

预测物种潜在分布区——比较 SVM 与 GARP

左闻韵^{1 3} 劳 逆² 耿玉英¹ 马克平^{1*}

(1 中国科学院植物研究所植被与环境变化国家重点实验室 北京 100093)

(2 清华大学软件学院 北京 100084) (3 中国科学院研究生院 北京 100049)

摘 要 物种分布与环境因子之间存在着紧密的联系,因此利用环境因子作为预测物种分布模型的变量是当前最普遍的建模思路,但是绝大多数物种分布预测模型都遇到了难以解决的“高维小样本”问题。该研究通过理论和实践证明,基于结构风险最小化原理的支持向量机(Support vector machine, SVM)算法非常适合“高维小样本”的分类问题。以20种杜鹃花属(*Rhododendron*)中国特有种为检验对象,利用标本数据和11个1 km × 1 km的栅格环境数据层作为模型变量,预测其在中国的潜在分布区,并通过全面的模型评估——专家评估,受试者工作特征(Receiver operator characteristic, ROC)曲线和曲线下方面积(Area under the curve, AUC)——来比较模型的性能。我们实现了以SVM为核心的物种分布预测系统,并且通过试验证明其无论在计算速度还是预测效果上都远远优于当前广泛使用的规则集合预测的遗传算法(Algorithm for rule-set prediction, GARP)预测系统。

关键词 物种分布预测模型 支持向量机 GARP ROC曲线 杜鹃花属 潜在分布区

PREDICTING SPECIES' POTENTIAL DISTRIBUTION—SVM COMPARED WITH GARP

ZUO Wen-Yun^{1 3}, LAO Ni², GENG Yu-Ying¹, and MA Ke-Ping^{1*}

¹Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China, ²School of Software, Tsinghua University, Beijing 100084, China, and ³Graduate University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Aims The most common method to build a predictive model of species' potential distribution is to use environmental factors, because they strongly affect species distribution. Unfortunately, most predictive models suffer from the “high dimension small sample size” problem, and cannot give satisfactory results in many cases. Support vector machine (SVM), which is based on structural risk minimization principle, has proven to be especially suitable for such data by both theory and abundant applications. Our objective was to implement a new predictive system of species' potential distribution based on the SVM method.

Methods We performed a country-scale case study using 20 Chinese endemic species of *Rhododendron*, employing herbarium specimen data and 11 layers of 1 km × 1 km digital environmental grid data. Through expert evaluation and receiver operator characteristic (ROC) curve, we compared SVM predictions with those of a commonly used modeling method, the genetic algorithm for rule-set prediction (GARP).

Important findings All scores of SVM's prediction are higher than GARP's in expert evaluation. For the statistical analysis of ROC curve, almost all the area under the curve (AUC) determinations of SVM are larger than that of GARP. Furthermore, SVM's prediction speed is much faster than GARP's. Through our experiment, comprehensive evaluation proved that SVM is much better than GARP in terms of both performance and accuracy on the “high dimension small sample size” problem.

Key words predictive model of species distribution, support vector machine (SVM), genetic algorithm for rule-set prediction (GARP), receiver operator characteristic (ROC) curve, *Rhododendron*, potential distribution

生物多样性的空间分布格局一直是生物地理学研究的主要问题(Brown & Lomolino, 1998)。学者希

望通过对生物多样性空间分布格局的研究,回答哪里是物种起源地?哪里是物种聚集中心?物种是按

收稿日期:2006-05-15 接受日期:2006-07-31

基金项目:国家科技基础条件平台工作项目(0246019B)

感谢中国科学院植物研究所喻梅副研究员在GIS系统方面给予的帮助以及工作上给予的大力支持

* 通讯作者 Author for correspondence E-mail: makp@brim.ac.cn

E-mail of the first author: margaretzy@ibcas.ac.cn

照什么规律分布在地球表面?在历史变迁的过程中物种是怎么实现目前的分布格局?不同物种的不同分布格局之间有什么关系?怎样制定物种多样性的保护策略?然而为了回答这些问题,我们首先要能清晰准确地了解物种的空间分布格局。过去,学者主要通过普查或者根据经验分析动、植物区系结构,在地图上粗略的勾画出物种大致的分布格局(方瑞征和闵天禄,1995)。但是随着现代计算机科学的进步,运算能力的增强,利用物种分布与环境因子之间存在的紧密联系(Brown & Lomolino, 1998),以环境因子作为预测模型变量的物种潜在分布区模型层出不穷。最早的有 CLIMEX (Sutherst & Maywald, 1985),基于生物气候数据的有 BIOCLIM (Busby, 1986; Nix, 1986; Peterson *et al.*, 1999; Farber & Kadmon, 2003; Tellez-Valdes & DaVila-Aranda, 2003)和 DOMAIN (Carpenter *et al.*, 1993; Manrique *et al.*, 2003),基于生态位的有 GARP (Genetic algorithm for rule-set prediction) (Stockwell *et al.*, 2006; Stockwell & Noble, 1992; Stockwell, 1997; Peterson *et al.*, 1999; Stockwell & Peters, 1999; Stockwell & Peterson, 2002, 2003; Peterson & Robins, 2003) ENFA (Environmental-niche factor analysis) (Hirzel *et al.*, 2001) 和 WhereWhy (Stockwell, 2006),以及基于特有种分布特点的特有性简约分析 (Parsimony analysis of endemism, PAE) (Manrique *et al.*, 2003)。

