Bangladesh J. Bot. **51**(4): 883-888, 2022 (December) Special DOI: https://doi.org/10.3329/bjb.v51i40.63830

PREDICTING POTENTIAL GEOGRAPHICAL DISTRIBUTION OF TOBACCO BLUE MOLD IN CHINA

GUOMING DU[1*]

School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, P.R. China

Keywords: Ecological niche model, Alien species, Support Vector Machine (SVM), GIS, Cross validation

Abstract

Tobacco blue mold is a serious tobacco disease, which has high risk of invasion in many areas, therefore, it is of particular importance to strengthen the quarantine work and to make a good prediction. Since each species has its own particular relatively stable ecological niche, one can predict the potential geographical distribution of tobacco blue mold in target areas like China. Based on the known distribution data of tobacco blue mold, 20 environmental factors were selected and combined support vector machine (SVM) with ecological niche model was used to predict the potential spatial distribution of tobacco blue mold and to obtain the optimal combination of SVM parameters through the brute force search algorithm. Search algorithm overcame the shortage of empirical method and improved the accuracy of prediction. The method was proved to be effective and feasible by cross validation which provided a robust theoretical basis for preventing the intrusion and spread of tobacco blue mold.

Introduction

Tobacco blue mold an oomycete plant pathogen causes yearly epidemics in tobacco in the United States and Europe (Blanco-Meneses *et al.* 2018). It is a rapidly spreading, devastating disease on tobacco, severely reducing the yield and quality of tobacco leaves and is a major disease. Blue mold, first discovered in Australian cultivated tobacco in 1891, had spread to 65 countries and regions other than China in the past 100 years and continued to harm tobacco leaves in America, Australia, Europe, the Middle East and North Africa. The loss caused by this disease has been extremely serious. For example, the tobacco blue mold epidemic in Cuba in 1980 wiped out 90 per cent of the cigar production, forcing the government to shut down temporarily the National Cigar Factories (Shi *et al.* 1996). China is a big tobacco producing country with the highest yield in the world. The cultivation covers all the climatic ecological regions in northern and southern China. Because tobacco blue mold has high adaptability and strong spreading ability in different ecological regions around the world, it poses a great threat to tobacco production in China (Zhang 1995). As a consequence, it is very important to strengthen the quarantine work and to make a good prediction of alien species. Since each species has its own particular niche, which is usually relatively stable, it is the prerequisite for ecological niche model to be able to predict the potential geographic distribution of species (Wiens *et al*. 2005). Therefore, depending on the specific habitat of each species, based on the known distribution areas of the target species, the niche demand of the target species can be induced or simulated by mathematical model, and then the suitable distribution of the target species can be obtained by projecting the niche demand to the target areas.

The prediction of the suitable areas of tobacco blue mold by ecological niche model is helpful for early warning and provides robust technical support for decision makers to draw up scientific management measures. The commonly used ecological niche models are, CLIMEX (Byeon *et al.*

[*]Author for correspondence: <eesdgm@mail.sysu.edu.cn>. [1]School of Geography and Planning, Sun Yat-sen University, 135 West Xingang RD., Guangzhou 510275, P.R. China.

2020, Stoeckli *et al*. 2020), BIOCLIM (Sindel and Michael 1992, Booth 2018), MaxEnt (Phillips *et al*. 2006, Wan *et al*. 2020, Saha *et al*. 2021), ANN (Gevrey *et al*. 2006, Oliveira *et al*. 2021), GARP (Stockwell *et al*. 1999, Ray *et al*. 2018) and SVM (Sadeghia *et al*. 2012, Baral and Haq 2020, Makmuang *et al*. 2020, Avolio and Fuduli 2021).

Support vector machines (SVM) proposed by Vapnik (Vapnik 1995) provides a new way to solve the problem of nonlinear modeling. SVM is an efficient method of machine learning with strict theoretical foundation. It has become a hot technology in the fields of computer learning, pattern recognition, computation intelligence, prediction and so on, which has been widely concerned at home and abroad.

