第31卷，第11期
2011年11月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 31，No. 11，pp3010-3013
November，2011

基于 Mann-Whitney 非参数检验和 SVM 的竹类高光谱识别

陈永刚，丁丽霞，葛宏立，张茂震，胡 芸

浙江农林大学，浙江省森林生态系统碳循环与固碳减排重点实验室，环境科技学院，浙江 临安 311300

摘 要 以毛竹、雷竹和孝顺竹野外高光谱数据为基础，在非参数统计理论和模式识别的基础上，提出了利用 Mann-Whitney 非参数检验筛选竹类间最佳特征区分波段及利用支持向量机识别竹类的问题。研究结果表明：(1) 毛竹与雷竹之间的最佳区分波段为 503～655，689～732，757～1 000，1 038～1 084，1 238～1 311，1 404～1 591，1 682～1 800，1 856～1 904 和 1 923～2 500 nm，毛竹与孝顺竹之间的最佳区分波段为 350～386，731～1 430，1 584～1 687，1 796～1 873 nm，雷竹与孝顺竹之间的最佳区分波段为 355～356，498～662，689～745 和 1 344～2 500 nm；利用 Mann-Whitney 非参数检验方法可以分别消除 30.0%，57.7%和 35.8% 的无效区分波段。(2) 在最佳区分波段内，利用支持向量机的 SMO 算法进行高光谱竹类识别，模型精度分别为 98.4%，93.5%和 95.1%，模型泛化精度分别为 93.3%，90.0%和 86.7%，表明此方法可有效区分和识别竹亚科中的不同竹类。

关键词 Mann-Whitney 检验；高光谱；SVM；竹亚科

中图分类号：S771.8 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)11-3010-04

引 言

高光谱数据波段多、数据冗余度大，利用高光谱数据进行数据识别时，选择提取特征光谱至关重要[1,2]。王渊、王志辉等分别利用光谱微分技术对油菜、南方常见树种进行高光谱特征选择及识别并取得了较好的效果。张健楠等利用非参数回归与最近邻方法实现了一种有效的恒星光谱自动分类方法。Rick Archibald 等利用 SVM 算法对 AVIRIS 影像进行特征提取和识别[3]。邢飞、王圆圆等采用支持向量机进行光谱特征提取和识别，发现支持向量机识别精度高且有较强的泛化能力。但是迄今还无利用 Mann-Whitney 非参数检验进行竹类高光谱特征波段提取和 SVM 光谱识别竹类的报道。本文提出了在 Mann-Whitney 非参数检验筛选竹类间最佳区分波段的基础上，而后，利用支持向量机 SMO 算法来区分识别同一竹亚科中三种不同的竹类，并在实验中验证了该方法的有效性和可行性。

1 数据与方法

1.1 仪器设备与数据获取

竹类数据采用 ASD Fieldspec Pro FR 野外光谱辐射仪测量。光谱仪波长范围为 350～2 500 nm，共 2 151 个波段，波长精度为±1 nm。在浙江省浙江农林大学校院翠竹园内采集竹类光谱数据。在竹叶新鲜状态下，使用光谱仪的植被高密度探头进行光谱测量，每测完 10 片叶面进行一次标准白板的校正，以保证数据的准确性。以竹亚科中毛竹、雷竹、孝顺竹 3 种外形较相似的竹类进行试验，每种竹类采集 45 份叶片样本，其中 30 份记录作为学习样本数据，15 份作为测试样本数据。

1.2 Mann-Whitney 非参数检验

非参数检验是一种与总体分布状况无关的检验方法，它不依赖于总体分布的形式，不考虑被研究对象为何种分布及分布类型是否已知。当样本观测值的总体分布类型未知或知之甚少，无法肯定其性质，不具备参数检验的应用条件时，非参数检验极具应用价值[4,5]。平均值是数据中心位置的一种很自然的测度，但其缺点是对极端值异常敏感，而中位数却对极端值不敏感，是数据中心位置的稳定测度[4]。1947 年 Mann 和 Whitney 在 Wilcoxon 非参数检验的基础上，提出了针对独立非成对样本总体间是否存在显著差异的 Mann-Whitney 检验[6]。Mann-Whitney 检验以中位数为测度，其假设检验原假设为 H_0：$X_1=Y_2$ 与备择假设为 H_1：$X_1\neq Y_2$，其中 X_1 和 Y_2 为两个数据总体的中位数。其有结点时大样本近似的 Z 值修正公式为式(1)所示[5]。