不同模型中运用的统计方法大相径庭。线性回归模型 (Generalized linear models, GLMs) (Hirzel *et al.*, 2001), 广义相加模型 (Generalized additive models, GAMs) Logistic 回归 (Logistic regression) (Manel *et al.*, 1999; Ozesmi & Ozesmi, 1999; Bolliger *et al.*, 2000; Osborne *et al.*, 2001), 神经网络 (Neural networks) (Manel *et al.*, 1999; Ozesmi & Ozesmi, 1999), 决策树 (Decision trees) (Stockwell & Noble, 1992), 主成分分析 (Principle components analysis, PCA) (Robertson *et al.*, 2001), 马氏距离 (Mahalanobis distance) (Farber & Kadmon, 2003), 最大熵法 (Maximum entropy method) (Phillips *et al.*, 2006), 遗传算法 (Genetic algorithm) (Stockwell *et al.*, 2006; Stockwell & Noble, 1992; Stockwell, 1997; Peterson *et al.*, 1999; Stockwell & Peters, 1999; Stockwell & Peterson, 2002, 2003; Peterson & Robins, 2003) 和回归树分析 (Regression tree analysis) (Iverson & Prasad, 1998) 等都是过去研究工作中运用过的数学方法。

以上这些方法虽然较为广泛的应用于各个科学

领域,但是在物种分布预测上却遇到了一个共同的困难——高维小样本数据。这是由物种分布预测的特性造成的。物种分布预测所基于的样本通常是由科学家实地考察采样获得,数量非常有限。然而潜在影响物种分布的环境因子却非常多,包括降水、温度、相对湿度、光照强度的年、月均值和年、月极值、地形上的海拔、坡度、坡向、土壤类型、植被类型等等特征,可达近百种。一般的统计方法都是基于大样本的假设,即样本数远大于模型参数的个数。然而以上方法参数个数一般与特征维数成一次或二次方关系,所以对于高维小样本数据这些方法所基于的假设不再成立。其后果表现为,虽然在训练样本上这些模型可以取得很高的分类正确率,但是在新的数据上表现却很差。在统计上这种现象称为过拟合。一些降维的方法例如 PCA 也只能在一定程度上缓解这个问题。

支持向量机 (Support vector machine, SVM) (Vapnik, 1995) 是针对分类和回归问题提出的统计学习理论。它基于结构风险最小化原理 (Structural risk minimization principle), 使得模型对期望数据做出最佳的界定,达到最佳的分类效果。由于它具有适用于高维小样本问题的独特性质,近几年来被非常广泛的应用于科学技术的各个领域。我们实现了以 SVM 为核心的物种分布预测系统,并且通过试验证明其无论在计算速度还是预测效果上都远远优于当前广泛使用的 GARP 预测系统 (Stockwell *et al.*, 2006; Stockwell & Noble, 1992; Stockwell, 1997; Peterson & Cohoon, 1999; Peterson *et al.*, 1999; Stockwell & Peters, 1999; Peterson, 2001; Peterson *et al.*, 2001, 2002; Stockwell & Peterson, 2002; Anderson *et al.*, 2003; Peterson & Robins, 2003; Stockwell & Peterson, 2003; Anderson & Martinez-Meyer, 2004; Martinez-Meyer *et al.*, 2004)。当然在未来进一步的研究中 SVM 还需要与其它模型进行比较。

1 方 法

1.1 SVM 原理

SVM 是 AT&T Bell 实验室的 Vapnik (1995) 提出的针对分类和回归问题的统计学习理论。SVM 方法基于结构风险最小化原理,明显优于传统的基于经验风险最小化原理 (Empirical risk minimization principle) 的分类方法。由于许多引人注目的特点和出众的实验性能 (Gunn, 1998), SVM 方法越来越受到重视。

SVM 模型主要适用于两类样本的二分类问题。例如在物种分布预测中有物种分布的点即为正例，没有该物种分布的点即为负例。SVM 利用一个超平面在高维特征空间中将近正负两类样本分开。超平面的法向量 w 以及截距 w_0 即为模型参数。决策函数为 $f(x) = w^T x + w_0$ ， $f(x) = 0$ 即为决策面(图 1)。其中 x 为样本点的特征向量。若 $f(x) \geq 0$ 则该样本被模型预测为正例，若 $f(x) < 0$ 则该样本被模型预测为负例。

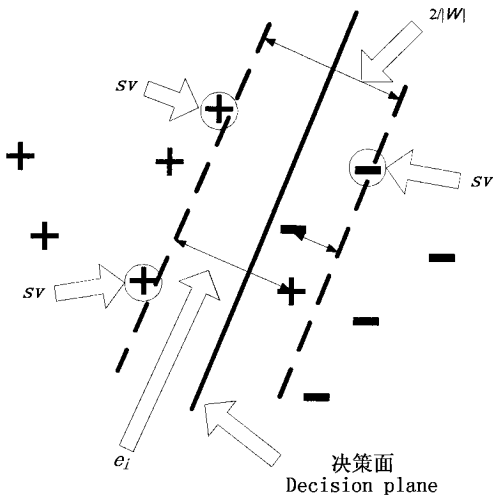


图 1 支持向量机(SVM)原理图

Fig.1 The principle of support vector machine (SVM)
 e_i : 错分样本惩罚 Punishment to wrongly classified samples SV: 支持向量 Support vectors $2/||w||$: 正负样本被分开的间距 Gap between positive and negative samples

