基于参数优化支持向量机的林下参净光合速率预测模型

武海巍[1, 2], 于海业[1*], 张 蕾[1]

1. 吉林大学工程仿生教育部重点实验室, 吉林 长春 130022
2. 北华大学电气信息工程学院, 吉林 吉林市 132021

摘 要 使用K-fold交叉验证方法, 通过两种支持向量机函数, 四种核函数, grid-search算法, 遗传算法, 粒子群算法, 建立对个体净光合速率预测拟合程度最高和最佳惩罚参数c的支持向量机模型。将可见光光谱组成成分配比关系归为一个*P*粒子, 将叶温、散射辐射、气温等归为一个ε粒子。通过信息粒子化技术对影响个体净光合速率的因子进行降维处理, 使得分析光合有效辐射、可见光光谱组成成分和个体净光合速率之间的相关关系成为了可能。试验结果表明, epsilon-SVR-RBF-Genetic Algorithm模型, nu-SVR-linear-grid-search模型和nu-SVR-RBF-Genetic Algorithm模型对光合有效辐射和P粒子组成预测集的拟合程度均达到97%以上, nu-SVR-linear-grid-search模型的惩罚参数*c*值最小, 泛化能力最强, 最终采用该模型对光合有效辐射、*P*粒子和ε粒子组成的预测集进行预测分析, 拟合程度达到96%以上。

关键词 可见光光谱; 支持向量机; 参数优化; 信息粒子化
中图分类号: O657.3 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2011)05-1414-05

引 言

在光谱分析中, 人工神经网络是人们常采用的方法, 其中又以BP神经网络应用的最多。BP神经网络要求大样本, 并且容易陷入局部极小值, 在实际应用中其预测的效果并不十分理想。支持向量机(SVM)是近年来机器学习研究的重大成果, 泛化能力强和全局寻优的特点令其在光谱分析中越来越受到重视。SVM利用松弛变量和核函数, 将低维空间的输入变换到一个高维空间, 在高维空间求样本数据线性不可分情况的最优分类面[1-3]。SVM计算和存储数据不受输入维数的限制[4, 5], 这更符合林下参复杂光环境的实际情况。

光作为植物最必需的资源, 是影响其形态和功能的重要因子, 而净光合速率P_n体现了植物有机物的积累。将人参个体P_n作为研究对象, 通过测定可见光光谱成分组成, 结合光合有效辐射, 预测净光合速率变化规律。由于影响模型的因素有很多[6, 7], 需要对模型进行优化[8]。本文详细介绍利用不同公式、不同核函数、不同相关参数的支持向量机建立净光合速率预测模型, 同时详细讨论参数优化、模型组合、信息粒子化的问题, 提高了模型的准确率和泛化能力。

1 实验及数据处理

1.1 样本来源

研究样本来源于吉林省梅河口林场林下参种植基地, 地理位置东经125.5°, 北纬42.2°, 海拔514 m。全年日照时数2 556 h, 年平均气温4.6°, 年平均降水量798 mm。所选试验样地为20年生同龄人工落叶松林, 位于海拔514 m, 坡向172°, 坡度13°, 林分郁闭度0.8, 平均树高14.8 m。以林下生长8年龄人参个体为研究对象, 测定净光合速率、光合有效辐射和以该人参根部为中心、以20 cm为半径所划生态位的可见光光谱组成成分配比关系。

1.2 光谱测定

采用美国生产的MSR-16型便携式多光谱辐射仪。根据仪器性质, 应用辐射计上方入射辐射的460~710 nm波段光谱分析林下可见光成分和比例改变。MSR-16多波段光谱仪测定的可见光波段能量以电压单位mV表示。时间跨度为2009年7月1日至7月28日。为了提高模型的泛化能力, 能在较大范围内正确预测人参个体净光合速率的值, 测试原则是所测的净光合速率和光合有效辐射值逐日升高, 具体方法如下。测量时间为试验当日9:00~16:00, 每一小时测定

收稿日期: 2010-09-26, 修订日期: 2010-12-20
基金项目: 国家自然科学基金项目(30871452)资助
作者简介: 武海巍, 1978年生, 吉林大学工程仿生教育部重点实验室博士研究生 e-mail: wuhwjlu08@mails.jlu.edu.cn
* 通讯联系人 e-mail: haiye@jlu.edu.cn

