure and the prediction of missing links in networks. Nature, 453: 98-101.

Bernhard, S. and C.P. John, 2001. Estimating the support of a high-dimensional distribution. Neur. Comput., 3(7): 1443-1471.

Chih, P.W. and I.T. Chiu, 2002. Turning telecommunications call details to churn prediction: A data mining approach. Exp. Syst. Appl., 23(2): 103-112.

Kim, H.S. and C.H. Yoon, 2004. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommun. Polic., 28(9): 751-765.

Li, X. and B. Luo, 2005. Optimal model selection for support vector machines. J. Comput. Res. Dev., 42(4): 576-558.

Manevitz, L. and M. Yousef, 2001. One-class SVMs for document classification. J. Mach. Learn. Res., 1(2): 139-154.

Muller, K.R., A. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen and V. Vapnik, 1999. Predicting Time Series with Support Vector Machines. In: Scholkopf, B., C.J.C. Burges and A.J. Smola, (Eds.), Advances in Kerne Methods-Support Vector Learning, MIT Press, Cambridge, MA, pp: 243-254.

Nath, S.V., 2003. Data warehousing and mining:. Customer churn analysis in the wireless industry. M.B.A Thesis, Faculty of the College of Business in Florida Atlancit University.

Rosset, S. and E. Neumann, 2003. Integrating customer value considerations into predictive modeling. 3rd IEEE International Conference on Data Mining, pp: 1-8.

Ruiz, A. and P.E. Lopez-de-Teruel, 2001. Nonlinear kernel-based statistical pattern analysis. IEEE T. Neur. Network., 12(1): 16-31.

Yi, M., W. Hui and L. Lei, 2004. Multi dimensional model based clustering for user behavior mining in telecommunications industry. Proceeding of the Third International Conference on Machine Learning and Cybernetics, Shanghai, pp: 26-29.