Materials and Methods

The principle of SVM is to project points in low dimensional spaces to high dimensional spaces, making them linearly separable and then the principle of linear partition is used to judge the classification boundary. The training sample set is (x_i,y_i), among which, $x_i \in Rn$, $y_i \in R$. The general form of linear discriminant function in n-dimension space is $g(x) = (w \cdot x)+b$, and all data in the set can be divided correctly by the classifying plane $(w \cdot x) + b = 0$. The classifying plane is the optimal hyperplane and the nearest heterogeneous vector to the optimal hyperplane is the support vector.

The quadratic optimization problem for optimal classification is as follows (Vapnik 1995).

$$\min \|w\|^2 + C\sum_{i=1}^{n}\tau_i \quad (1)$$

The constraint is: $y_i(w \cdot x_i - b) \geq 1 - \tau_i$.

In Formula (1), C is called penalty coefficient, which has great influence on the result and is an important parameter. τ_i is the relaxation parameter; $w=\sum_{i=1}^{n}\alpha_i y_i x_i$; $1 \leq i \leq n$. $||w||$ is the norm of the vector w, which is a measure of the length of the vector.

Its Lagrange function is:

$$L(w,b,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\tau_i - \sum_{i=1}^{n}a_i[[y_i[\varphi(x_i)\cdot w] + b] - 1 + \tau_i] - \sum_{i=1}^{n}\beta_i\tau_i$$

Its dual problem is expressed by

$$\max W(a) = \sum_{i=1}^{n}a_i - \frac{1}{2}\sum_{i,j=1}^{n}y_i y_j a_i a_j[\varphi(x_i)\times\varphi(x_j)] = \sum_{i=1}^{n}a_i - \frac{1}{2}\sum_{i,j=1}^{n}y_i y_j a_i a_j K(x_i,x_j) \quad (2)$$

The constraints are: $1 \leq a_i \leq C$, $\sum_{i=1}^{n} y_i a_i = 0$.

In Formula (2), $K(x_i,x_j) = \varphi(x_i)\ \varphi(x_j)$, which is the kernel function. The key of SVM is the kernel function. It is difficult to divide low dimensional space vector sets. The solution is to project them to high dimensional spaces. But the difficulty of this approach is the increase in computational complexity, and the kernel function neatly solves this problem. That is to say, so long as the appropriate kernel function is chosen, the classification function of high-dimensional spaces can be obtained. The radial basis function is dominated by

$$K(x_i, x_j) = \exp(-\gamma\ \|x_i - x_j\|^2) \quad (3)$$

In Formula (3), if the parameter $\gamma = 1/(2\sigma^2)$ is too large or too small, the performances of SVM will be reduced. So choosing the appropriate value requires a trade-off.

The final decision function is expressed by

$$f(x,a) = \text{sign}\left[\sum_{i,j=1}^{n}y_i a_i K(x_i,x_j) + b\right] \quad (4)$$

According to the regression theory of SVM, the penalty coefficient C and the γ value of radial basis function have great influence on the result, so it is necessary to select the appropriate C and γ to obtain the superior performance of SVM. The traditional empirical method is by trial and error, so it is subjective. In this paper, the exact C and γ can be obtained by the brute force search algorithm.

Because each species has its own particular relatively stable niche, one can predict the potential geographic distribution of tobacco blue mold in China. In the present study based on ecological niche of the known distribution sites of tobacco blue mold were selected. There are 56 known distribution sites of tobacco blue mold (Fig. 1). The ecological niche is closely related with the environmental factors. The environmental data are mainly composed of three factors, i.e., precipitation, temperature and topography. Twenty variables, which specifically include 19 bioclimatic variables from the WorldClim and a global Digital Elevation Model (DEM) were selected. The 19 bioclimatic variables include annual mean temperature, monthly mean temperature difference between day and night, ratio of diurnal temperature difference to annual temperature difference, temperature variation variance, maximum temperature of the warmest month, minimum temperature of the coldest month, range of annual temperature variation, mean temperature of the wettest season, mean temperature of the driest temperature, mean temperature of the hottest season, mean temperature of the coldest season, annual mean humidity, humidity of the wettest month, humidity of the driest month, variance of humidity, humidity of the wettest season, humidity of the driest season, mean humidity of the warmest season and mean humidity of the coldest season. The spatial resolution of the environmental data set is 2.5 min.

