Journal of
Plant Ecology

VOLUME 1, NUMBER 2,
PAGES 143–145

JUNE 2008

doi: 10.1093/jpe/rtn005

Advanced Access published
on 15 April 2008

available online at
www.jpe.oxfordjournals.org

GeoSVM: an efficient and effective tool to predict species' potential distributions

Wenyun Zuo[1], *Ni Lao*[2], *Yuying Geng*[1] *and Keping Ma*[1,*]

[1] *State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China*

[2] *School of Software, Tsinghua University, Beijing 100084, China*

*Correspondence address. State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. E-mail: kpma@ibcas.ac.cn

Patterns of species distribution have long been one of the important topics of ecological study (Brown and Lomonilo 1998). In this brief communication, we introduce a new program—GeoSVM—that uses support vector machine (SVM) to predict species' potential distributions. (GeoSVM is now available at http://www.unm.edu/~wyzuo/GEO.htm.) Here, we also give the results of our evaluation of the performance of GeoSVM. We used data for 30 species of Rhododendron *in China as a case study to compare GeoSVM and Genetic Algorithm for Rule-Set Prediction (GARP), one of the most popular models to predict species' potential distributions. We found that GeoSVM is more accurate and efficient than GARP. Furthermore, GeoSVM can handle more environmental information, which significantly improves the prediction accuracy.*

Patterns of species distribution can potentially answer a bunch of fundamental questions in ecology, such as where are the original habitats of the species; how do the species distribute on earth; how do species achieve their distribution patterns; what is the relationship between distribution patterns of different species and how to set up a policy to conserve endangered species. The development of computer technology and machine learning methods enables the use of environmental factors to simulate species' potential distribution.

Various statistical models have been explored in previous works for predicting species distributions, e.g. generalized linear models, generalized additive models, logistic regression, neural networks, decision trees, principle components analysis (PCA), Mahalanobis distance, maximum entropy method, genetic algorithm and regression tree analysis (see a survey in Zuo *et al.* 2007). These statistical models have been commonly used in wide range of other applications. However, when applied to the prediction of potential species distributions, a common problem arises—the high dimensionality and small sample size problem. This problem is caused by the nature of the task—the prediction of potential species distributions generally depends on the specimen data. These data are accumulated by fieldwork. Fieldwork, being an expensive and difficult process, limits the quantity of data available. We have >400 species of *Rhododendron* in China, but only 161 of them have >20 location samples (the lower limit of sample size for GARP). On the other hand, there are >100 environmental factors that can potentially affect species distribution, such as meteorological factors like annual, monthly, maximum and minimum values of temperature, precipitation and relative humidity as well as geographical factors like altitude and slope and soil and vegetation type. Most statistical methods rely on the big sample assumption that 'the number of samples is much larger than the number of parameters'. As we can see, however, this assumption does not hold anymore for species distribution data. Under this situation, these models usually perform well on training samples, but badly on new testing data. This phenomenon is called 'over training'. Some dimension-reducing methods, such as PCA, can mitigate this problem but only to some extent.

SVM is a model for classification and regression based on statistical learning theory created by Vapnik (1995) at AT&T Bell Labs. It is based on structural risk minimization principle, an improvement over the traditional empirical risk minimization principle. Because of its outstanding empirical performance, SVM has been well accepted by many scientific communities (Gunn 1998). We implemented a potential species distribution predicting system, called GeoSVM, based on SVM. Detailed system architecture of GeoSVM is described in Zuo *et al.* (2007). First, GeoSVM randomly generates negative sample points that are five times the number of positive ones. GeoSVM assumes that the species do not exist at negative sample points. Weight 1/5 is given to each negative sample and Weight 1 is given to each positive sample. Environmental features are extracted from the environmental digital map based on the training samples' locations. These environmental

features and labels for each point (positive or negative) comprise the training data. Then, an SVM model is trained using the training data. Finally, GeoSVM iterates through every single grid point on the map and uses SVM to predict whether the species exist on the point based on environmental features extracted at this point. We used the open source library LibSVM (Lin 2006) developed by Chang and Lin (2006).

To test the performance of GeoSVM, we chose *Rhododendron* L., a genera rich in China. It has ~970 natural species (subspecies not included) around the world. It is one of the representative genera of Hengduan Mountain and east Himalayas—two of the most important hot spots for biodiversity in the world. We used 30 species of *Rhododendron* L. (29 of them are Chinese endemic species) as objects of study to compare the performance of GeoSVM with GARP. The predictions for these species were evaluated by both expert evaluation scores and statistical metrics. The sample size of all these species was >20—the minimum sample size required by GARP (Stockwell and Peters 1999). Specimen data came from seven major herbaria in China (Zuo *et al.* 2007). Our environmental data included 11 layers used in the paper by Zuo *et al.* (2007) and 72 more meteorological layers (Yu *et al.* 2004): monthly average temperature, monthly maximum temperature, monthly minimum temperature, monthly average precipitation, monthly average relative humidity and monthly average total radiation. These 83 environmental variable layers were on 1 × 1 km grid maps. GARP cannot hold >11 environmental layers at such fine resolution. Thus, we compared GARP and GeoSVM by using only 11 layers as in the paper by Zuo *et al.* (2007). We also tested whether more environmental information can result in better prediction by using all 83 layers in GeoSVM.