收稿日期：2011-02-22，修订日期：2011-06-29

基金项目：国家自然科学基金项目(30771725，30972360)和浙江省教育厅项目(Y201017891)资助

作者简介：陈永刚，1980 年生，浙江农林大学环境科技学院讲师 e-mail：cyg_gis@163.com

$$Z=\frac{W_{xy}-mn/2}{\sqrt{\frac{mn(m+n+1)}{12}-\frac{mn\left(\sum_{i=1}^{g}\tau_i^3-\sum_{i=1}^{g}\tau_i\right)}{12(m+n)(m+n-1)}}} \quad (1)$$

因在每个波段上竹类反射率测量样本的统计分布未知，无法进行传统的参数检验，根据非参数检验的特点，分别对 350～2 500 nm 范围内的共 2 151 个波段，竹类两两之间的反射率测量样本组进行 Mann-Whitney 检验，以判定在此波段下两种竹类是否可分。如果接受 H_0，则表示在此波段下两种竹类不可分；如果接受 H_1 则表示在此波段下，两种竹类可分，此波段即是这两种竹类的最佳区分波段。

1.3　支持向量机 SVM 和 SMO 最小序列化算法

支持向量机 SVM 是 Vapnik 和 Boser 于 1992 年首先提出，它在解决小样本、非线性及高维模式识别中有其独特的优势[7-10]。支持向量机方法是建立在统计学习理论、最大间隔超平面、Mercer 核、凸二次规划和松弛变量等多项技术和原理基础上的，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以获得最好的模型泛化能力，其对非线性决策边界具有高度准确的全局建模能力且不会产生过拟合[8]。SVM 的关键在于核函数，低维空间向量集通常难于划分，将它们映射到高维空间而使其得以解决。核函数巧妙地解决了高维计算复杂度的问题。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。常用的核函数有：多项式核、高斯径向基核、Sigmoid 核等。

非线性 SVM 可用经典的二次规划方法求解，但同时求解 n 个拉格朗日乘子涉及多次迭代，计算量开销巨大，因而一般采用 SMO 最小序列化算法。其基本思路是每次只更新两个乘子，迭代获得最终解。SMO 算法的优点在于，只有两个变量的二次规划问题存在解析解。它可保证解的全局最优性，不存在陷入局部极小解的问题；分类器复杂度由支撑向量的个数，而非特征空间的维数决定，因此较少会因维数灾难而发生过拟合现象[11]。其缺点是需要求解二次规划问题，其规模与训练量成正比，因此计算复杂度高且存储开销大[12]。

2　试验与分析

2.1　以 Mann-Whitney 非参数检验筛选竹类最佳区分波段

对野外测量得到的毛竹、雷竹和孝顺竹各 30 份光谱学习样本数据，分别按毛竹与雷竹、毛竹与孝顺竹、雷竹与孝顺竹两两成一组，共分成 3 组，对每组合计 60 份样本数据分别进行分析。在 R 语言下编写程序，调用 wilcox. test()函数分别对此 3 组学习样本数据进行 Mann-Whitney 非参数检验，以判定其在不同波段下组内两种竹类的光谱反射率是否具有显著差异，以确定其在此波长下是否可分。其中，0 表示不可分，1 表示可分。结果如表 1—表 3 和图 1—图 3 所示，图中黑色部分表示在置信度为 99%下竹类间两两可分。

Table 1　Mann-Whitney test of MaoZhu and LeiZhu

波长/nm	350	351	352	…	2 498	2 499	2 500
统计量 W	375	389	347	…	640	626	627
概率 P	0.271	0.371	0.130	…	0.005	0.009	0.008
是否可分	0	0	0	…	1	1	1