中将得到惩罚。 e_i 不等于 0 则说明样本点 i 越过了边界，模型将受到 γe_i 的惩罚。其中 γ 是可以调节的惩罚参数。公式(1)中第二项 $\sum e_i$ 则代表已知训练样本被模型错分的惩罚，体现了模型的经验风险。整个式子代表了 SVM 算法在结构和经验风险间寻求最佳平衡的过程。

解该优化问题的过程可以参见数值优化的理论。基本方法是先将公式(1)转化为对偶问题，然后在对偶空间计算二次规划问题。数值优化的理论可以证明该优化问题达到最优解时必然有一部分样本 x_i 满足： $e_i = 0$ ， $y_i(w^T x + w_0) = 1 - e_i$ 。这些样本即被称为支持向量(Support vectors)。正是这些支持向量决定了决策面参数(w, w_0) (图 1) (Vapnik, 1995)。SVM 模型的参数 w 是一个向量， w_0 是一实数。 w 每一维对应了某个环境因子的权重。 w_0 为分类平面的截距。

1.2 GARP

GARP 是通过遗传算法创建的生态位模型，该模型描述了维持物种种群数量所需的环境条件。它用已知物种分布点数据和带有与物种存活能力相关的环境参数层作为模型输入参数，通过不断迭代的遗传算法实现 4 种规则模型——原子规则(Atomic)、逻辑回归(Logistic regression)、生物气候包络(Bioclimate envelope)和逆生物气候包络(Negated bioclimate envelope)的独立和组合分析，探索物种存在与否与环境参数之间的非随机相关性，并预测和估算物种的潜在分布区(Stockwell & Noble, 1992; Peterson & Cohoon, 1999)。

由于遗传算法的运算原理导致该算法每次的输出结果都存在随机差异，都需要生成最优集合。综合最优集合中的结果，可以得到概率分布的物种潜在分布区(Stockwell & Noble, 1992; Anderson et al., 2003)。参考 Anderson 等(2003)讨论的结果，基于以下参数设置——运行次数 runs = 2 000，收敛值(Convergence limit) = 0.01，最大重复 = 1 000，同时选择所有 4 种规则(不使用其所有组合)：开启最优化规则集合参数选项(Best subset selection parameters)；遗漏(预测但实际存在的百分比)测量(Omission measure)外部(Extrinsic, 针对测试数据的)；遗漏阈值(Omission threshold)：强制(Hard) 0 遗漏(Omission)，在强制遗漏阈值下的模型数(Total models under hard omission threshold) = 100，预测面积比(已知记录以外的预测面积占区域总面积的百分比)阈值(Commission threshold) = 分布中心区的 10%(10% of distribu-

$$\min_w M(\vec{w}) = \frac{1}{2} \|\vec{w}\|^2 + \gamma \sum_{i=1}^N e_i \quad (1)$$

约束条件 $e_i \geq 0, y_i(w^T x_i + w_0) \geq 1 - e_i, i = 1, \dots, N$ 。

w 为超平面的法向量， w_0 为超平面的截距， x_i 为样本点 i 的特征向量， y_i 为样本类别($y_i = +1$ ，若样本点 i 对应正例， $y_i = -1$ ，若样本点 i 对应反例)， e_i 为样本点 i 越过分类缓冲带的距离， γ 为可以调节的惩罚参数， N 为样本个数。

模型(w, w_0)训练过程的核心是求解公式(1)的带约束的优化问题。其中 $y_i = +1$ 若样本点 i 对应正例， $y_i = -1$ 若样本点 i 对应反例。 e_i 是优化过程中的变量，对应着样本点 i 越过分类缓冲带的距离。公式(1)中第一项 $\|w\|^2$ 即为决策面两侧缓冲带宽度的倒数(图 1)。缓冲带越窄， $\|w\|^2$ 越大，即体现了模型对未来数据预测的风险(结构风险)。没有越过该缓冲带的正例或负例则视为被模型正确分类。越过该缓冲带的正例或负例在优化目标函数 $M(w)$

tion)。运行 GARP 1.1.6,在最优规则集合(Best-subset)中得到 10 个模拟结果,将 10 个结果叠加得到一个 10 等级划分的物种潜在存在概率分布图。

GARP 模型是一个庞大复杂的模型,在其官方网站(The University of Kansas Center for Research, 2002)上可以获得详细的用户手册和算法手册。本文由于篇幅问题不能一一讨论 GARP 模型中每个规则的数学公式和理论,如有兴趣可进一步阅读 GARP 模型的算法手册。

1.3 数据

1.3.1 研究对象——中国特有的 20 种杜鹃花属植物

杜鹃花属(*Rhododendron*),约含 970 种(不包括种以下分类等级),是横断山区与东喜马拉雅这两个具有世界意义的生物多样性关键区域的代表属之一,该属的多数种类是构成当地高山、亚高山灌丛生态系统的键种,并为亚高山针叶林、针阔混交林下层优势种或主要伴生种。杜鹃花属的种类分布很广,欧、亚、北美及大洋洲均可见其踪迹,但主要分布在亚洲。东亚与马来西亚的种类最多,占世界总数的 90% 以上,仅特有种就有 850 多种。据最近的资料(Wu *et al.*, 2005)表明,我国共有杜鹃花约 570 种(不包括种以下分类等级),其中特有种约 405 种,占世界种类的 40% 以上,是中国种子植物中最大的属。