For each species, the predictions were blindly scored by expert into five grades. The scores were mostly based on past research experience, individual ecology, known distribution areas, habitats, flora and climate envelopes. Furthermore, we used Receiver Operator Characteristic (ROC) curve as a statistical test for the prediction. ROC is one of the most common statistical methods to evaluate the performance of classification model (Mozer *et al.* 2002). Area under the curve (AUC), the area under ROC curve, is the metric used for comparison, which is positively related to the performance of the classification model.

The predictions of 30 species' potential distribution from GARP and GeoSVM were significantly different. Moreover, using 11 layers and 83 layers in GeoSVM also showed significant difference. The one-tailed *t*-test in R showed that average expert scores (2.45 ± 0.23) of predictions by GeoSVM with 11 layers were significantly ($P < 0.0001$) higher (mean of difference was 2.12) than those by GARP (0.33 ± 0.07) and significantly ($P < 0.0001$) lower (mean of difference was −1.92) than those by GeoSVM with 83 layers (4.37 ± 0.11) (Table 1). We drew ROC for each prediction and calculated their AUC values. Results showed that AUC of predictions by GeoSVM with 11 layers were larger than that by GARP for all 30 species and smaller than those by GeoSVM with 83 for 29 species (Fig. 1).

Table 1 expert scores of the predicted potential distribution maps of 30 species of *Rhododendron* in China

Latin name	GARP	$GeoSVM_{11}$	$GeoSVM_{83}$
Rhododendron aganniphum Balf. f et K. Ward	0	3	5
Rhododendron argyrophyllum Franch.	1	3	4
Rhododendron augustinii Hemsl.	1	4	5
Rhododendron brachyanthum Franch.	0.5	3	5
Rhododendron calophytum Franch.	0.5	3	3.5
Rhododendron davidii Franch.	0.5	3	4
Rhododendron decorum Franch.	0	2	5
Rhododendron delavayi Franch.	1	3.5	4.5
Rhododendron dendrocharis Franch.	0.5	2.5	3.5
Rhododendron fulvum Balf. f. et W. W. Smith	1.5	1.5	4.5
Rhododendron haematodes Franch.	0.5	0.5	4
Rhododendron heliolepis Franch.	0	2	5
Rhododendron irroratum Franch.	0.5	2.5	5
Rhododendron lutescens Franch.	0	3	4
Rhododendron mariesii Hemsl. et Wils.	0.5	4	5
Rhododendron nivale Hook. f.	0	3.5	5
Rhododendron ovatum (Lindl.) Planch. ex Maxim.	0	5	5
Rhododendron phaeochrysum Balf. f. et W. W. Smith	0	3.5	5
Rhododendron protistum Balf. f. et Forrest	0	1.5	4.5
Rhododendron racemosum Franch.	0.5	2.5	5
Rhododendron rex Lévl.	0.5	0.5	4.5
Rhododendron saluenense Franch.	0.5	0.5	3.5
Rhododendron sanguineum Franch.	0	1.5	3.5
Rhododendron selense Franch.	0	0	4
Rhododendron simsii Planch.	0	5	5
Rhododendron spinuliferum Franch.	0.5	2.5	3.5
Rhododendron strigillosum Franch.	0	1.5	3.5
Rhododendron uvariifolium Diels	0	2.5	4.5
Rhododendron wardii W. W. Smith	0	1.5	4
Rhododendron stewartianum Diels[a]	0	1.5	3.5

[a] This species is not an endemic species of *Rhododendron* in China.
Average expert scores (±standard error, SE) of GARP prediction is 0.33 ± 0.07. Average expert score (±SE) of $GeoSVM_{11}$ prediction is 2.45 ± 0.23. Average exports score (±SE) of $GeoSVM_{83}$ prediction is 4.37 ± 0.11. The one-tailed *t*-test shows that the difference between GARP and $GeoSVM_{11}$ (mean of difference is −2.12) and the difference between $GeoSVM_{11}$ and $GeoSVM_{83}$ (mean of difference is −1.92) are significant (P value < 0.0001).

