文章编号: 1007-4619(2007)01-0069-08

基于决策树的高光谱数据特征选择及其对分类结果的影响分析

王圆圆, 李 京
(北京师范大学 资源学院 资源技术与工程研究所, 北京 100875)

摘 要: 本文利用 OMIS高光谱数据,研究了决策树算法 (Decision Tree, DT)特征选择的特点以及特征选择对决策树分类结果的影响。设计了三种特征选择方法: SEP, MDLM 和 RELIEF,将它们与 DT特征选择的结果以及特征选择后的分类精度 (考虑了三种分类器:最大似然法、后向传播神经网络、最邻近法)进行对比,并分析了这三种特征选择方法对决策树结构和分类精度的影响。结果显示, DT是一种比较好的特征选择方法;经过特征选择后再生成的决策树比直接生成的决策树,用到更少的特征 (平均减少了 43.36%)、有更多的节点 (平均增加了 18.61%)和更高的分类精度 (平均提高了 0.35%),当样本数量少时,分类精度的提高幅度最大,而树的大小却基本没有增加。

关键词: 决策树; 高光谱; 特征选择
中图分类号: TP751.1 **文献标识码**: A

Analysis of Feature Selection and Its Impact on Hyperspectral Data Classification Based on Decision Tree Algorithm

WANG Yuan-yuan, LI Jing
(College of Resources Science, Beijing Normal University, Beijing 100875, China)

Abstract: In this article, OMIS hyperspctral data was used to study feature selection ability of DT(Decision Tree) algorithm and the impacts of feature selection on DT. The DT was compared to three designed feature selection methods(SEP, MDLM and RELIEF) based on feature selection results and classification accuracy in which three different methods (ML、BPNN and 1-NN) were applied. Moreover, the impacts of the three designed feature selection methods on DT classification results at different training sample sizes were analyzed. Results indicated that DT was a good feature selection method. After feature selection, DT algorithm outputted to those classification trees that used fewer features(average decrease was 43.36%), had fewer tree nodes(average increase was 18.61%), and had higher classification accuracy(average increase was 0.35%). When the training sample size was small, accuracy improvement was the most significant and meanwhile the tree size scarcely changed.

Key words: decision tree; feature selection; hyperspectral data

1 引 言

高光谱遥感数据光谱分辨率高 (<10nm),波段数量大 (可达 200多),与一般遥感数据相比,具有数据量更大的特点,因此分析起来面临更大的困难和挑战。在监督分类中,由于 Hughes现象的存在,为了保证较高的精度,每一类的样本数量应该是特征数的 10倍到 100倍,这意味着样本量必须增加到成千上万个,而现实中要获得这么多的可靠样本是非常困难的,即使能够获得这么多样本,数据计算的时间和空间复杂度也是让人难以承受的。目前常用

收稿日期: 2005-04-26;**修订日期**: 2006-02-23
基金项目:国家高技术研究发展计划 (编号: 2002AA130020),科技部政府间科技合作项目 (编号: CHN-24/2004)。
作者简介:王圆圆 (1981—),女, 2003年毕业于北京师范大学资源与环境科学系,现为该校资源学院在读博士生。主要从事高光谱遥感研究。已发表论文 2篇。 E-mail: wangyuanyuan@ires.cn。

的解决途径有两种,改进分类算法和有效的特征(波段)处理。对分类算法的改进就是要使其在样本量小特征多的情况下仍可以得到好的结果,如支持向量机(Support Vector Machine)分类方法就是基于小样本统计学理论发展起来的,对高维空间有很好的推广能力,一些学者将其引入到高光谱数据分类中,获得了很好的效果[1,2]。特征处理主要包括特征选择和特征抽取,特征选择就是从所有波段中只选用一部分波段,去掉那些与分析目标无关的或是冗余的波段[3,4],特征抽取就是用部分或全部波段通过某种映射方式构造新的对目标可以更好解释的特征变量[5],如 PCA(Principle Component Analysis)和 MNF(Minimum Noise Fraction)等,此外还有不是很常见的特征块方法,它是通过对相邻波段的综合处理获得新的特征[6]。有效的特征处理可以在减少特征个数的同时尽量保持足够多的相关信息,这样不仅对样本量的要求降低了,也可以使分类算法处理速度更快,结果更简单,甚至更准确。

决策树学习(Decision Tree DT)算法作为一种数据挖掘方法,具有很多其他方法所没有的优点[7],如训练速度快,执行快;对数据分布形式不做假设,可以获得非线性的映射;非黑箱式操作,可以形成易于人们理解的规则;具有内置的特征选择能力。因此,不少遥感领域专家开始对决策树进行研究,但这些研究多是直接利用决策树算法对多光谱数据进行分类,或将其与别的分类算法相比较[8-12],较少涉及高光谱数据,更少有对决策树比较深入的研究,鉴于此,本文将利用高光谱数据研究两个问题:

(1)决策树一般被视为分类算法,其实它也可以做特征选择,与别的特征选择方法相比,它的效果怎么样。

(2)决策树有内置的特征选择能力,数据预处理中的特征选择步骤对决策树分类是否必要以及有何影响。

2 特征选择和决策树算法介绍

在数据挖掘概念出现以前,特征选择就已在机器学习和模式识别领域得到广泛关注,其定义是:选择一个特征子集并使它对目标问题的说明是足够而且是必须的[13],自动化的特征选择常常是海量高维数据快速有效处理的重要一环。一个特征选择方法由三个部分组成[14]:特征评价准则,特征搜索(产生)方式和停止搜索准则。特征选择的目的,其实就是找出使特征评价指标值最大(或最小)时的特征子集,读者可以参阅文献[14],以了解更多有关特征选择的知识。本文在以前出现的特征选择方法的基础上设计了三种方法(表 1),这三种方法的优点在于简单、速度快,而且适合分类问题。SEP和MDLM方法都是从空集开始,每次加入一个使评价指标增长最大的特征,直到满足停止准则,RELIEF方法是先获得每个特征的可分性权重,按权重从大到小的顺序依次选入特征,并且使得任何两个特征之间的相关系数都小于某阈值(本文经过试验一些后,认为 0.95比较合适)。

表 1 三种特征选择方法的描述

Table 1 Description of 3 designed feature selection methods

名称	评价准则	搜索方法	停止准则
SEP	可分性(用所有两类组合的 J-M距离之和表示)	顺序前向选择	可分性增加百分比 <0.01%
MDLM	信息论中的最小描述长度准则[14]	顺序前向选择	描述长度增加
RELIEF	可分性[14]	排序结合顺序前向选择	预先定义的最优的特征子集已经找到了

决策树学习算法是现在数据挖掘领域中最流行的算法之一,常用的算法有 ID3[15], C4.5[16]和 CART[17]等。决策树的工作过程,其实就是找出分类能力最好的属性变量,把数据分成多个子集,每个子集再用分类能力最好的属性进行划分,如此迭代一直进行到所有子集仅包含同一类型或子集包含的样本数小于某阈值。决策树特征选择的结果就是综合那些在每个子集里被评价为是分类能力最好的属性变量。这样的选择特征方法可能存在以下不足,首先,在生成树的过程中,随着深度的增加,到达节点的样本量会迅速减少,基于这些不充足样本的特征选择可能导致错误的结论;其次,决策树选择特征的机制和别的方法是不一样的,它不是搜索使评价函数极值化的特征组合,而只是在分裂产生的样本子集中找寻最好的分类属性,得到的特征选择结果其实是局部最优的个体组合[18];最后,决策树中出现的特征是有等级的,浅节点处的特征比深节点处的特征更重要,如果直接综合在每个节点出现的特

征作为特征选择结果,就完全忽略了这种等级性,有可能使特征子集的潜在功效达不到发挥。

3　数据及方法

3.1　数据及处理

本文选用的是 OMIS高光谱数据,此数据是 2001年 5月 9日在北京顺义地区获取的,共有 128个波段,其中 1—64波段为可见 近红外波段, 65—96波段为短波红外, 97—104为中红外, 105—112为热红外, 113—128为短波红外。去除严重受水汽吸收干扰的中红外波段和反应地物热辐射信息的热红外波段,剩下 112个波段。去除热红外波段的原因有两个,一是很多的手持光谱仪都没有这个波段,今后的研究主要针对地面数据,二是获得的 OMIS用户文件里没有提供这个波段区间相应的偏置值(Offset)和增益值(Gain)。对数据作了如下简单的预处理,首先是辐射定标,将 DN值转化为传感器处的辐射亮度值,然后是去除大气影响,由于缺少数据获取时大气状况信息,所以就采用了简单的 IARR(Internal Average Relative Reflectance)方法将辐射亮度值归一化。最后选了一个大小为 196行 381列的实验区。观察实验区,通过目视解译,选了 8种土地覆盖类别:耕地、浇过水的耕地、裸地、有少量植被的裸地、果园、水体、菜地和建筑用地,各类别的样本量分别为: 328, 254, 336, 362, 318, 223, 197和 462,其中菜地和建筑用地的光谱特征比较混杂,纹理破碎,其他类型的光谱特征比较单一,纹理均匀。