本研究选择 20 种中国特有的杜鹃花属植物(表 2)在中国的标本采样点作为存在样本点,对其预测结果进行专家评估和统计分析。这 20 种中国特有的杜鹃花属植物的标本采样点均大于 20 个,这是 GARP 模型的预测输入点最低要求(Stockwell & Peters, 1999; Peterson & Cohoon, 1999; Stockwell & Peterson, 2002)。标本数据来自于我国 7 个主要的标本馆——中国科学院植物园标本馆、中国科学院昆明植物园标本馆、中国科学院华南植物园标本馆、中国科学院武汉植物研究所标本馆、四川大学标本馆、四川林校标本馆、庐山植物园标本馆,以及英国爱丁堡皇家植物园中的部分采集地为中国的杜鹃花属标本。所有标本均经过再次鉴定,去除了具有同一采集号而在不同标本馆被定为不同物种的错误记录。

1.3.2 环境数据

预测系统中所运用到的环境数据包括中国数字气象数据(于贵瑞等,2004)中的年均温、年最高温、年最低温、年均降水、年均相对湿度和年均太阳总辐射,以及由 GARP 官方网站(The University of Kansas

Center for Research, 2002)提供的亚洲数字高程数据中的数字高程、坡度、坡向、降水流量累积量和降水流向。所有 11 个环境变量图层均是 1 km × 1 km 的栅格图。

1.4 SVM 物种分布预测系统流程

编程实现 SVM 物种分布预测系统,系统工作的流程如图 2 所示。首先由系统产生 5 倍于标本点的随机负例点(无某一物种分布的点)与标本点共同组成训练样本点,并给随机负例点赋等于标本点权重的 1/5 的权重。然后从环境变量图层中提取训练样本点对应的环境特征,生成训练数据。SVM 对训练数据进行训练和检验产生适合于该分类问题的模型,并存入模型文件。模型文件主要记录了每个环境因子在模型中的权重。对于一个待预测的栅格,其环境因子按照以上权重加权求和(求 x_i 与 w 的内积)得到的数值就是判别该栅格是否为存在点的依据。根据可以人为设定的阈值与该数值之间的大小关系即可确定该栅格为正例还是负例。SVM 用模型文件中的模型遍历整个研究区域的所有栅格,进行预测。这就是整个 SVM 物种分布预测系统的流程。“评估”是本次实验中专有的部分,是将运算结果放到评估系统中与 GARP 模型的预测结果进行对比。

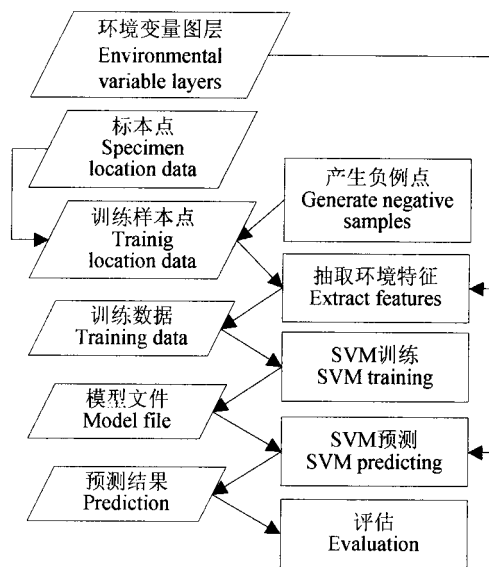


图 2 支持向量机(SVM)物种分布预测系统流程图

Fig. 2 The flow chart of support vector machine(SVM) predictive system

普遍的 SVM 模型已经被很多人实现过,可以在因特网上找到多个版本。SVM 的算法部分即 SVM 训练这个模块我们使用了 Chang 和 Lin(2006)开发

的 LibSVM 开放源码(Lin, 2006)。在此基础上我们开发了 SVM 物种分布预测系统。

1.5 模型评估

1.5.1 专家评估

我们邀请中国杜鹃花属专家对每个物种的两个模型预测结果进行 5 分制盲审。专家根据自己的研究经验, 综合杜鹃花属物种的个体生态学, 已知分布区, 主要气候和生物区的地理分布进行判断。所有预测出来的潜在分布区图, 只有编号, 没有模型名称, 专家按照编号对号打分。

1.5.2 统计方法评估——受试者工作特征(Receiver operator characteristic, ROC)曲线

ROC 曲线是目前常用的评价分类模型优劣的

方法之一(Mozer *et al.*, 2002)。分类模型的优劣可以通过混合矩阵(表 1)的参数计算。如果分类模型能将目标类别(Target class)以及非目标类别(Nontarget class)全部分类正确, 即 $FP = 0, FN = 0$, 则该分类模型是最佳分类模型。因此, TP、FP、TN 和 FN 所表示的敏感性(Sensitivity)与特异性(Specificity)就成了衡量一个分类模型性能的指标。

$$\text{敏感性} = TP / (TP + FN)$$

$$1 - \text{特异性} = 1 - TN / (TN + FP) = FP / (TN + FP)$$

将 $1 - \text{特异性}$ 与敏感性绘制在直角坐标系中即为 ROC 曲线, 根据敏感性与特异性的定义可知 ROC 曲线下方面积(Area under the curve, AUC)的大小与分类器模型的优劣正相关。

表 1 混合矩阵
Table 1 Confusion matrix

	实际真值 Actual positive	实际假值 Actual negative
预测真值 Predicted positive	真真 True positive (TP)	假真 False positive (FP)
预测假值 Predicted negative	假假 False negative (FN)	真假 True negative (TN)