Table 2　Mann-Whitney test of MaoZhu and XiaoShunZhu

波长/nm	350	351	352	…	2 498	2 499	2 500
统计量 W	223	181	154	…	426	439	446
概率 P	0.001	0.000	0.000	…	0.730	0.878	0.959
是否可分	1	1	1	…	0	0	0

Table 3　Mann-Whitney test of LeiZhu and XiaoShunZhu

波长/nm	350	351	352	…	2 498	2 499	2 500
统计量 W	366	307	302	…	242	248	262
概率 P	0.217	0.035	0.029	…	0.002	0.003	0.005
是否可分	0	0	0	…	1	1	1

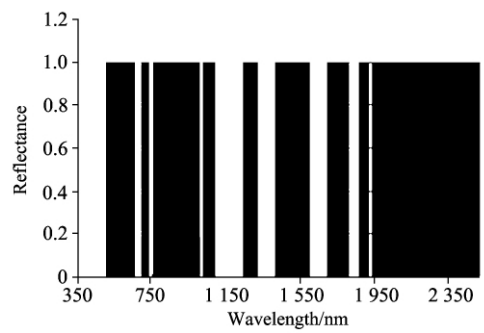

Fig. 1　Optimal discriminating band between MaoZhu and LeiZhu

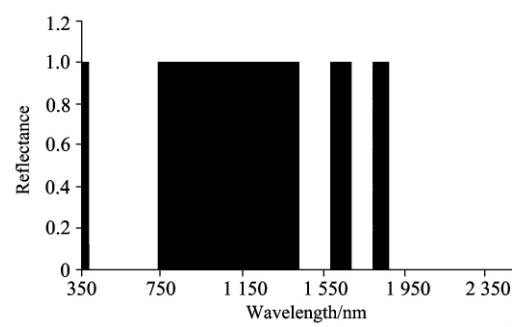

Fig. 2　Optimal discriminating band between MaoZhu and XiaoShunZhu

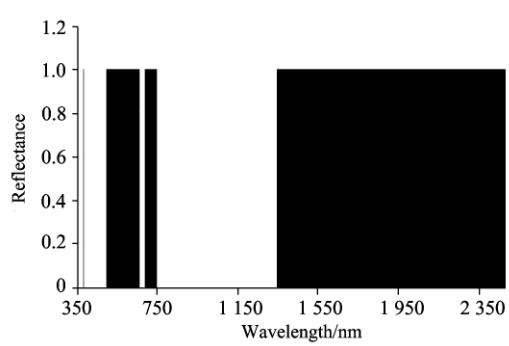

Fig. 3　Optimal discriminating band between LeiZhu and XiaoShunZhu

分析结果发现：在置信度为 99%下，毛竹与雷竹之间的最佳区分波段为 503～655，689～732，757～1 000，1 038～1 084，1 238～1 311，1 404～1 591，1 682～1 800，1 856～1 904和 1 923～2 500 nm，毛竹与孝顺竹之间的最佳区分波段为 350～386，731～1 430，1 584～1 687 和 1 796～1 873 nm，雷竹与孝顺竹之间的最佳区分波段为 355～356，498～662，689～745 和 1 344～2 500 nm。利用 Mann-Whitney 非参数检验有效的实现了特征区分波段的选择和无效区分波段的消除，无效区分波段消除率分别为 30.0%、57.7%和 35.8%。

2.2 基于 SVM 的竹类高光谱识别及精度评价

利用支持向量机的 SMO 最小序列化算法，对 3 组经 Mann-Whitney 非参数检验特征区分波段筛选后的光谱学习样本数据，进行 SVM 分类识别和精度评价。且利用 10-折交叉检验(10-folder cross-validation)方法，对竹类高光谱分类识别模型进行精度检验。混淆矩阵和模型精度评价计算结果如表 4—表 9 所示。

Table 4 Confusion matrix of MaoZhu and LeiZhu

	被分类归于“毛竹”	被分类归于“雷竹”
实际属于“毛竹”	30	0
实际属于“雷竹”	1	29

Table 5 Model accuracy evaluation of MaoZhu and LeiZhu by SVM classification method