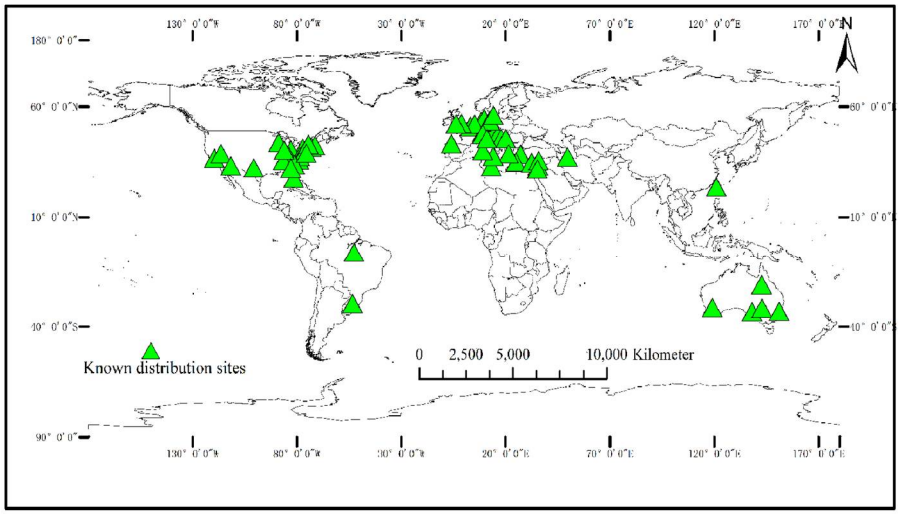

Fig. 1. Global geographic distribution of tobacco blue mold.

Methods of predicting tobacco blue mold's suitable areas in China based on SVM are given below: Before training, the original data are normalized within [0, 1), the aim was to transform data of different dimensions and orders of different magnitudes into comparable data of non-dimensions and orders of same magnitude that can be used to perform mathematical operations on each other. The commonly used methods include linear function, logarithmic function and so on. In the present study, the linear function is used, namely:

$$y(k)=(x(k) - \min(x(n)))/(\max(x(n)) - \min(x(n))),\ k=1,2,\ldots, N. \quad (5)$$

In Formula (5), the minimum and the maximum are represented by the 'min' and 'max', respectively.

The SVM parameters to be optimized are C and γ. According to the regression theory of SVM, the penalty coefficient C and γ value of radial basis function have great influence on the result. Therefore, to obtain the superior performance of SVM, it is necessary to select appropriate C and γ. The traditional trial and error method is subjective. In the present study, the training samples are used to train various types of SVM classifiers. In the training, the brute force search algorithm was used to determine the C and γ parameters of SVM; and then the test samples were used to verify the trained classifiers. According to the classification accuracy and the computation time, the suitable model was selected. Finally, the SVM classifier was used to classify the whole test area.

The ecological niche model was used to train and verify the training set to produce a model suitable for the classification problem, which was then stored in the model file. The model file mainly recorded the weight of each environment factor in the model.

By means of the secondary development technology of GIS, for any grid point P_i in the target areas, every environmental factor v_i was obtained corresponding to the point P_i. The value of environmental factors was summed according to the weighted method: $V = \sum_{i=1}^{m} w_i v_i$, and the value V was obtained. The possibility of invasive species was judged by comparing with the threshold value and the existence and non-existence were represented by 1 and 0, respectively.

For generation of the spatial distribution map of environmental factors, prediction was made by traversing all grid points (evenly spacing grids) in the whole target areas. The environmental factor Ef' was used to replace the elevation value of each grid point in DEM and then the potential spatial distribution map of invasive species in China was obtained by GIS technology.

For performing cross validation, Leave-One-Out cross validation (LOO-CV) method was used. Suppose N samples, each of which serves as an independent validation set and the remaining N-1 samples as the training set, the last one as test set, so N models are obtained by LOO-CV method. The training set and the test set are independent. The average classification accuracy of the final validation set of these N models was used as the performance criteria of the LOO-CV classifier.

Results and Discussion

Based on the above method, the potential geographical distribution of tobacco blue mold was China was obtained, as shown in Fig. 2.

