ICA与PCA在高光谱数据降维分类中的对比研究

臧 卓[1,2],林 辉[2],杨敏华[1]
(1.中南大学 信息物理工程学院,湖南 长沙 410083;
2.中南林业科技大学 林业遥感信息工程研究中心,湖南 长沙 410004)

摘 要: 利用独立主成分算法与主成分分析法分别对原始数据及3种预处理数据进行降维运算,再利用常用的4种分类算法分类,对比分类结果发现,独立主成分算法与主成分分析算法在乔木树种高光谱数据降维中并不具有非常明显的优势,且独立分量分析(ICA)算法提取用于分类的数据不如PCA算法稳定;从计算机的运行成本上来看,PCA算法优于ICA算法,基本上ICA算法平均成本是PCA算法的4到5倍;不同的数据预处理及降维方法组合对分类结果影响明显,d(log(*R*))结合PCA算法,log(*R*)结合ICA算法降维结果最理想;通过比较乔木树种高光谱数据的分类结果发现,Fisher判别法最适合对PCA和ICA降维结果进行分类。

关键词: 高光谱;降维;分类;独立主成分;主成分
中图分类号: S771.8 **文献标志码:** A **文章编号:** 1673-923X(2011)11-0018-05

Comparative study on descending dimension classification of hyperspectral data between ICA algorithm and PCA algorithm

ZANG Zhuo[1,2], LIN Hui[2], YANG Min-hua[1]
(1. School of Info-Physics and Geomatics Engineering, Central South University, Changsha 410083, Hunan, China;
2. Research Center of Forestry Remote Sensing & Information Engineering, Central South University of Forestry & Technology, Changsha 410004, Hunan, China,)

Abstract: The three kinds of pre-processed hyper-spectra data (first-derivative (d(*R*)), logarithms (log(*R*)), logarithms-first-derivative (d(log(*R*)))) and original data were reduced in dimension by ICA and PCA algorithm, and then classified by Support Vector Machine (SVM)-Gaussian Raial Basis Function (RBF), Support Vector Machine (SVM)-Liner, Back Propagation(BP)neural network and Fisher classification method. The results show that compared with PCA, ICA did not have obvious advantage in dimension reduction for hyperspectral data of trees. ICA was less stable than PCA. Judging by the running cost of computer, PCA algorithm was better than ICA, ICA's average cost was 4 to 5 times more than PCA's. Various pre-processing methods and various classification methods showed the different influences on classification results. By comparing several combination methods, the study has found that d (log(*R*))-PCA combination and log (*R*)-ICA combination were usually considered to be the most ideal. According to the results of the classification, it is found that Fisher classification method is suitable for the classification of the data preprocessed by PCA and ICA.
Key words: hyperspectral; dimension reduction; classification; Independent Component Analysis Principle Component

收稿日期: 2011-05-10
基金项目: 国家自然科学基金资助项目(30871962);高等学校博士学科点专项科研基金项目(200805380001);国家林业局林业公益项目专题(201104028);中南林业科技大学林业遥感信息工程研究中心开放性研究基金项目(RS2008K01)
作者简介: 臧 卓(1978-),男,辽宁锦州人,讲师,博士,主要研究方向:林业遥感及地理信息系统

独立分量分析(ICA)是近年发展起来的一种信号分解技术[1-3]。该方法以非高斯源信号为研究对象，在统计独立的假设下，对多路观测到的混合信号进行盲源分离，分离出隐含在混合信号中的独立信号。

独立主成分与主成分分析(PCA)[4-7]本质都是矩阵变换，区别在于，主成分分析算法是通过光谱数据的矩阵变换尽量保存原有数据的信息，当然原有数据信息既包括不同地物的特征，也包括外来的干扰噪声；而独立主成分首先假设信号源(观测数据)是有误差的，通过矩阵变换去除干扰信息，获取原始信号特征。从理论的出发点上可以发现独立主成分算法优于主成分分析算法，但独立主成分用于提取原始信号特征的数据是观测数据，但观测数据误差的产生原因是多种多样的，从各种影响因素中提取原始信号也是非常困难的。本研究主要是将该方法应用于乔木树种的高光谱数据降维，因为不同树种间的光谱数据差异比不同地类之间差异小的多，因此独立主成分是否能有效提取不同乔木树种类别间的差异，并且独立主成分与主成分分析算法相比较那种更具优势，本文通过后面的降维分类结果进行证明。