一次，测杆高度 1.7 m，每日随机选取一个测试地点，每试验单位5个定点测量，每点测量两次取平均，5点平均，若比前一日同一时间的光合有效辐射值大，则作为该试验单位的光谱组成，否则换新的试验单元，重新测量。每日所测对应数据取均值作为该日该试验单元相关数据，如图1所示。

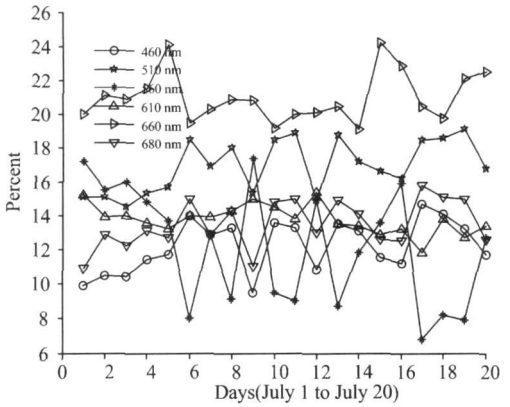

Fig 1　Percent of visible spectrum composition

2　建立净光合速率预测模型

2.1　影响净光合速率因素分析与数据处理

林下参光环境复杂，影响人参个体净光合速率的因素很多，并且很多因素之间有着互相制约的关系，很难确定个体净光合速率 P_n 和各个因素之间关系模型。经试验研究[9, 10]表明，个体净光合速率 P_n 与光合有效辐射 PAR、可见光光谱组成成分配比关系、叶温 T_1、散射辐射 PFDdif、气温 T_a、空气中 CO_2 浓度 c_a、直射辐射 PFDdir、空气相对湿度 RH、气孔导度 G_s、胞间 CO_2 浓度 c_i、蒸腾速率 T_r 等具有一定的关系。可见，影响 P_n 的因子维数众多，本研究采用信息粒子化方法进行降维处理。将可见光光谱组成成分配比关系归为一个粒子，称为 P 粒子。将 T_1，PFDdif，T_a，c_a，PFDdir，RH，G_s，c_i，T_r 归为一个粒子，称为 ε 粒子。由于 PAR、P 粒子和 ε 粒子中各个测试数据的单位不相同，采取数据同步归一化方法。归一化后 P 粒子数据如图2所示。

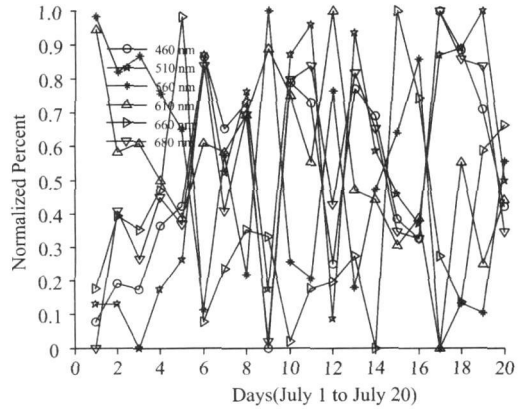

Fig 2　Normalized percent of visible spectrum composition

由图2曲线可见，P 粒子数据在[0, 1]区间变化。由于同步归一化，ε 粒子数据也在[0, 1]区间变化。本研究主要分析 PAR 和可见光光谱组成成分配比关系与 P_n 之间对应关系的模型，对 ε 粒子数据采取在[0, 1]区间随机赋值的数据处理方法。ε 粒子数据的不同取值将对预测模型的准确性有不同影响。归一化后预测集 P 粒子和随机 ε 粒子如图3所示。