项目	TP	FP	Precision	Recall	F-Measure	ROC
毛竹	1.000	0.033	0.968	1.000	0.984	0.983
雷竹	0.967	0.000	1.000	0.967	0.983	0.983
加权平均值	0.983	0.017	0.984	0.983	0.983	0.983

Table 6 Confusion matrix of MaoZhu and XiaoShunZhu

	被分类归于“毛竹”	被分类归于“孝顺竹”
实际属于“毛竹”	29	1
实际属于“孝顺竹”	3	27

Table 7 Model accuracy evaluation of MaoZhu and XiaoShunZhu by SVM classification method

项目	TP	FP	Precision	Recall	F-Measure	ROC
毛竹	0.967	0.100	0.906	0.967	0.935	0.933
孝顺竹	0.900	0.033	0.964	0.900	0.931	0.933
加权平均值	0.933	0.067	0.935	0.933	0.933	0.933

Table 8 Confusion matrix of LeiZhu and XiaoShunZhu

	被分类归于“雷竹”	被分类归于“孝顺竹”
实际属于“雷竹”	29	1
实际属于“孝顺竹”	2	28

经计算发现，三组竹类高光谱分类识别模型的 Kappa 统计量分别为 0.967，0.867 和 0.900，竹类高光谱分类识别精度分别为 98.4%，93.5%和 95.1%。可以认为，SVM 竹类高光谱识别模型具有较好的一致性和精度。

Table 9 Model accuracy evaluation of LeiZhu and XiaoShunZhu by SVM classification method

项目	TP	FP	Precision	Recall	F-Measure	ROC
雷竹	0.967	0.067	0.935	0.967	0.951	0.950
孝顺竹	0.933	0.033	0.966	0.933	0.949	0.950
加权平均值	0.950	0.050	0.951	0.950	0.950	0.950

2.3 以测试样本检验竹类识别模型精度

在实际应用之前，不但需从学习样本评价模型精度，还必须经过测试样本测试以确认其分类精度。只有确认分类精度达到预定要求后，才能实际应用该模型。测试是以学习得到的模型在“测试样本”上进行识别检验，以此评价分类器的性能。根据所建立的三个 SVM 竹类高光谱识别模型，分别用 15 个样本进行测试。测试样本检验的模型精度结果如表 10 所示。

Table 10 Model accuracy for testing samples

区分竹类	测试样本	样本总数	分类正确数	分类错误数	样本分类精度/%	模型平均分类精度/%
毛竹～雷竹	毛竹	15	15	0	100.0	93.3
	雷竹	15	13	2	86.7	
毛竹～孝顺竹	毛竹	15	12	3	80.0	90.0
	孝顺竹	15	15	0	100.0	
雷竹～孝顺竹	雷竹	15	11	4	73.3	86.7
	孝顺竹	15	15	0	100.0	

从表 10 可以看出，用 SVM 竹类高光谱识别模型对毛竹、雷竹和孝顺竹测试样本进行识别分类，识别平均精度分别为 93.3%，90.0%和 86.7%。结果证明，利用 Mann-Whitney 非参数检验提取特征区分波段和利用 SVM 算法进行竹类高光谱识别是可行的和有效的。

3 结 论

(1)以不依赖于统计总体分布形式和不易被极端值扰动的 Mann-Whitney 非参数检验方法为基础，选择竹类间的最佳区分特征波段。实验发现，利用 Mann-Whitney 非参数检验方法可以有效消除无效区分波段，消除率分别为 30.0%，57.7%和 35.8%，有效的约减了数据的维度。从而为竹类特征区分波段的筛选和提取提供了适用可行的波段选择方法。

(2)利用支持向量机处理非线性和高维数据的特点，以及 SMO 全局最优解的特性，对三组竹类学习样本数据进行支持向量机识别，10-折交叉检验结果发现模型的 Kappa 统计量分别为 0.967，0.867 和 0.900；模型精度分别为 98.4%，93.5%和 95.1%，模型具有较好的一致性和精度。