1　材料与方法

1.1　试验时间、地点

野外高光谱数据采集试验分别于 2004、2005、2006、2010、2011 年，在湖南省株洲市攸县黄丰桥国有林场进行。

1.2　试验材料

选择黄丰桥林场主要用材林树种，即，杉木 *Cunninghamia lanceolata*、马尾松 *Pinus massoniana*、樟树 *Cinnamomum camphora* 的冠层叶片作为观测对象。

1.3　试验方法

1.3.1　试验设计

分别于不同年份观测马尾松、杉木及樟树林的冠层高光谱数据，并利用独立组成分算法(ICA)和主成分分析算法(PCA)对采集的高光谱数据进行降维，最后利用支持向量机(SVM－RBF 和 SVM－liner)、BP 神经网络 Fisher 分类法对降维数据进行分类，并比较分类结果。

1.3.2　试验仪器

ASD FieldSpec HandHeldTM 地物光谱仪，美国 ASD(analytical spectral device)公司新产品，波长范围 325～1 075 nm，光谱分辨率为 3.5 nm，视场 25°，重量 1.2 kg。

1.3.3　数据采集

高光谱数据采集选在距树顶端 1～3 m，且不受阳光遮挡处。具体见表 1。

表 1　2004 年～2011 年数据采集†

Table 1　Data collected from 2004 to 2011

树种	时间					合计
	2004	2005	2006	2010	2011	
马尾松	0	0	0	79	14	93
杉木	0	0	0	149	22	171
樟树	17	11	18	0	0	46
合计	17	11	18	228	36	310

† 在本研究中从以上 310 条数据中随机选取 250 条作为训练集，60 条作为测试集。

1.4　数据处理与分析

1.4.1　数据预处理　首先对数据采用 S. Golay 方法进行平滑处理[8]，从而降低噪声对最终分类结果的影响，结果如图 1 所示。另外，由于数据所采集的时间不同，光照条件自然会有一些差异，而且观测的乔木树种的高度也略有差异，不同树种的叶片的疏密程度也各不相同，这些都会增加背景土壤对观测数据的影响。为了消除这些因素的影响，本研究对平滑后的高光谱数据(R)进行 3 种变换操作[9]，即：一阶微分变换(d(R))、对数变换(log(R))、对数一阶微分变换(d(log(R)))。

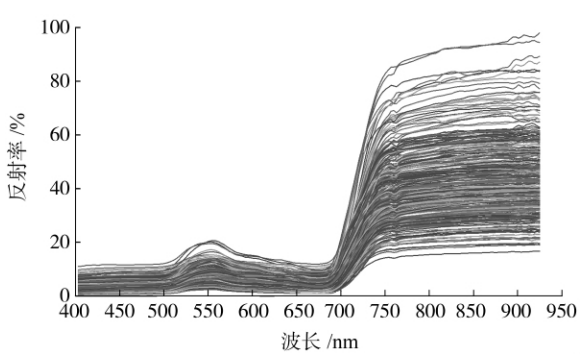


图 1　S. Golay 方法滤波后光谱反射率曲线

Fig. 1　Hyper-spectral filter curves by S. Golay method

1.4.2　ICA 与 PCA 数据降维　选择训练集的 250 条数据分别进行 PCA 变换和 ICA 变换，两种算法主成分个数分别从 1 取到 37，ICA 通过迭代得出各主成分的线性方程，PCA 算法只要获取前 1 至 37 个变换矩阵，最后用测试集的 60 条数据，与前面训练集所建立的方程进行运算，最终得到降维的数据。并对两种变换方法所得数据利用遥感中常用的 4 种分类方法，进行分类对比研究。

1.4.3　分类方法的选择　本研究分别将 PCA 与 ICA 降维结果利用支持向量机[10]（SVM－RBF 和 SVM－Liner）、BP 神经网络[11]、Fisher 分类法[12]，4 种分类算法进行分类，通过对比分类结果，分析 PCA 与 ICA 算法各自的优缺点，并寻求最优的数据预处理－＞降维－＞分类的方法组合。