2 实验结果

2.1 预测结果

GARP 和 SVM 两个模型对 20 个物种潜在分布区的预测结果差异明显。由于篇幅所限, 以马银花(*Rhododendron ovatum*)潜在分布图(图 3)为例。如图 3 可见 GARP 所预测的潜在分布区显著大于 SVM 的预测结果。

2.2 专家评估结果

专家对于 20 个物种潜在分布图进行盲审的结果(表 2)通过 ANOVA 分析表明 SVM 预测的 20 个物种平均分(2.5 ± 1.1)显著($p < 0.0001$)高于 GARP 的平均分(0.2 ± 0.5)。

2.3 统计结果

对 GARP 和 SVM 的模型预测结果分别做 ROC 曲线, 以马银花为例(图 4)并计算其 AUC。结果显示 20 个物种中, AUC_{SVM} 均明显大于 AUC_{GARP} (图 5)。

2.4 运算时间对比

模型运算所需的时间(从模型数据输入到数据输出整个过程所消耗的时间)是衡量模型运算效率的一个重要指标。在 DELL Precision 470 DT 工作站——双核 CPU, 两个都为英特尔(R)至强(R)处理

表 2 20 种杜鹃花属中国特有种潜在分布区专家评估结果
Table 2 Expert's scores of the predicted potential distribution maps of 20 endemic species of *Rhododendron* in China

	GARP	SVM
雪山杜鹃 <i>Rhododendron aganniphum</i>	0	3
短花杜鹃 <i>Rhododendron brachyanthum</i>	0.5	3
美容杜鹃 <i>Rhododendron calophyllum</i>	0.5	3
腺果杜鹃 <i>Rhododendron davidii</i>	0.5	3
大白杜鹃 <i>Rhododendron decorum</i>	0	2
树生杜鹃 <i>Rhododendron dendrocharis</i>	0.5	2.5
似血杜鹃 <i>Rhododendron haematodes</i>	0.5	0.5
亮鳞杜鹃 <i>Rhododendron helirolepis</i>	0	2
露珠杜鹃 <i>Rhododendron irroratum</i>	0.5	2.5
黄花杜鹃 <i>Rhododendron lutescens</i>	0	3
雪层杜鹃 <i>Rhododendron nivale</i>	0	3.5
马银花 <i>Rhododendron ovatum</i>	0	5
栎叶杜鹃 <i>Rhododendron phaeochrysum</i>	0	3.5
大树杜鹃 <i>Rhododendron protistum</i> var. <i>giganteum</i>	0	1.5
血红杜鹃 <i>Rhododendron sanguineum</i>	0	1.5
多变杜鹃 <i>Rhododendron selense</i>	0	0
杜鹃 <i>Rhododendron simsii</i>	0	5
芒刺杜鹃 <i>Rhododendron strigillosum</i>	0	1.5
紫玉盘杜鹃 <i>Rhododendron uvarifolium</i>	0	2.5
黄杯杜鹃 <i>Rhododendron wardii</i>	0	1.5

GARP: 规则集合预测的遗传算法 Algorithm for rule-set prediction
SVM: 支持向量机 Support vector machine