3.2　研究方法

分层随机选取训练样本,第一份训练样本包含每个类别的 10%的样本,以后下一份训练样本总在上一份训练样本的基础上另外加上分层随机选入的每个类别的 10%的样本,这样可以得到大小逐渐规律增加的 10份样本(10%, 20%,…, 100%)。对每份训练样本实施特征选择和决策树算法(本文选用的是现在最常用的 C4.5决策树算法),一方面将决策树看成一种特征选择方法,与本文选用的其他特征选择方法做出比较分析,并采用最大似然法(ML)、后向传播神经网络法(BPNN)和最邻近法(1-NN)对特征选择后的数据进行分类,考察几种特征选择方法的效果;另一方面将决策树看成一种分类算法,来研究特征选择对其结果的影响,以及此种影响与样本数量之间的关系,影响主要考虑两方面:(1)决策树的精度,以十折交叉验证精度(10-Fold Cross Validation Accuracy)衡量,因为在训练样本大小不同的情况下,适宜的检验样本的大小是变化的,而且总样本数量也较少,所以就没有留出一个独立的检验样本集;(2)复杂度,以树的总节点个数衡量。

4　结果与分析

4.1　特征选择结果

表 2显示了 4种不同方法(SEP, MDLM, RELIEF, DT)的特征选择结果,从中可以发现:

(1)不同方法选择的特征子集差异很大。为了定量化描述两种结果的相似性,本文设计了相似性指数 $R=\frac{CARD(A\cap B)}{CARD(A\cup B)}$,其中 CARD(A∩B)表示两种方法结果(即 A和 B)的交集所包含的元素个数,CARD(A∪B)表示两种方法结果的并集所包含的元素个数,R的含义就是两种方法的特征选择结果中相同波段所占的比例。对每一份样本都计算两种方法特征选择结果的相似度,再对 10份样本做平均即可得到表 3中的结果。从中可以看出,结果最相近的是 MDLM和 SEP, RELIEF和其他方法的结果差别都很大。之所以不同方法的特征选择结果很不一样,主要是由于评价指标不同造成的。

(2)从表 2中可以看出,随着样本量的增加, SEP, MDLM, DT三种方法选择出的特征个数也逐渐增加,其中 DT的增加最明显。当样本量较大时, SEP, MDLM, RELIEF三种方法的特征选择结果趋于稳定,而 DT的特征选择结果仍有较大的波动,显示了该算法的不稳定性。

(3)为了进一步分析 4种方法特征选择的结果,本文作了如下的操作:对每种方法综合 10份样本的特征选择结果,统计出被选中频次大于或等于 5的特征,然后再确定这些特征所属的波谱区间,结果见表 4。从中可以看出 SEP和 RELIEF方法选出的特征主要都位于短波红外(尤其是 RELIEF方法),MDLM和 DT方法选出的特征在 7个波谱区间上分布的比较均匀,在比较重要的红谷波段区域, 4种方法中只有 DT选出了一个特征,由此可以认为 DT的特征选择效果是不错的。

(4)最大似然法(ML)是一种常用的统计分类器。图 1是经过 4种方法选择特征后,采取 ML分类得到的 10折交叉验证精度(由于当用全部特征时, ML中要计算的方差协方差矩阵近似奇异,所以