2　结果与分析

2.1　分类结果对比

支持向量机（SVM－RBF 和 SVM－liner）、BP 神经网络、Fisher 分类法对经过 PCA 与 ICA 变换后的 4 种数据分类结果见图 2。

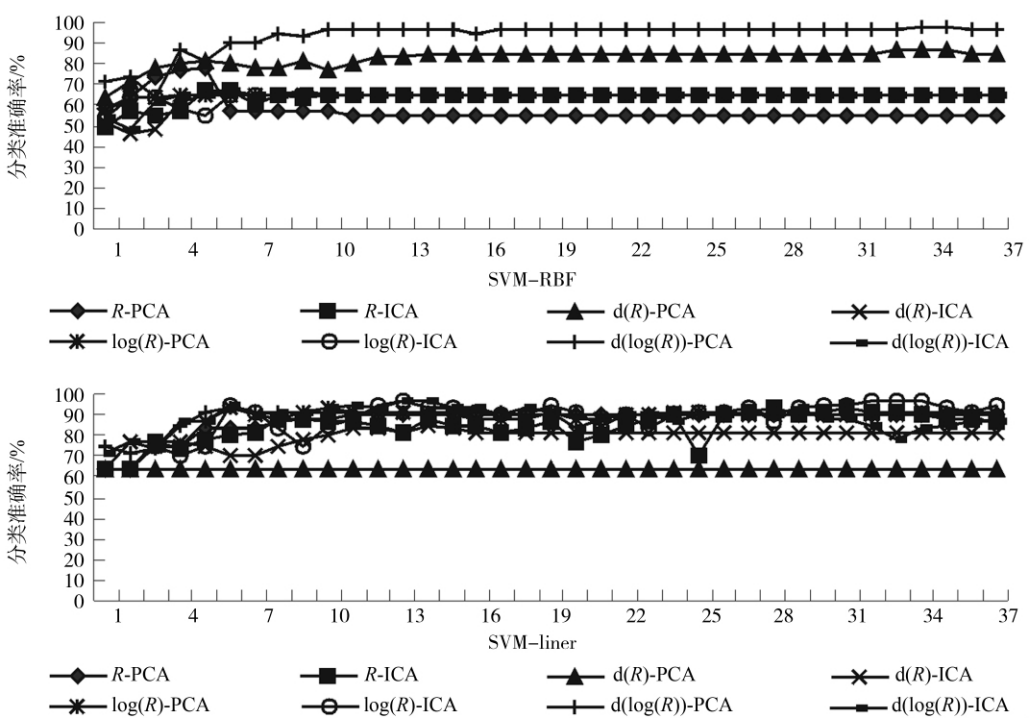
对比图 2 中 32 种方法组合的 1 184 次分类结果可以发现以下特征：

（1）通过对 4 种分类器的分类结果的比较，发现 ICA 降维算法在乔木树种高光谱数据的分类中并不优于 PCA 算法，并且在 R－ICA－SVM－liner 以及 R－ICA－Fisher 等方法组合中，分类结果的波动明显比 R－PCA－SVM－liner 以及 R－PCA－Fisher 方法组合的波动大，这说明 ICA 算法在某些情况下稳定性不如 PCA 算法。

（2）高光谱数据预处理方法对降维分类结果影响较大，log(R)与 d(log(R))变换方法明显优于其它数据预处理方法。

（3）4 种分类方法中最不稳定的是 BP 神经网络，从图形中可以看出 BP 神经网络的分类结果抖动最大，且大多数的分类精度均低于 90%，最低为 23.33%。

（4）利用 SVM－RBF 算法的分类结果受数据预处理以及降维算法影响较大，其中最合适的方法组合是 d(log(R))－PCA－SVM－RBF，最高分类精度为 98.33%。