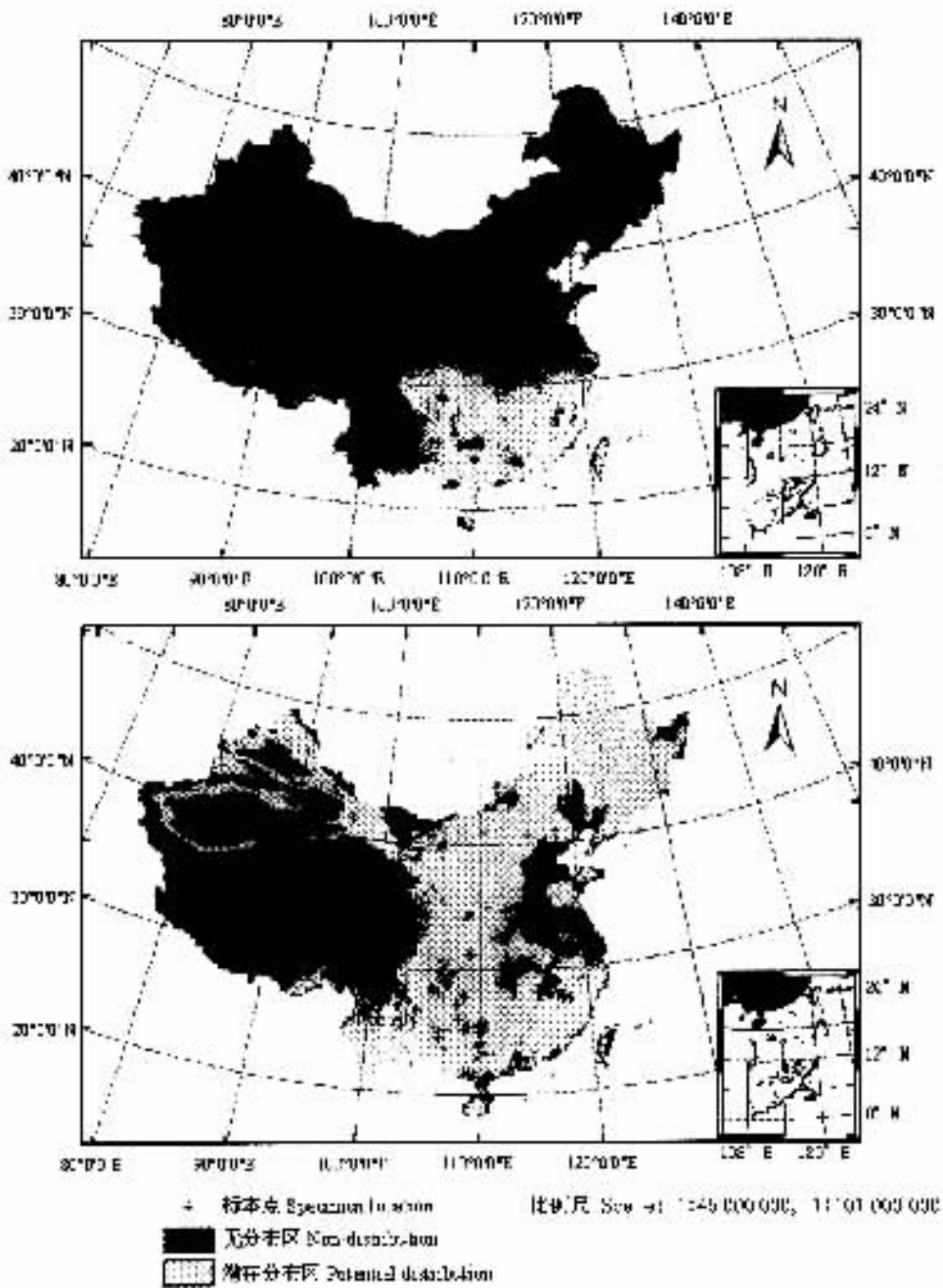


图3 马银花的潜在分布图(上图为 SVM 预测结果,下图为 GARP 预测结果)

Fig.3 The potential distribution maps of *Rhododendron ovatum* (The above is the result of SVM ; the bottom is the result of GARP) SVM, GARP 见表 2 See Table 2

器 3.0 GHz/800 MHz/2 MB, 英特尔(R)EM64T——的运行环境中,该研究数据 GARP 系统的连续运行总时间为每个物种 30~40 h,而 SVM 系统的运行总时间为每个物种 2.5 h。两种方法的系统运行总时间主要花在数据文件的读写上。如果单独比较模型训练的时间(从模型数据输入到训练出适于进行预测的模型所消耗的时间),则 GARP 的模型训练时间为

每个物种 9~10 h,而 SVM 的训练时间每个物种小于 1 s。训练时间的巨大差距主要是由于遗传算法采用的是大量的随机搜索过程,计算量繁重。另外由于搜索复杂度与环境特征数呈指数相关,环境特征数量的增加将显著降低模型的计算率。与之相反 SVM 的计算复杂度与环境特征数呈一次方关系,故可以处理环境特征数非常多的问题。

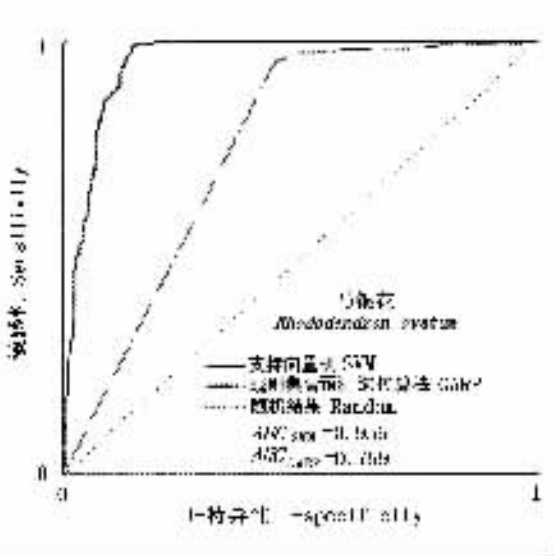


图4 GARP 和 SVM 预测马银花的 ROC 曲线与 AUC
 Fig.4 ROC curve and AUC of *Rhododendron ovatum* predicted with GARP and SVM

Random 是随机预测结果的 ROC 曲线 Random is the ROC curve of stochastic prediction ROC: 受试者工作特征 Receive operator characteristic AUC: 曲线下方面积 Area under the curve SVM、GARP: 见表 2 See Table 2

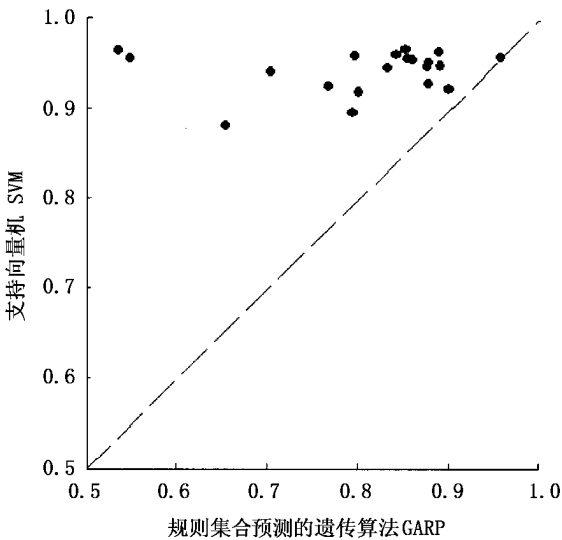


图5 GARP 和 SVM 预测结果的 AUC 对比
 Fig.5 Comparison of AUC for GARP and SVM
 SVM、GARP: 见表 2 See Table 2 AUC: 见图 4 See Fig. 4