Results showed that the potential geographic distribution of tobacco blue mold is mainly in the entire south, east and middle China most of southwest China except Tibet, north China and Taiwan, and a few areas in northwest and northeast China except Shanxi province. The South China Sea was not included for the lack of data. As can be seen from Fig. 2, tobacco blue mold is a potential threat to most tobacco producing areas in China. It is important to take effective quarantine and control measures to prevent the invasion and spread of tobacco blue mold.

The appropriate values for C and γ play an important role on preventing SVM from deterioration in the generalization performance (Sadeghia *et al*. 2012). In the present experiment, empirical method was adopted to predict the fitness of known sample points with $C = 0.02$ and $\gamma = 0.01$, respectively. SVM was used to predict the suitable areas of known sample points. Results showed that three out of the 56 sample points (True) were False with 94.6% accuracy. The parameters of SVM were then optimized and a group of optimal combination was selected by the brute force search algorithm from 2500 combinations of C and γ, which was: $C = 0.019$, $\gamma = 0.002$. Only one of the 56 sample points (True) had a false prediction value and the accuracy was 98.2%.

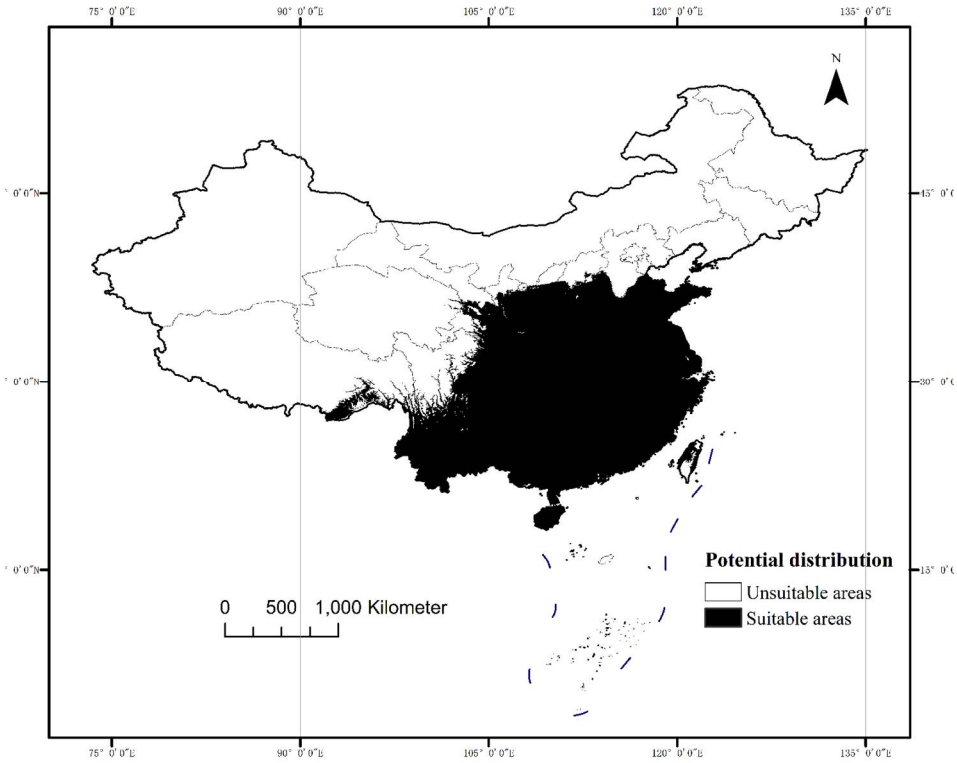

Fig. 2. Potential distribution areas of tobacco blue mold in China.

The LOO-CV classifier is used for cross validation. Results showed that the prediction accuracy of empirical method was 83.9% and the prediction accuracy of SVM parameters optimized by the brute force search algorithm was 91.1%. In the LOO-CV cross validation, each training sample does not participate in the prediction, so the evaluation results are reliable; in addition, there was no random factor in the experimental process that would affect the experimental data, ensuring that the experimental process can be replicated. However, because the number of models needed to be built was the same as the number of the original data samples, the calculation cost of this method will be higher when the number of original data samples was quite large. For the number of samples studied was small. So this method is effective.

The present study showed that the global geographical distribution of tobacco blue mold has not yet reached its maximum potential geographical distribution, so tobacco blue mold has a high risk of invasion in many areas in China. Once tobacco blue mold enters into these areas, it would be a huge loss to the local tobacco industries. As a consequence, it is very necessary to strengthen the quarantine of tobacco blue mold.