表 2 4种特征选择方法的结果

Table 2 Feature selection results of 4 different methods

方法	选出的波段
SEP	20, 22, 27, 51, 76, 115, 116
	13, 19, 22, 25, 34, 48, 76, 114, 125, 128
	6, 13, 18, 22, 29, 48, 76, 113, 117, 118, 123, 125
	4, 13, 23, 29, 35, 47, 71, 114, 115, 116, 119, 124, 126, 127
	4, 13, 18, 23, 25, 28, 41, 48, 76, 114, 115, 116, 119, 123, 124, 127
	4, 10, 13, 18, 22, 27, 38, 47, 54, 76, 115, 116, 119, 122, 124, 127, 128
	3, 9, 17, 23, 27, 31, 38, 47, 51, 76, 115, 116, 119, 121, 123, 124, 127
	3, 5, 11, 20, 23, 29, 34, 38, 47, 51, 76, 115, 116, 117, 119, 121, 123, 124, 127
	3, 11, 20, 22, 23, 29, 33, 38, 45, 47, 51, 76, 114, 117, 119, 121, 123, 125, 127
	3, 5, 16, 21, 22, 28, 34, 38, 47, 51, 76, 115, 116, 117, 119, 123, 124, 125, 127, 128
MDLM	17, 24, 36, 117, 125
	20, 22, 24, 30, 114, 120, 125
	6, 19, 22, 24, 32, 47, 114, 120, 125
	11, 19, 22, 25, 28, 47, 114, 117, 120, 125
	11, 20, 22, 24, 28, 47, 113, 115, 117, 125, 127
	6, 13, 20, 23, 25, 31, 47, 114, 117, 122, 125, 127
	6, 11, 18, 22, 24, 28, 47, 115, 117, 120, 123, 125, 127
	6, 11, 20, 22, 24, 28, 47, 113, 115, 117, 123, 125, 127
	6, 11, 20, 22, 24, 28, 47, 113, 115, 117, 119, 123, 125, 127
	6, 11, 20, 23, 25, 28, 47, 113, 115, 117, 119, 123, 125, 127
RELIEF	1, 4, 11, 23, 47, 62, 63, 64, 65, 66, 73, 81, 82, 83, 84, 85, 86, 87, 88, 91, 122, 127
	1, 3, 11, 23, 47, 62, 63, 64, 65, 73, 81, 82, 83, 84, 85, 86, 87, 88, 91, 122, 127
	1, 2, 11, 24, 56, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 122, 127
	1, 2, 14, 47, 62, 63, 64, 65, 75, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 127
	1, 4, 17, 47, 62, 63, 64, 65, 75, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 127
	1, 4, 17, 47, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 93, 127
	1, 4, 10, 19, 47, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 93, 127
	1, 4, 10, 19, 47, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 127
	1, 4, 10, 19, 47, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 120, 125
	1, 4, 10, 19, 47, 62, 63, 64, 65, 68, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 120, 121
DT	2, 3, 6, 55, 58, 59, 96
	9, 16, 36, 57, 58, 59, 61, 73, 117
	3, 10, 13, 22, 47, 55, 64, 71, 72, 123
	3, 4, 11, 12, 13, 15, 54, 55, 58, 61, 70, 76, 122
	4, 11, 12, 13, 43, 53, 55, 58, 63, 76, 128
	2, 3, 4, 11, 12, 15, 18, 55, 58, 63, 72, 76, 122, 123
	2, 3, 4, 11, 12, 15, 18, 41, 51, 52, 54, 58, 63, 69, 72, 76, 93, 122
	3, 4, 5, 6, 11, 12, 15, 18, 20, 28, 41, 51, 55, 58, 63, 69, 72, 76, 121, 122
	3, 4, 5, 6, 9, 12, 15, 18, 24, 30, 41, 47, 51, 54, 55, 58, 63, 72, 76, 89, 121, 122, 123
	1, 3, 4, 5, 9, 11, 15, 16, 18, 27, 30, 33, 47, 48, 51, 53, 55, 58, 63, 72, 76, 95, 121, 122, 123, 127, 128

注:其中用斜体标出的是在特征选择后,用决策树分类时所用到的特征(由于决策树本身具有特征选择的能力,输入的特征可能不会全都被选择)。

表 3　不同方法特征选择结果的相似性

Table 3　Similarity of feature selection results for different method-pair

	MDLM/%	RELIEF/%	DT/%
SEP	18.73	3.81	12.27
MDLM		3.99	6.87
RELIEF			5.21

表 4　4种方法在不同波谱区间选入的特征个数

Table 4　The number of features selected at different spectrum range by four methods

波谱区间	相应波谱区间中的特征个数			
	SEP	MDLM	RELIEF	DT
蓝谷	0	1	2	2
绿峰	0	0	0	0
黄边	1	1	0	3
红谷	0	0	0	1
红边	2	3	0	0
近红反射平台	3	2	4	4
短波红外	7	3	14	3

结果不予考虑)。从图 1中可以看出,经过 SEP和 MDLM方法选择特征后,ML分类精度最高而且稳定,这可能主要是因为 SEP和 MDLM的特征评价函数与 ML分类函数相近的缘故,SEP中用来度量可分性的 Jeffries-Matusita距离和 ML中用到的马氏距离很相近,最小描述长度函数和 ML函数也是可以相互转化的[19]。DT特征选择的效果不如 SEP和 MDLM,分类精度偏低,而且随样本变化的波动较大。基于 RELIEF特征选择的 ML分类精度最低。

(5)图 2是经过 4种方法选择特征后,采取后向反馈神经网络(BPNN)分类得到的 10折交叉验证精度(由于利用全部特征时的网络过于复杂,暂不考虑)。网络结构的参数对精度有很大影响,但这些参数的选择又缺乏理论依据,本文则凭经验对网络按如下标准设计:网络分为三层,输入层节点个数为输入特征的个数,输出层 8个节点(对应 8个类别),中间层节点数为输入层和输出层的节点数之