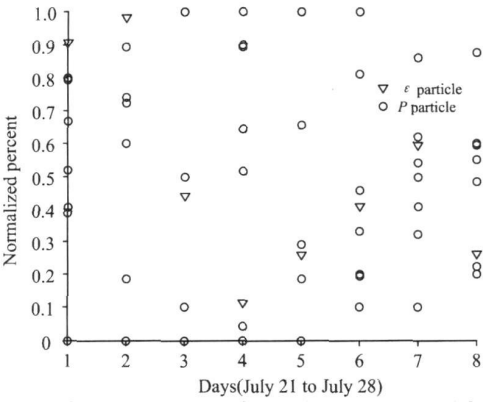

Fig 3　Normalized *P* particle and random ε particle

2.2　支持向量机建模

本研究采用 epsilon-SVR 公式，nu-SVR 公式，linear 核函数，polynomial 核函数，radial basis function 核函数，sigmoid 核函数，惩罚参数 c 和 g 值采用优化算法进行参数寻优。以上多种组合建立不同的支持向量机模型，进行交叉试验，寻求最优模型。利用该模型对加入 ε 粒子的林下人参个体净光合速率进行预测。

2.3　参数寻优

采用 K-fold 交叉验证方法，训练集分为 10 组，分别用 grid-search，Genetic Algorithm(GA)，Particle Swarm Optimization(PSO) 对惩罚参数 c 和 g 值进行参数寻优。

Table 1　Various combinations algorithm correlation coefficient *R* based on epsilon-SVR

	Grid-search/%	GA/%	PSO/%
Linear	96.785 7	96.560 5	96.076 1
Polynomial	86.648 3	87.296 2	86.648 3
RBF	95.511 6	97.998 2	95.708 6
Sigmoid	94.157 6	95.969 9	94.204 7

Table 2　Various combinations algorithm correlation coefficient *R* based on nu-SVR

	Grid-search/%	GA/%	PSO/%
Linear	98.045 5	92.058 5	93.251 7
Polynomial	84.671 3	84.671 3	84.671 3
RBF	95.018 4	98.960 7	95.510 4
Sigmoid	94.464 1	97.189 3	93.997 2

2.4　结果分析

寻求最优模型阶段，训练集为7月1日至7月20日的 P 粒子和 PAR，预测集为7月21日至7月28日 P 粒子和 PAR。支持向量机中多种公式、核函数、参数优化方法的组合，会产生不同的预测效果。如表1和表2中所示。

由表 1 可见，epsilon-SVR-RBF-Genetic Algorithm 模型，称为 ERGA 模型，使得预测集的拟合程度最高，达到了 97.998 2%。ERGA 模型预测结果如图 4—图 6 所示。

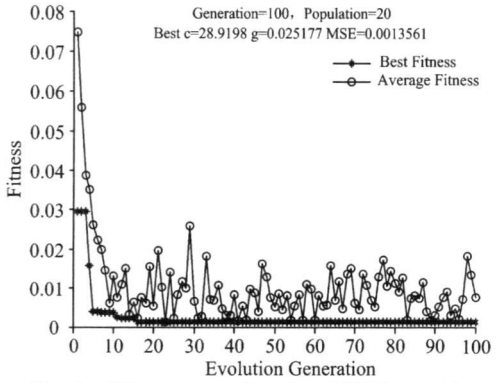

Fig. 4 Fitness curve based on ERGA model

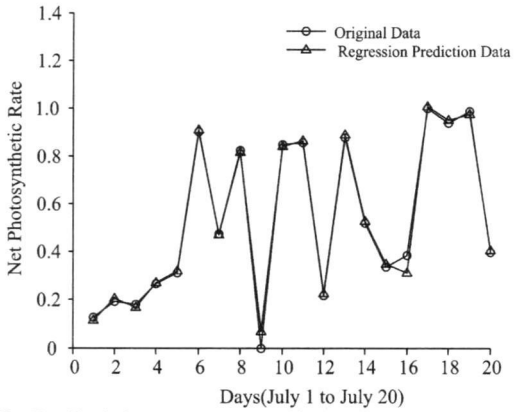

Fig. 5 Training set regression forecasting curve of net photosynthetic rate based on ERGA model