Acknowledgments

This work was financially supported by the National Program on Key Research Projects of China (no. 2017YFC1502706).

References

Avolio M and Fuduli A 2021. A semiproximal support vector machine approach for binary multiple instance learning. IEEE T. Neur. Net. Lear. **32**(8): 3566-3577.

Baral P and Haq MA 2020. Spatial prediction of permafrost occurrence in Sikkim Himalayas using logistic regression, random forests, support vector machines and neural networks. Geomorphol. **371**: 107331.

Blanco-Meneses M, Carbone I and Ristaino JB 2018. Population structure and migration of the tobacco blue mold pathogen, *Peronospora tabacina*, into North America and Europe. Mol. Ecol. **27**: 737-751.

Booth T 2018. Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. Austral. Ecol. **43**(8): 852-860.

Byeon DH, Jung JM, Jung S and Lee WH 2020. Effect of types of meteorological data on species distribution predicted by the CLIMEX model using an example of *Lycorma delicatula* (Hemiptera: Fulgoridae). J. Asia-Pacific Biodiv. **13**(1): 1-6.

Gevrey M, Worner S, Kasabovb NK and Pitt J 2006., Estimating risk of events using SOM models: A case study on invasive species establishment. Ecol. Model. **197**(3-4), 361-372.

Makmuang D, Wangkeeree R, Nattee C and Khamsemanan N 2020. A novel twin parametric support vector machine for large scale problem. Thai J. Math. **18**(4): 2107-2127.

Oliveira DVD, Rode R, Neto RRDO, Gama JRV and Leite HG 2021. Use of artificial neural networks for predicting volume of forest species in the Amazon Forest. Sci. For. **49**(131): e3610.

Phillips SJ, Anderson RP and Schapire RE 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. **190**(3-4): 231 - 259.

Ray D, Behera MD and Jacob J 2018. Evaluating ecological niche models: a comparison between Maxent and GARP for predicting distribution of *Hevea brasiliensis* in India. Proceedings of the Indian Nation. Sci. Acad. Part B Biol. Sci. **88**(4): 1337-1343.

Sadeghia S, Zarkamib R and Sabetraftarb K 2012. Use of support vector machines (SVMs) to predict distribution of an invasive water fern *Azolla filiculoides* (Lam.) in Anzali wetland, southern Caspian Sea, Iran. Ecol. Model. **244**: 117 - 126.

Saha A, Rahman S and Alam S 2021. Modeling current and future potential distributions of desert locust *Schistocerca gregaria* (Forskal) under climate change scenarios using MaxEnt. J. Asia-Pacific Biodiv. **14**(3): 399-409.

Shi J and Shi JK 1996. Studies on tobacco blue mold. Chinese Tobacco Sci. **2**: 9-15.

Sindel BM and Michael PW 1992. Spread and potential distribution of *Senecio madagascariensis* Poir. (Fireweed) in Australia. Australian J. Ecol. **17**: 21-26.

Stockwell D and Peters D 1999.The GARP modeling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Syst. **13**(2): 143-158.

Stoeckli S, Felber R and Haye T 2020. Current distribution and voltinism of the brown marmorated stink bug, *Halyomorpha halys*, in Switzerland and its response to climate change using a high-resolution CLIMEX model. Int. J. Biometeorol. **64**(12): 2019-2032.

Vapnik VN 1995. The nature of statistical learning theory. Berlin: Springer. pp. 988-999.

Wan J, Qi GJ, Ma J, Ren YL, Wang R and Mckirdy S 2020. Predicting the potential geographic distribution of *Bactrocera bryoniae* and *Bactrocera neohumeralis* (Diptera: Tephritidae) in China using MaxEnt ecological niche modeling. J. Integr. Agr. **19**(8): 2072-2082.

Wiens JJ and Graham CH 2005. Niche conservatism: Integrating evolution, ecology, and conservation biology. Ann. Rev. Ecol. Syst. **36**: 519-539.

Zhang Z 1995. Tobacco blue mold. Acta Tabacaria Sinica **2**(3): 1-10.

(Manuscript received on 10 January, 2022; revised on 7 October, 2022)