3 讨论

对比 GARP 和 SVM 预测的潜在分布图,如马银花(图 3)除了预测的区域存在明显差异,其预测面积也显著不同。虽然有学者认为所谓潜在分布区,是指物种可能的分布区,并不表示物种现在一定在该区域内分布,该区域可能是物种的历史分布区,或者可能是具有维持生存所需的环境因子但物种还未

扩散到的未来分布区(Anderson *et al.*, 2003)。本研究邀请的专家则认为,GARP 对于 20 个所预测的潜在分布区明显超出了这些物种可能分布的范围。而 SVM 对这些物种的预测结果比较合理。SVM 是基于结构风险最小化原理的方法,分类目标是对期望数据做出最佳的界定(即泛化误差的最小化)(Vapnik, 1995),这较 GARP 的对于训练数据和检验数据的误差最小化的目标更加合理,因为我们不能保证训练数据和检验数据就是预测的期望数据。而 GARP 的预测结果明显超出物种的可能分布区,很有可能是由于经验风险最小化原理导致的,因为足够大的面积可以使得“假假值(False negative)” $FN = 0$ 。 $FN/(TP + FN)$ 是否趋近于 0 正是 GARP 判定规则模型合理性的一个参数——外部遗漏值(Extrinsic omission)(Anderson *et al.*, 2003)。

GARP 在当前的物种分布预测领域中已经成为一个比较广泛认可的预测模型,在不同地区,不同数据中也都得到过较好的验证(Stockwell *et al.*, 2006; Stockwell & Noble, 1992; Stockwell, 1997; Peterson & Cohoon, 1999; Peterson *et al.*, 1999, 2001, 2002; Stockwell & Peters, 1999; Peterson, 2001; Stockwell & Peterson, 2002, 2003; Anderson *et al.*, 2003; Peterson & Robins, 2003; Anderson & Martinez-Meyer, 2004; Martinez-Meyer *et al.*, 2004)。但是在我们的数据中,专家评估(表 2)和 ROC 曲线统计结果 AUC(图 5)都一致表明, SVM 物种分布预测系统的结果优于 GARP。当然,由于 SVM 在物种潜在分布区预测上的应用较少(Guo *et al.*, 2005),没有更多的数据支持,我们不能肯定对于其它地域和其它类型的数据, SVM 会有普遍的优势,这是我们需要进一步研究的内容之一。到目前为止,众多预测模型(Busby, 1986; Nix, 1986; Stockwell & Noble, 1992; Carpenter *et al.*, 1993; Stockwell, 1997, 2006; Hirzel *et al.*, 2001)并存主要是因为没有任何一种方法具有普适性,专家评分虽然显示 SVM 物种分布预测系统的预测效果显著强于 GARP(表 2),但是专家认为 SVM 物种分布预测系统并不是对每个物种都有很好的预测,我们将针对这些问题对 SVM 物种分布预测系统继续研究和改进。

不同的模型对应着不同的数据类型和数据量,因为对于不同的算法来说数据类型和数据量的变化可能会导致运算量的猛增,而运算量的增加直接导致模型运算时间的延长,这是模型运行必须考虑的重要因素。目前 GARP 的技术研究和众多应用研究

(Stockwell *et al.*, 2006; Stockwell & Noble, 1992; Stockwell, 1997; Peterson & Cohoon, 1999; Peterson *et al.*, 1999, 2001; Stockwell & Peters, 1999; Peterson, 2001; Stockwell & Peterson, 2002, 2003; Anderson *et al.*, 2003; Peterson & Robins, 2003; Anderson & Martínez-Meyer, 2004; Martínez-Meyer *et al.*, 2004) 中都没有提到模型的运算时间。由于我们的研究区域覆盖整个中国, 数据层精度高达 $1 \text{ km} \times 1 \text{ km}$ 的栅格, 如此庞大的数据量是前人研究所未触及的。因此, 由遗传算法支持的 GARP 遇到了计算时间和内存限制两方面的挑战。由于遗传算法的迭代运算使得运算量随输入数据量的增加呈级数增长, 所以对于我们的数据 GARP 的运算时间远远超出了研究工作可以接受的范围。而且由于遗传算法的运算原理导致该算法每次的输出结果都存在随机差异, 所以必须生成多个模型结果从中选择最优的 10 个模型结果, 将其叠加得到概率分布的物种潜在分布区 (Anderson *et al.*, 2003)。这样就进一步加重了模型的运算负担, 增加了模型的运算时间。而 SVM 算法本身就是适用于高维数据的学习训练, 因此对于本研究的数据来说, SVM 算法的优点得到很好的发挥。

众所周知, 对于模型, 在预测效果一致的基础上, 运算速度快的无疑会更受使用者的青睐。就本研究中的数据来说, SVM 不仅预测效果显著优于 GARP, 而且总的运算效率为 GARP 的 12~16 倍, 算法训练效率更可达 GARP 的 3 万倍。因此, 在未来大数据量物种潜在分布区预测研究中, SVM 算法无疑具有更广泛的应用前景。