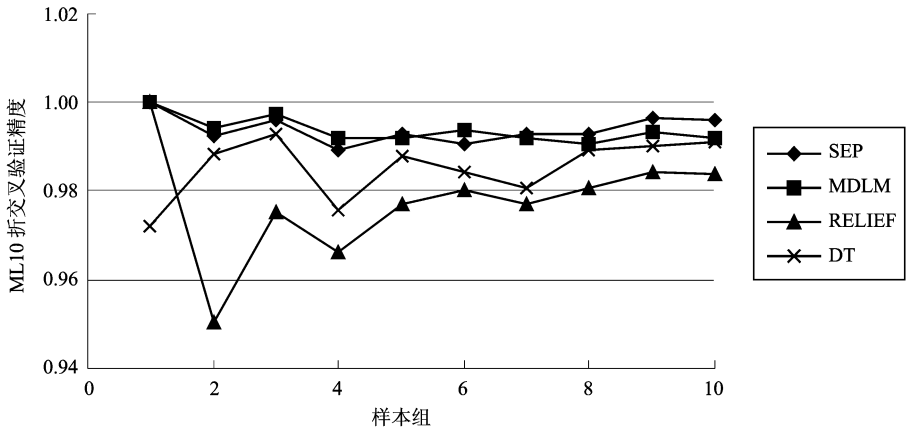

图 1　特征选择后 ML分类的 10折交叉验证精度随样本容量增加的变化趋势

Fig 1　The change trend of 10-fold cross validation accuracy of ML after feature selection through 4 methods

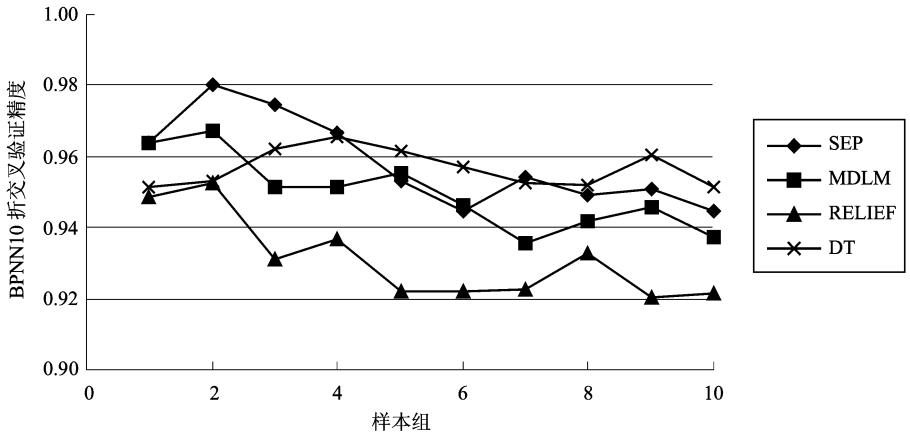

图 2　特征选择后 BPNN分类的 10折交叉验证精度随样本容量增加的变化趋势

Fig 2　The change trend of 10-fold cross validation accuracy of BPNN after feature selection through 4 methods

和,学习率为 0.01,动量项为 0.9,最大迭代次数 2000。由于没有优化设置网络参数,故而得不到很高精度的结果(比 ML获得的精度低),但由于只是特征选择之间的比较,所以不同分类方法的精度差异可以暂时忽略。从图 2中可以看出,基于 DT和 SEP特征选择的 BPNN分类效果比较好,MDLM方法次之,RELIEF方法最差,而且 DT方法获得的结果在不同大小样本的情况下都维持在较高的水平上,这也体现了 DT的确可以作为一种有效的特征选择方法。

(6)图 3是经过 4种方法选择特征后,采取最邻近法(1-NN)分类得到的 10折交叉验证精度,经过 SEP,MDLM和 DT三种方法选择特征后,1-NN的精度在各种大小样本下都维持在非常高的水平上,其中 MDLM的结果起伏稍大,SEP和 DT的结果非常相近且起伏较小,经 RELIEF特征选择后获得的分类精度明显非常低而且不稳定。