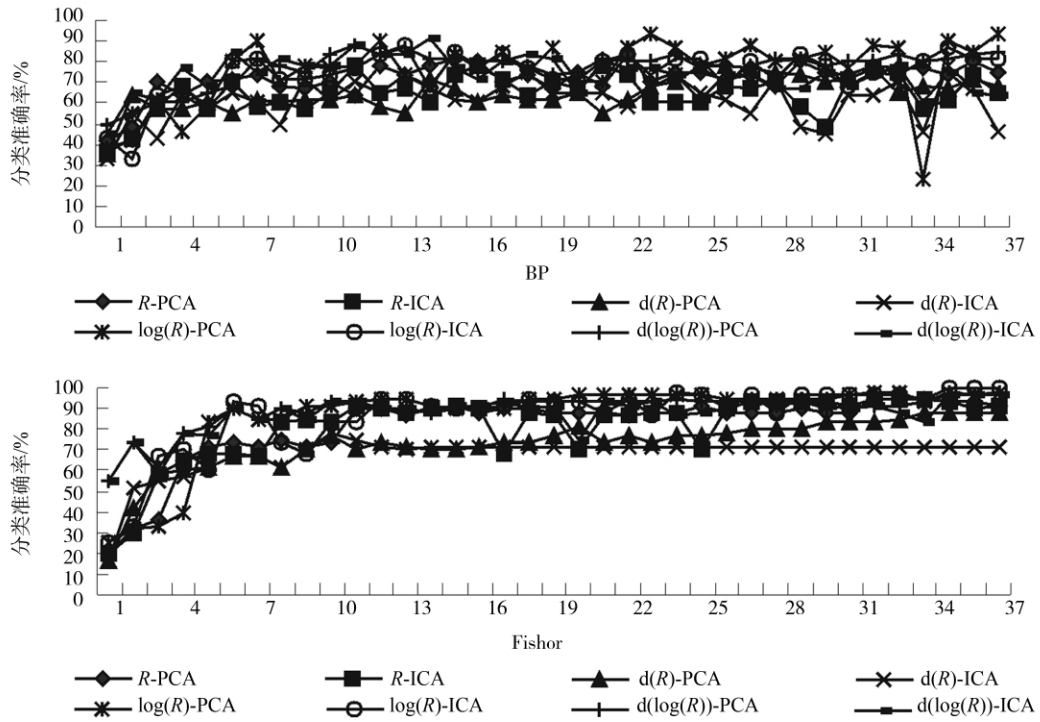


图 2　4 种分类算法分类结果

Fig. 2　Results of 4 kinds classification algorithm

2.2　分类组合方法对比

从上面的分类结果中可以发现，当主成分个数小于 10 个时，随着主成分个数的增加，分类精度与主成分个数基本呈现线性关系，即随主成分个数的增加，分类精度不断提高；当主成分个数大于 10 个时，分类精度基本稳定，波动明显减小，为了对比几种方法组合的稳定性，本研究计算了从 11 个主成分至 37 个主成分之间的分类精度的最小值、最大值、平均值以及标准差，结果如下表所示。

从统计学上可以知道，标准差越小，分类结果越稳定，平均分类精度越高，分类效果越好。因此通过对上面数据的比较分析发现，虽然 log(R)－ICA－Fisher 最高分类精度达到了 100%，但它的最低分类精度为 83.33%，且标准差处于中等水平 0.0427，其分类结果的波动要大于 log(R)－PCA－Fisher 和 d(log(R))－PCA－SVM－RBF 组合，其它部分分类算法虽然标准差为 0，但是因为其最大最小及平均分类精度均较低，因此并不能作为优选组合方式。比较 PCA 与 ICA 降维后数据的分类结果发现，在 32 个方法组合中，当数据预处理方法与分类方法相同时，ICA 算法的标准差有 30 次大于 PCA 算法，这说明 ICA 算法的稳定性不如 PCA 算法。

表 2　分类结果的最大值、最小值、平均值及标准差

Table 2　The maximum and minimum value, mean and standard deviation of various classification and transform

分类法	量和函数	变换法	最大值	最小值	平均值	标准差
SVM-RBF	R	PCA	55.00%	55.00%	55.00%	0.000 0
		ICA	65.00%	65.00%	65.00%	0.000 0
	d(R)	PCA	86.67%	80.00%	84.88%	0.012 2
		ICA	65.00%	65.00%	65.00%	0.000 0
	log(R)	PCA	65.00%	65.00%	65.00%	0.000 0
		ICA	65.00%	65.00%	65.00%	0.000 0
	d(log(R))	PCA	98.33%	95.00%	96.73%	0.005 6
		ICA	65.00%	65.00%	65.00%	0.000 0
BP	R	PCA	83.33%	63.33%	74.57%	0.047 7
		ICA	78.33%	48.33%	67.28%	0.076 0
	d(R)	PCA	78.33%	55.00%	67.47%	0.066 9
		ICA	78.33%	45.00%	63.52%	0.088 1
	log(R)	PCA	93.33%	23.33%	80.06%	0.133 1
		ICA	88.33%	66.67%	78.15%	0.058 9
	d(log(R))	PCA	88.33%	70.00%	80.49%	0.038 3
		ICA	91.67%	60.00%	74.51%	0.081 2
Fisher	R	PCA	95.00%	86.67%	90.06%	0.024 7
		ICA	95.00%	68.33%	88.52%	0.073 3
	d(R)	PCA	88.33%	70.00%	78.33%	0.061 1
		ICA	75.00%	70.00%	71.73%	0.007 3
	log(R)	PCA	98.33%	90.00%	95.19%	0.024 6
		ICA	100.00%	83.33%	93.95%	0.042 7
	d(log(R))	PCA	96.67%	88.33%	92.78%	0.021 7
		ICA	96.67%	75.00%	90.86%	0.043 7