(3)利用测试样本对支持向量机分类模型的泛化精度进行检验，平均精度分别为 93.3%，90.0%和 86.7%，结果证明，利用 Mann-Whitney 非参数检验提取特征区分波段和利用支持向量机 SMO 算法进行竹类高光谱识别是可行、有效

的。本方法为竹类高光谱精细识别及波段探索分析提供了一种快捷、方便的研究手段如何消除竹亚科光谱试验数据中异常样本的影响是下一步研究的主要内容。

References

[1] De Backer S, Kempeneers P, Debruyn W, et al. Geoscience and Remote Sensing Letters, IEEE, 2005, 2(3): 319.
[2] Guo B, Gunn S R, Damper R I, et al. Geoscience and Remote Sensing Letters, IEEE, 2006, 3(4): 522.
[3] Archibald R, Fann G. IEEE Geoscience and Remote Sensing Letters, 2007, 4(4): 674.
[4] Rosner B A. Fundamentals of Biostatistics: Duxbury Resource Center, 2006.
[5] Wasserman L. All of Nonparametric Statistics: Springer-Verlag New York Inc, 2006.
[6] Mann H B, Whitney D R. The Annals of Mathematical Statistics, 1947, 18(1): 50.
[7] Soman K P, Diwakar S, Ajay V. Insight into Data Mining: Theory and Practice: PHI Learning Pvt. Ltd., 2006.
[8] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining: Pearson Addison Wesley Boston, 2006.
[9] Han J, Kamber M. Data Mining: Concepts and Techniques 2nd ed: Morgan Kaufmann, 2006.
[10] Cortes C, Vapnik V. Machine Learning, 1995, 20(3): 273.
[11] Shevade S K, Keerthi S S, Bhattacharyya C, et al. IEEE Transactions on Neural Networks, 2002, 11(5): 1188.
[12] MENG Geng-xiang, FANG Jing-long(蒙庚祥, 方景龙). Computer Engineering and Design(计算机工程与设计), 2005, 26(6): 1592.

Hyperspectral Bambusoideae Discrimination Based on Mann-Whitney Non-Parametric Test and SVM

CHEN Yong-gang, DING Li-xia, GE Hong-li, ZHANG Mao-zhen, HU Yun
Zhejiang Provincial Key Laboratory of Carbon Cycling in Forest Ecosystems and Carbon Sequestration, College of Environmental Science and Technology, Zhejiang Agriculture and Forestry University, Lin'an 311300, China

Abstract In the present study, based on the leaf-level hyperspectral data of MaoZhu, LeiZhu and XiaoShunZhu, We come up with two solutions to discrimination through the theory of non-parametric test and pattern recognition; the first one is that optimal discriminating band between bambusoideae species is extracted by Mann-Whitney non-parametric test, the other is that bambusoideae species is discriminated by the support vector machine. The research results showed that (1) the optimal discriminating band between MaoZhu and LeiZhu is around 503～655, 689～732, 757～1 000, 1 038～1 084, 1 238～1 311, 1 404～1 591, 1 682～1 800, 1 856～1 904, and 1 923～2 500 nm; the optimal discriminating band between MaoZhu and XiaoShunZhu is around 350～386, 731～1 430, 1 584～1 687, and 1 796～1 873 nm; the optimal discriminating band between LeiZhu and XiaoShunZhu is around 355～356, 498～662, 689～745, and 1 344～2 500 nm; and it can eliminate 30.0%, 57.7%, and 35.8% of the invalid distinction between bands by Mann-Whitney non-parametric test method. (2) In these optimal discriminating bands, we found that the accuracy of bambusoideae discrimination is 98.4%, 93.5%, and 95.1%, the generalization accuracy is 93.3%, 90.0%, and 86.7% by sequential minimal optimization algorithm. It indicates that this method is valid for selecting feature band and discriminating bambusoideae species.

Keywords Mann-Whitney test; Hyperspectral; SVM; Bambusoideae

(Received Feb. 22, 2011; accepted Jun. 29, 2011)