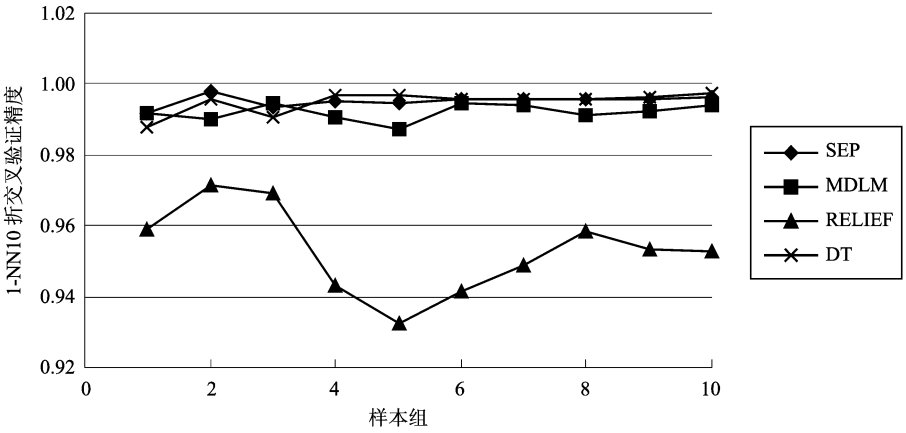

图 3 特征选择后 1-NN分类的 10折交叉验证精度随样本容量增加的变化趋势

Fig 3 The change trend of 10-fold cross validation accuracy of 1-NN classification after feature selection through 4 methods

4.2 决策树分类结果的差异

4.2.1 决策树的精度分析

(1)随着样本数量的增加,4种情况下生成的决策树精度均呈现波动式的增加,而且彼此之间差别逐渐减少(图 4(a))。由此可见,特征选择对精度的影响主要发生在样本量比较少的时候。

(2)有效的特征选择可以一定程度上提高精度,但提高幅度不明显,本文中,三种特征选择方法对精度的提高平均为 0.35%,在样本量最少的时候,精度的提高幅度最大,为 1.71%。SEP特征选择对精度的提高效果最好,MDLM方法其次,RELIEF方法最差,事实上它使得 8个样本组的精度都降低,如果不考虑 RELIEF的结果,SEP和 MDLM对精度提高的幅度平均为 0.59%。

4.2.2 决策树的结构分析

(1)三种经过特征选择生成的决策树用到的特征个数较少,彼此相近,并且随样本量的增加缓慢增长,而直接生成的决策树用到的特征个数会随样本量的增大而快速增长(图 4(b))。

(2)决策树的复杂度(以树的节点数衡量)基本上是随着样本数量增加而上升(图 4(c))。没有特征选择时,决策树的节点个数呈现一种线性增加趋势,但有了特征选择后,节点个数变化的波动性更大,如从第 8份到第 9份样本,MDLM+DT和 SEP+DT方法得到的决策树节点个数都下降了,这可能是因为特征选择相当于把数据投影到一个低维子空间,生成决策树时受到子空间和样本的双重变化影响。

(3)有了特征选择后,树的节点数平均增加了 18.61%,增幅最大的是采用了 RELIEF特征选择(平均增加了 25.33%),其次是 MDLM(18.7%)和 SEP(11.8%),在样本量少的时候,4种情况下生成的决策树的节点个数差别比较小。

5 结论与讨论

本文利用高光谱数据,研究了决策树特征选择的特点和特征选择对决策树分类的影响,主要结论有以下几点:(1)决策树是一种比较好的特征选择方法,而且更适合在样本量比较少的情况下使用,因为样本多时,选择出的特征数量过多,而且不如其他方法获得的结果稳定;(2)特征选择对决策树分类