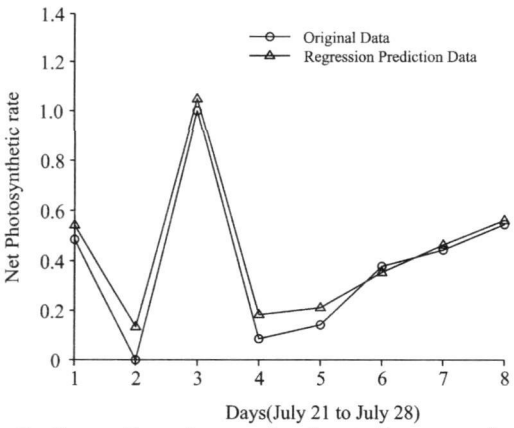

Fig. 6 Forecasting set regression forecasting curve of net photosynthetic rate based on ERGA model

由表 2 可见，nu-SVR-linear-gris-search 模型，称为 NLGS 模型，nu-SVR-RBF-Genetic Algorithm 模型，称为 NRGA 模型，这两个模型下，预测集均有较高的拟合程度，分别为 98.045 5% 和 98.960 7%。NLGS 模型预测结果如图 7—图 9 所示。

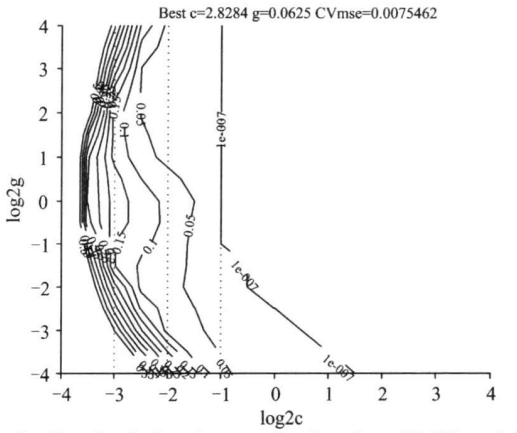

Fig. 7 Result of choosing c and g based on NLGS model

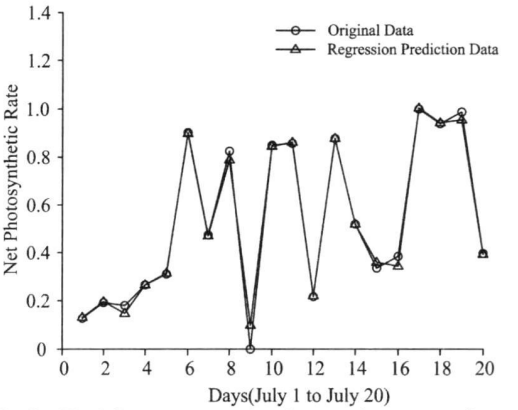

Fig. 8 Training set regression forecasting curve of net photosynthetic rate based on NLGS model

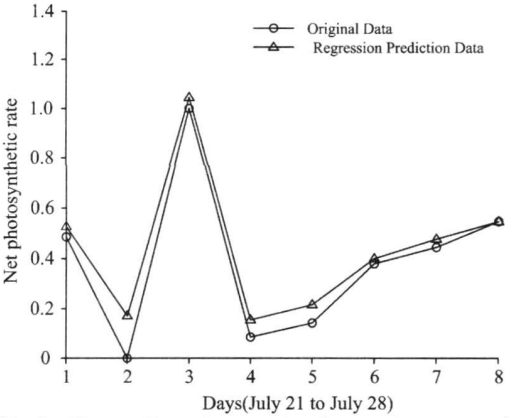

Fig. 9 Forecasting set regression forecasting curve of net photosynthetic rate based on NLGS model

NRGA 模型预测结果如图 10—图 12 所示。

三种模型各自所采用的惩罚参数 c 和对应的相关系数 R 如表 3 所示。

模型中惩罚参数 c 的值越小，说明该模型的泛化能力越强。由表 3 可见，三个模型对预测集拟合程度水平相当，但惩罚参数 c 相差较大，其中 NLGS 模型的惩罚参数 c 值最小。