续表 2

Continuation of table 2

分类法	量和函数	变换法	最大值	最小值	平均值	标准差
SVM-liner	R	PCA	90.00%	90.00%	90.00%	0.000 0
		ICA	93.33%	70.00%	86.67%	0.053 3
	d(R)	PCA	63.33%	63.33%	63.33%	0.000 0
		ICA	85.00%	81.67%	81.98%	0.008 1
	log(R)	PCA	91.67%	88.33%	90.06%	0.010 8
		ICA	96.67%	83.33%	92.22%	0.035 2
	d(log(R))	PCA	91.67%	90.00%	90.74%	0.008 4
		ICA	96.67%	78.33%	89.57%	0.043 3

从计算机的运行成本上来看，PCA 算法优势较为明显，以提取 37 个主成分为例，运算 10 次取平均值，PCA 算法用时 0.106 1 秒，ICA 算法用时 0.611 5秒，基本上 ICA 算法平均成本是 PCA 算法的 4 到 5 倍。

3　结论与讨论

通过对独立主成分和主成分分析算法的降维分类结果的比较，得出以下结论。

(1)在乔木树种高光谱数据降维分类中，虽然 ICA 变换后数据的最高分类精度达到 100%，PCA 变换后数据的最高分类精度为 98.33%，但从全局来看，ICA 算法并不比 PCA 算法具有明显优势，而且 ICA 算法提取用于分类的数据不如 PCA 算法稳定。

(2)从计算机的运行成本上来看，PCA 算法优于 ICA 算法，基本上 ICA 算法平均成本是 PCA 算法的 4 到 5 倍。

(3)不同的数据预处理及降维方法组合对分类结果影响明显，d(log(R))结合 PCA 算法，log(R)结合 ICA 算法降维结果最理想。

(4)通过比较乔木树种高光谱数据的分类结果发现，Fisher 判别法最适合对 PCA 和 ICA 降维结果进行分类。

(5)不同的分类算法对 ICA 和 PCA 算法的敏感度不同，在比较 ICA 降维分类结果中，Fisher、SVM－Liner 两种分类算法表现最为理想；而对 PCA 降维分类结果比较中发现，Fisher、SVM－RBF 分类算法表现最理想。

(6)研究通过反复试验发现，在将 ICA 与 PCA 算法应用于乔木树种高光谱数据的降维分类的时候，获取 ICA 或 PCA 分析后的前 20 个主成分可以充分表达原有乔木树种的特征，再结合较好的数据预处理方法和分类方法，使树种的识别达到 90%以上是可以实现的。

参考文献：

[1] 冯　燕，何明一，宋江红，等. 基于独立成分分析的高光谱图像数据降维及压缩[J]. 电子与信息学报，2007，29(12)：2871－2875.

[2] 苏令华，衣同胜，万建伟. 基于独立分量分析的高光谱图像压缩[J]. 光子学报，2008，(5)：973－976.

[3] 刘良春，冯　燕. 结合纯像元提取和 ICA 的高光谱降维方法[J]. 计算机应用研究，2011，3(28)：1184－1185.

[4] 张　亮. 基于 PCA 和 SVM 的高光谱遥感图像分类研究[J]. 光学技术，2008，(12)：184－187.

[5] 祝诗平. 基于 PCA 与 GA 的近红外光谱建模样品选择方法[J]. 农业工程学报，2008，(9)：126－130.

[6] 刘智深，丁　宁，赵　朝，等. 主成分分析法在油荧光光谱波段选择中的应用[J]. 地理空间信息，2009，6(7)：12－15.

[7] 赵春晖，胡春梅，石　红. 采用选择性分段 PCA 算法的高光谱图像异常检测[J]. 哈尔滨工程大学学报，2011，(1)：109－113.

[8] 蔡天净，唐　瀚. Savitzky_Golay 平滑滤波器的最小二乘拟合原理综述[J]. 数字通信，2011，(1)：63－68.

[9] 浦瑞良，宫　鹏. 高光谱遥感及其应用[M]. 北京：高等教育出版社，2000.

[10] Matlab 中文论坛. MATLAB 神经网络 30 个案例分析[M]. 北京：北京航空航天大学出版社，2010，4.

[11] 魏海坤. 神经网络结构设计的理论与方法[M]. 北京：国防工业出版社，2005.

[12] 谢中华. Matlab 统计分析与应用：40 个案例分析[M]. 北京：北京航空航天大学出版社，2010.

[本文编校：吴　毅]