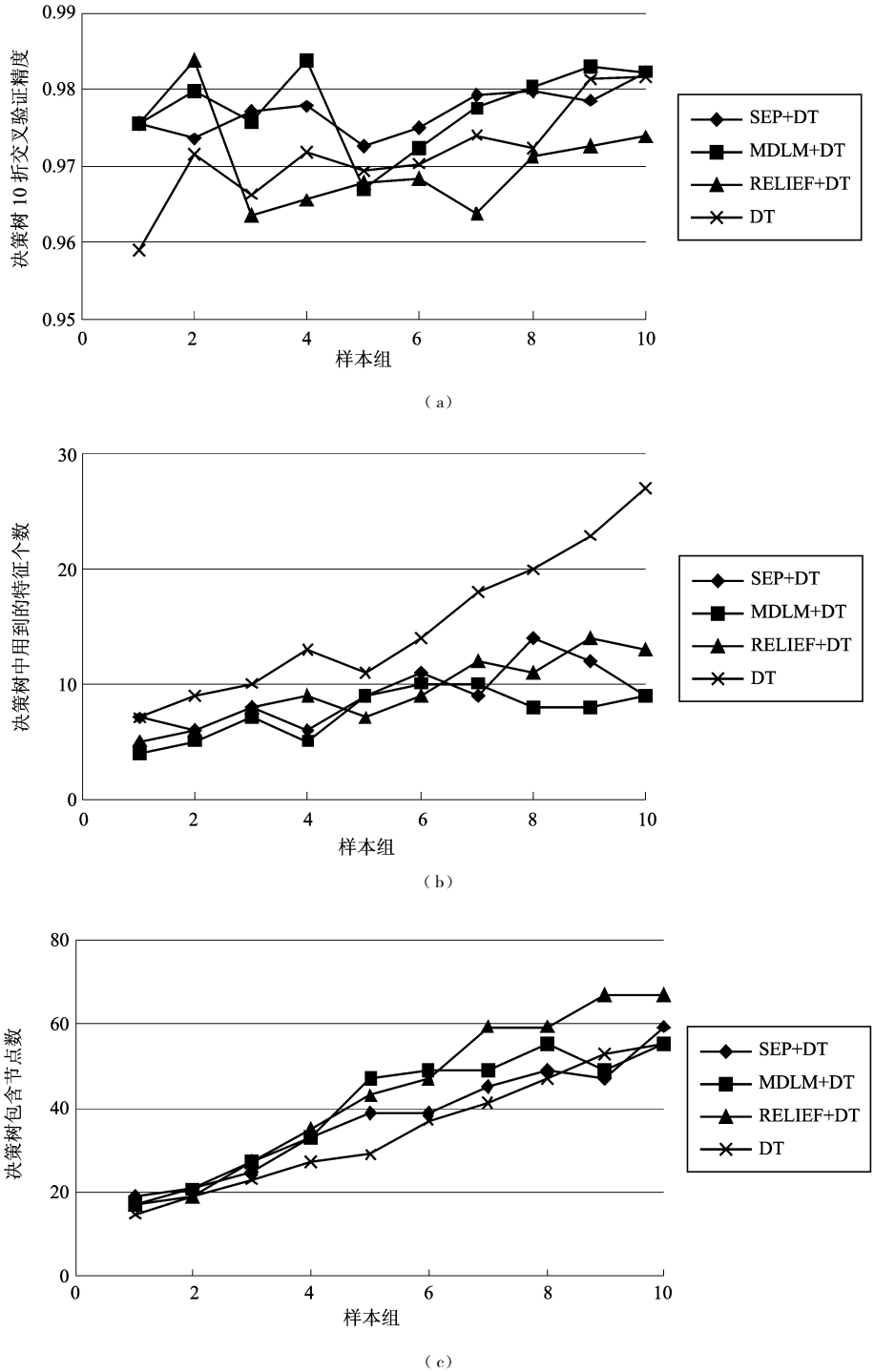

图 4　4种情况下决策树精度(a)、用到的特征个数(b)、节点数(c)随样本容量的变化

Fig 4　The change trend of classification accuracy (a), the number of features used (b) and the number of tree nodes(c) with sample increase under 4 different conditions

结果的影响主要在树的结构上,经过特征选择后生成的决策树用到的特征个数平均减少了 43.36%,节点数平均增加了 18.61%;(3)特征选择对决策树分类精度的影响较小,平均提高幅度为 0.35%;(4)在小样本量情况下,特征选择对于高光谱数据决策树分类是很有必要的,此时决策树的精度提高幅度最大,而树的节点数基本没有增加。

为了提高分类精度,选择或设计比较好的特征选择方法是很重要的,如本文设计的 RELIEF方法效果就不太好,它的特征选择结果与其他方法的结果相差

很大,而且获得的分类精度低而不稳,这是由本文设计的 RELIEF方法特点决定的,它本质上是对每个特征单独地进行分析,然后排序,选出排序位次高而且彼此相关小的个体组合,不像 MDLM 和 SEP那样有考虑特征之间的相互作用对评价函数的影响。

一般认为,在分类前进行特征选择常常会获得更简单更易于理解的分类模式,但本文中却得出了特征选择使决策树节点数增加的结果,其原因在于高光谱数据全是数值型变量, C4.5决策树算法可以使一个数值型变量出现在树中一条路径的不同位置,采用不同的阈值构成节点分裂标准,使树的大小受到变量个数和变量重复出现次数两个因素的影响,经计算表明,本文中有特征选择的决策树,变量平均出现次数为 2.06,而直接生成的决策树,变量平均出现次数为 1.1,变量重复出现次数的增加导致了决策树更加复杂。特征选择对决策树的影响还体现在分类精度上,虽然决策树用到的特征个数减少了,但节点数增加,分类边界变复杂,因此提高了分类精度,当然这其中也有无关和冗余特征被去除的缘故。精度虽有提高,但幅度小,这可能有两方面原因,首先是特征选择对高光谱数据决策树分类的影响的确是不大,其次是用到的10折交叉验证精度常常比用独立的检验样本计算出来的精度要高,使不同情况下的计算结果更接近。