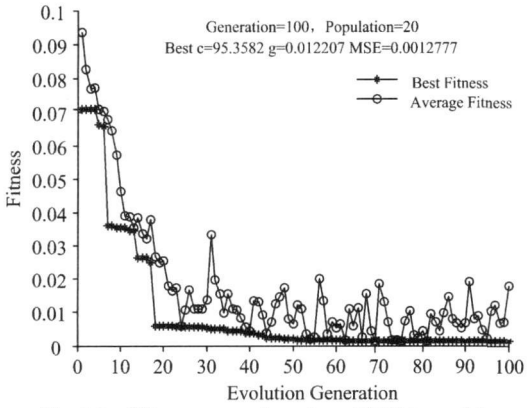

Fig. 10　Fitness curve based on NRGA model

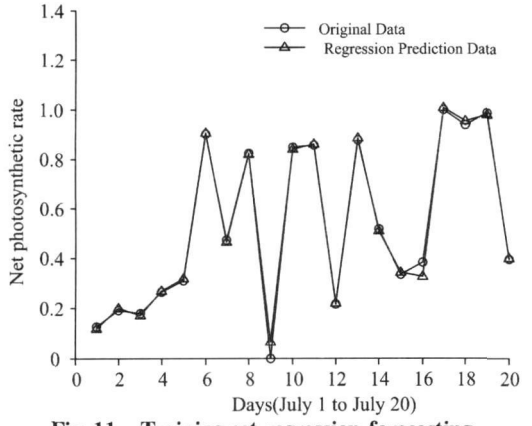

Fig. 11　Training set regression forecasting curve based on NRGA model

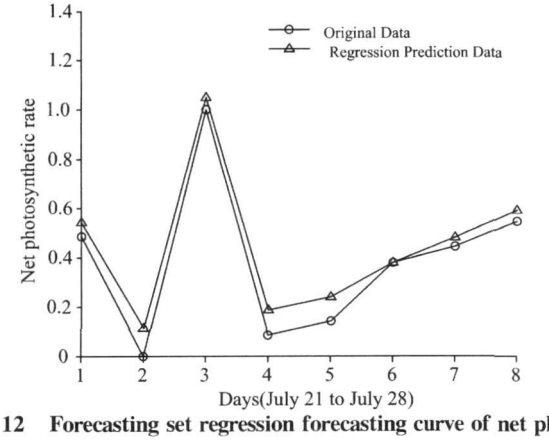

Fig. 12　Forecasting set regression forecasting curve of net photosynthetic rate based on particle swarm optimization

Table 3　Three models *c* and *R*

	ERGA model	NLGS model	NRGA model
c	28 919 8	2 828 4	95 358 2
R	97. 998 2%	98. 055 5%	98 960 7%

综合考虑，本研究采用了 NLGS 模型。为了更真实体现其他自然环境因子对 *P*n 的影响，在模型应用阶段，预测集 *P* 粒子中加入 ε 粒子。由于进行了同步归一化处理以及 ε 粒子数据的不确定性，将 ε 粒子数据从 0 逐步增加到 1，步长为 0 01。经过交叉试验，ε 粒子中对 NLGS 模型准确性影响最大的数据如表 4 所示。

Table 4　Worst correlation coefficient *R* after mixing ε particle

$\varepsilon_{particle}$	R/%
0 829 6, 0 463 9, 0 085 6, 0 324 4, 0 544 1, 0 670 7, 0 222 0, 0 043 6	96 549 5

加入该 ε 粒子后，NLGS 模型的预测效果如图 13 所示。

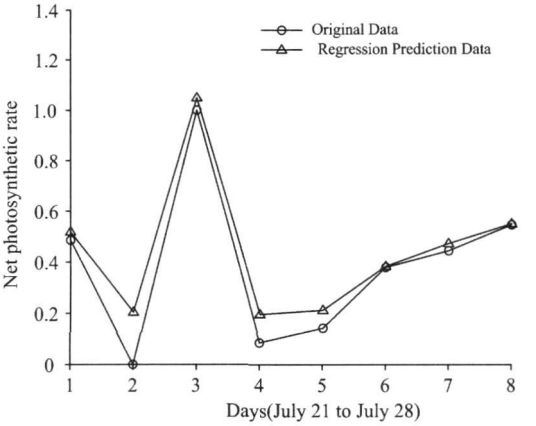

Fig 13　Forecasting set with mixed ε particle regression forecasting curve based on NLGS model