参 考 文 献

- Anderson RP, Lew D, Peterson AT (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, 162, 211–232.
- Anderson RP, Martínez-Meyer E (2004). Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biological Conservation*, 116, 167–179.
- Bolliger J, Kienast F, Bugmann H (2000). Comparing models for tree distributions: concept, structures, and behavior. *Ecological Modelling*, 134, 89–102.
- Brown JH, Lomolino MV (1998). *Biogeography* 2nd edn. Sinauer Associates, Sunderland, Massachusetts.
- Busby JR (1986). A biogeographical analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology*, 11, 1–7.
- Carpenter G, Gillison AN, Winter J (1993). DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals. *Biodiversity and Conservation*, 2, 667–680.
- Chang CC, Lin CJ (2006). LIBSVM: a library for support vector machines. Available via DIALOG. <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>. Cited 15 Jan 2006.
- Fang RZ (方瑞征), Min TL (闵天禄) (1995). The floristic study on the genus *Rhododendron*. *Acta Botanica Yunnanica* (云南植物研究), 17, 359–379. (in Chinese with English abstract)
- Farber O, Kadmon R (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, 160, 115–130.
- Gunn SR (1998). Support vector machines for classification and regression. Available via DIALOG. <http://www.ecs.soton.ac.uk/srg/publications>. Cited 21 Jul 2006.
- Guo Q, Kelly M, Graham CH (2005). Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling*, 182, 75–90.
- Hirzel AH, Helfer V, Metral F (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145, 111–121.
- Iverson LR, Prasad AM (1998). Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, 68, 465–485.
- Lin CJ (2006). LIBSVM. <http://www.csie.ntu.edu.tw/~cjlin>. Cited 21 Jul 2006.
- Manel S, Dias JM, Ormerod SJ (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120, 337–347.
- Manrique CE, Duran R, Argaez J (2003). Phylogeographic analysis of taxa endemic to the Yucatán Peninsula using geographic information systems, the domain heuristic method and parsimony analysis of endemism. *Diversity and Distributions*, 9, 313–330.
- Martínez-Meyer E, Peterson AT, Hargrove WW (2004). Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography*, 13, 305–314.
- Mozier MC, Dodier R, Colagrosso MD, Guerra-Salcedo C, Wolniewicz R (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*. MIT Press, Cambridge, MA, 1409–1415.
- Nix HA (1986). A biogeographic analysis of Australian elapid snakes. In: Longmore R ed. *Atlas of Elapid Snakes of Australia*. Australian Government Publishing Service, Canberra, 4–15.
- Osborne PE, Alonso JC, Bryant RG (2001). Modelling landscape-scale habitat use using GIS and remote sensing: a case study with

- great bustards. *Journal of Applied Ecology*, 38, 458 – 471.
- Ozesmi SL, Ozesmi U (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, 116, 15 – 31.
- Peterson AT (2001). Predicting species' geographic distributions based on ecological niche modeling. *Condor*, 103, 599 – 605.
- Peterson AT, Cohoon KP (1999). Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*, 117, 159 – 164.
- Peterson AT, Ortega-Huerta MA, Bartley J, Sanchez-Cordero V, Soberon J, Buddemeier RH, Stockwell DRB (2002). Future projections for Mexican faunas under global climate change scenarios. *Nature*, 416, 626 – 629.
- Peterson AT, Robins CR (2003). Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology*, 17, 1161 – 1165.
- Peterson AT, Sanchez-Cordero V, Soberon J, Bartley J, Buddemeier RW, Navarro-Siguenza AG (2001). Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*, 144, 21 – 30.
- Peterson AT, Soberon J, Sanchez-Cordero V (1999). Conservatism of ecological niches in evolutionary time. *Science*, 285, 1265 – 1267.
- Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231 – 259.
- Robertson MP, Caithness N, Villet MH (2001). A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, 7, 15 – 27.
- Stockwell D, Peters D (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, 13, 143 – 158.
- Stockwell D, Peterson AT (2003). Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. *Ecological Indicators*, 3, 213 – 221.
- Stockwell DRB (1997). Generic predictive systems: an empirical evaluation using the learning base system (LBS). *Expert Systems with Applications*, 12, 301 – 310.
- Stockwell DRB (2006). Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 193, 188 – 196.
- Stockwell DRB, Beach JH, Stewart A, Vorontsov G, Vieglais D, Pereira RS (2006). The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling*, 195, 139 – 145.
- Stockwell DRB, Noble IR (1992). Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, 33, 385 – 390.
- Stockwell DRB, Peterson AT (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148, 1 – 13.
- Sutherst RW, Maywald GF (1985). A computerised system for matching climates in ecology. *Agriculture, Ecosystems & Environment*, 13, 281 – 299.
- Tellez-Valdes O, DaVila-Aranda P (2003). Protected areas and climate change: a case study of the cacti in the Tehuacan-Cuicatlan Biosphere Reserve, Mexico. *Conservation Biology*, 17, 846 – 853.
- The University of Kansas Center for Research (2002). User's manual. <http://www.lifemapper.org/desktopgarp/Default.asp?Item=1&Lang=1>. Cited 21 Jul 2006.
- Vapnik VN (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.
- Wu ZY, Peter HR, Hong DY (2005). *Flora of China* 14. Science Press, Beijing, China.
- Yu GR (于贵瑞), He HL (何洪林), Liu XA (刘新安), Liu D (刘栋), Wang QF (王秋凤), Ren CY (任传友), Li ZQ (李正泉), Su W (苏文), Yue YZ (岳燕珍), Fan LS (范辽生), Guo XB (郭学兵), Zhu QL (祝青林), Cai F (蔡福) (2004). *Atlas for Spatialized Information of Terrestrial Ecosystem in China—Volume of Climatological Elements (中国陆地生态系统空间化信息研究图集——气候要素分卷)*. China Meteorological Press, Beijing. (in Chinese)