高光谱遥感即将进入航天时代,成为遥感应用的主要信息源之一,对高光谱数据挖掘亦将成为高光谱应用的关键环节。决策树算法作为一种重要的数据挖掘技术,具有很多优良特性,在构建分类模型的同时选择有用特征,尤其应该在高光谱遥感领域得到充分深入的研究。今后的研究计划包括考虑特征选择对多变量决策树的影响,最好能设计出一种适合决策树算法的特征选择方法,另外,目前已有不少学者提出了专门针对全数值型变量的决策树的生成算法,如 Berzal等人设计的 Multi-way Decision Tree,就是通过变量局部离散化来获得结构更简单的决策树[20],这些新型的决策树在高光谱遥感领域的适用性如何,也亟需得到研究论证。

致 谢 感谢 Ian H. Witten, Eibe Frank在网上免费提供的 weka-3-4软件,帮助作者利用决策树算法做研究。

参 考 文 献 (References)

[1] Melgani F, Bruzzone L. Classification of Hyperspectral Remote Sensing Images with Support Vector Machines[J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, **42**(8): 1778—1789.

[2] Pal M, Mather P M. Assessment of the Effectiveness of Support Vector Machines for Hyperspectral Data[J]. Future Generation Computer Systems, 2004, **20**: 1215—1225.

[3] Thenkabail P S, Enclona E A, Mark S A. Accuracy Assessment of Hyperspectral Waveband Performance for Vegetation Analysis Application[J]. Remote Sensing of Environment, 2004, **91**: 354—376.

[4] Serpico S B, Bruzzone L. A New Search Algorithm for Feature Selection in Hyperspectral Remote Sensing Images[J]. IEEE Transaction on Geoscience and Remote Sensing, 2001, **39**: 1360—1367.

[5] Özkan C, Erbek F S. Comparing Feature Extraction Techniques for Urban Land-use Classification[J]. International Journal of Remote Sensing, 2005, **26**(4): 747—757.

[6] Jia X, Richards J A. Efficient Maximum Likelihood Classification for Imaging Spectrometer Data Sets[J]. IEEE Transactions on Geosciences and Remote Sensing, 1994, **32**: 274—281.

[7] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementation[M]. Beijing: China Machine Press, 2003.

[8] Friedl M A, Brodley C E. Decision Tree Classification of Land Cover from Remotely Sensed Data[J]. Remote Sensing of Environment, 1997, **61**: 399—409.

[9] Brown de Colstoun E C, Story M H, Thompson C, et al. National Park Vegetation Mapping Using Multitemporal Landsat 7 Data and a Decision Tree Classifier[J]. Remote Sensing of Environment, 2003, **85**: 316—327.

[10] Wessels K J, De Fries R S, Dempewolf J, et al. Mapping Regional Land Cover with MODIS Data for Biological Conservation: Examples from the Greater Yellowstone Ecosystem USA and Para State Brazil[J]. Remote Sensing of Environment, 2004, **92**: 67—83.

[11] Pal M, Mather P M. An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification[J]. Remote Sensing of Environment, 2003, **86**: 554—565.

[12] Goel P K, Prasher S O, Patel R M, et al. Classification of Hyperspectral Data by Decision Trees and Artificial Neural Networks to Identify Weed Stress and Nitrogen Status of Corn[J]. Computers and Electronics in Agriculture, 2003, **39**: 67—93.

[13] Kira K, Rendell L. A Practical Approach to Feature Selection[A]. Proceedings of the 9th Int Conf On Machine Learning[C]. 1992.

[14] Dash M, Liu H. Feature Selection for Classification[J]. Intelligent Data Analysis, 1997, **1**: 131—156.

[15] Quinlan J. Introduction of Decision Trees[J]. Machine Learning, 1986, (1): 81—106.

[16] Quinlan J. C4.5: Programs for Machine Learning[M]. California: Morgan Kaufmann, 1993.

[17] Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees[M]. Belmont: Wadsworth, 1984.

[18] Pemer P, Apte C. Empirical Evaluation of Feature Subset Selection Based on a Real World Data Set[J]. Engineering Application of Artificial Intelligence, 2004, **17**: 285—288.

[19] Tom M. Mitchell. Machine Learning[M]. Beijing: China Machine Press, 2003.

[20] Berzal F. Building Multi-way Decision Tress with Numerical Attributes[J]. Information Sciences, 2004, **165**: 73—90.