可见，引入 ε 粒子后，拟合程度由 98 045 5% 下降到 96 549 5%。但是整体来看，NLGS 模型对净光合速率变化的预测结果令人满意。

3　结　论

本文采用多种支持向量机公式、核函数和参数优化算法进行了 K-fold 交叉试验，通过信息粒子化技术，对影响人参个体净光合速率的因素进行了粒子化处理，解决了林下光环境复杂的关键问题。在进行相应数据预处理前提和 NLGS 模型作用下，对林下参光环境中的个体净光合速率进行有效预测，拟合程度达到了 96% 以上。由试验结果可见，个体净光合速率和光合有效辐射、可见光光谱组成成分配比关系之间具有相当的相关性，为林下参光环境中人参个体和群体动态受光动态模型的进一步分析提供了理论依据、技术支持和试验方法。

References

[1] XIN Zhi-yun, GU Ming(辛治运, 顾 明). Journal of Tsinghua University(Sci. & Tech.)(清华大学学报·自然科学版), 2008, 48(7): 1147.

[2] Manevitz L M, Yousef M. Journal of Machine Learning Research, 2001, 2: 139.

[3] HUANG Qian, WANG Zhen, WEI Tao, et al(黄 谦, 王 震, 韦 韬, 等). Computer Engineering(计算机工程), 2006, 32(16): 127.

[4] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer Press, 1995.

[5] Evqeniou T, Pontil M. Machine Learing and Its Application. Advanced Lectures, Springer Publisher, 2005. 249.

[6] YU Ke, CHENG Yi-yu(虞 科, 程翼宇). Chinese Journal of Analytical Chemistry(分析化学), 2006, 34(4): 561.

[7] HOU Zhen-yu, CAI Wen-sheng, SHAO Xue-guang(侯振雨, 蔡文生, 邵学广). Chinese Journal of Analytical Chemistry(分析化学), 2006, 34(5): 617.

[8] YANG Jian-lei, ZHU Tuo, XU Yan, et al(杨建磊, 朱 拓, 徐 岩, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2010, 30(1): 243.

[9] WU Haiwei, YU Haiye, YANG Haoyu, et al. 2010 International Conference on Machine Vision and Human-Machine Interface, 2010. 109.

[10] YU Hai-ye, ZHANG Lei, ZHOU Li-na, et al(于海业, 张 蕾, 周丽娜, 等). Journal of Jilin Agricultural University(吉林农业大学学报), 2007, 29(3): 237.

Prediction Model of Net Photosynthetic Rate of Ginseng under Forest Based on Optimized Parameters Support Vector Machine

WU Hai-wei[1, 2], YU Hai-ye[1*], ZHANG Lei[1]

1. Key Laboratory of Bionics Engineering, Ministry of Education, Jilin University, Changchun 130022, China
2. School of Electrical and Information Engineering, Beihua University, Jilin 132021, China

Abstract Using K-fold cross validation method and two support vector machine functions, four kernel functions, grid-search, genetic algorithm and particle swarm optimization, the authors constructed the support vector machine model of the best penalty parameter c and the best correlation coefficient. Using information granulation technology, the authors constructed P particle and ε particle about those factors affecting net photosynthetic rate, and reduced these dimensions of the determinant. P particle includes the percent of visible spectrum ingredients. ε particle includes leaf temperature, scattering radiation, air temperature, and so on. It is possible to obtain the best correlation coefficient among photosynthetic effective radiation, visible spectrum and individual net photosynthetic rate by this technology. The authors constructed the training set and the forecasting set including photosynthetic effective radiation, P particle and ε particle. The result shows that epsilon-SVR-RBF-genetic algorithm model, nu-SVR-linear-grid-search model and nu-SVR-RBF-genetic algorithm model obtain the correlation coefficient of up to 97% about the forecasting set including photosynthetic effective radiation and P particle. The penalty parameter c of nu-SVR-linear-grid-search model is the minimum, so the model's generalization ability is the best. The authors forecasted the forecasting set including photosynthetic effective radiation, P particle and ε particle by the model, and the correlation coefficient is up to 96%.

Keywords Visible spectrum; Support vector machine; Optimized parameters; Information granulation

(Received Sep. 26, 2010; accepted Dec. 20, 2010)

* Corresponding author