Running time of a program is an important evaluation criterion. We ran both GARP and GeoSVM on a DELL Precision 470 workstation. For our data, the experiment took GARP ~35 hr per species, compared with 2.5 hr per species for GeoSVM. For both methods, most of the running time was spent doing prediction in order to produce a distribution map. If we

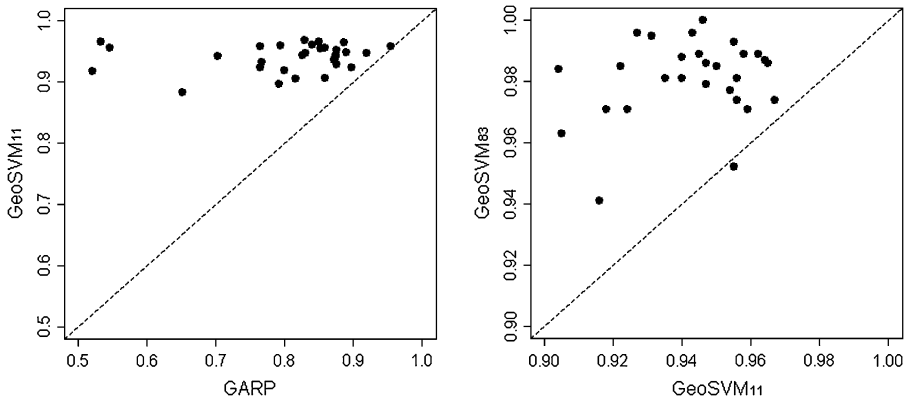

Figure 1 comparison of AUC for GARP and $GeoSVM_{11}$, and $GeoSVM_{11}$ and $GeoSVM_{83}$. Average exports score (±standard error, SE) of GARP prediction is 0.804 ± 0.0202. Average exports score (±SE) of $GeoSVM_{11}$ prediction is 0.939 ± 0.00407. Average exports score (±SE) of $GeoSVM_{83}$ prediction is 0.977 ± 0.00357. The one-tailed *t*-test shows that the mean of the difference between GARP and $GeoSVM_{11}$ is −0.137 (*P* value < 0.0001) and the mean of difference between $GeoSVM_{11}$ and $GeoSVM_{83}$ is −0.0385 (*P* value < 0.0001).

subtract the time used to draw the distribution map for both models, the training time of GARP was ~9.5 hr per species and that of GeoSVM was <1 s per species. The huge difference of training time between GARP and GeoSVM was possibly due to the characteristics of Genetic Algorithms (GA). GA is based on random search processes to find the convergent points, which involves huge calculations. Furthermore, the time complexity increases exponentially for GA as the number of environmental features increases. On the contrary, the time complexity of SVM generally increases with the number of environmental features. Therefore, SVM opens a gate for us to predict efficiently the species distribution with a large number of environmental features.

So far, the reason that many predicting models coexist is that none of them can work well for all situations, and GARP is by far one of the most popular models to predict species' potential distribution. In this study, however, we found that GeoSVM performs much better than GARP in terms of both efficiency (14 times faster on total running time and 30,000 times faster on training time) and effectiveness of analysis for our data. We also found that using more environmental information can significantly improve the prediction accuracy. In conclusion, SVM opens for us a gate for using very high dimensional data with very small sample size. To the best of our knowledge, there are very few studies in which SVM was applied to predict species' potential distribution (Guo *et al.* 2005). Our study demonstrated better performance of prediction in GeoSVM for our data. Further studies are needed to examine whether SVM is suitable for other types of data or situations.

References

Brown JH, Lomolino MV (1998) *Biogeography*, 2nd edn. Sunderland, MA: Sinauer Associates.

Chang CC, Lin CJ (2006) *LIBSVM: A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf (15 January 2006, date last accessed).

Gunn SR (1998) *Support Vector Machines for Classification and Regression*. http://www.ecs.soton.ac.uk/~srg/publications (21 July 2006, date last accessed).

Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecol Model* **182**:75–90.

Lin CJ (2006) *LIBSVM*. http://www.csie.ntu.edu.tw/~cjlin (21 July 2006, date last accessed).

Mozer MC, Dodier R, Colagrosso MD, et al (2002) Prodding the ROC curve: constrained optimization of classifier performance. In: Dietterich T, Becker S, Ghahramani Z (eds). *Advances in Neural Information Processing Systems XIV*. Cambridge, MA: MIT Press, 1409–15.

Stockwell D, Peters D (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *Int J Geogr Inf Sci* **13**:143–58.

Vapnik VN (1995) *The Nature of Statistical Learning Theory*. Berlin: Springer.

Yu GR, He HL, Liu XA, et al (2004) *Atlas for Spatialized Information of Terrestrial Ecosystem in China—Volume of Climatological Elements*. Beijing, China: China Meteorological Press (in Chinese).

Zuo W, Lao N, Geng Y, et al (2007) Prediction species' potential distribution—SVM compared with GARP. *J Plant Ecol* **31**: 711–19.