一种基于特征选择的不完整数据分类方法

陈景年[1,2],黄厚宽[1],田凤占[1],薛小平[3]
CHEN Jing-nian[1,2],HUANG Hou-kuan[1],TIAN Feng-zhan[1],XUE Xiao-ping[3]

1.北京交通大学 计算机与信息技术学院,北京 100044
2.山东财政学院 信息与计算科学系,济南 250014
3.北京交通大学 电子信息工程学院,北京 100044
1.School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China
2.Department of Information and Computing Science,Shandong University of Finance,Ji'nan 250014,China
3.School of Electronics and Information Engineering,Beijing Jiaotong University,Beijing 100044,China
E-mail:jnchen06@163.com

CHEN Jing-nian,HUANG Hou-kuan,TIAN Feng-zhan,et al.Classification method for incomplete data based on feature selection.Computer Engineering and Applications,2007,43(31):23-24.

Abstract: Feature selection is an important policy to simplify data,reduce necessary memory and improve the accuracy and efficiency of classification.Data are often incomplete because of various kinds of reasons.For incomplete data,methods of constructing selective classifiers can also reduce necessary memory and improve the accuracy and efficiency of classification.So developing selective classifiers for incomplete data is an important problem.In this paper a method of constructing selective Bayes classifiers from incomplete data is presented.Experiments on twelve benchmark incomplete data sets show that not only is the classification accuracy of the selective classifier proposed much higher than that of the very efficient RBC classifier,but also its performance is more robust.

Key words: feature selection;classification;Bayesian method;incomplete data

摘 要:特征选择(也称作属性选择)是简化数据表达形式,降低存储要求,提高分类精度和效率的重要途径。实际中遇到的大量的数据集包含着不完整数据。对于不完整数据,构造选择性分类器同样也可以降低存储要求,提高分类精度和效率。因此,对用于不完整数据的选择性分类器的研究是一项重要的研究课题。有鉴于此,提出了一种用于不完整数据的选择性贝叶斯分类器。在12个标准的不完整数据集上的实验结果表明,给出的选择性分类器不仅分类准确率显著高于非常有效地用于不完整数据的RBC分类器,而且分类性能更加稳定。

关键词:特征选择;分类;贝叶斯方法;不完整数据
文章编号:1002-8331(2007)31-0023-02 **文献标识码:**A **中图分类号:**TP391

1 前言

在分类过程中,并非每个特征(也称作属性)都对分类起积极作用,甚至有的属性甚至会降低分类效果。因此,产生了许多以提高分类效果为目的的特征选择方法。大量的基于属性选择的分类器应运而生。然而这些选择性分类器大都是针对完整数据的。由于各种原因,在实际应用中遇到的大量的数据集都包含着不完整数据。对于不完整数据,在属性选择的基础上构造分类器同样也可以降低存储要求,提高分类精度和效率。有时甚至效果会比完整数据的情况更显著。因此,对用于不完整数据的选择性分类器的研究是一项重要的研究课题。由于不完整数据的复杂性,对不完整数据的处理同时又是一个比较困难的问题,至今仍然没有十分有效的用于不完整数据的选择性分类器。本文提出的用于不完整数据的选择性分类器,是一种以包装法(wrapper)[1]为主要思想的选择性分类器,并且基于一种对不完整数据进行分类的贝叶斯分类器RBC [2](对RBC后面有介绍)。

本文内容是这样安排的:在第2章中,简要介绍对不完整数据进行分类的非常高效的RBC分类器,以及包装法的主要思想。第3章介绍本文提出的用于不完整数据的选择性分类器。在第4章,将提出的算法在12个不完整数据集上进行实验,并对实验结果加以讨论。第5章对本文进行总结,并对下一步的工作加以展望。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60503017,No.60673089)。
作者简介:陈景年(1970-),博士生,副教授,CCF学生会员,主要研究领域为模式识别、机器学习、数据挖掘;黄厚宽(1940-),教授,博士生导师,CCF高级会员,主要研究领域为人工智能、模式识别、数据仓库、数据挖掘以及多智能体系统;田凤占,博士,副教授,CCF会员,主要研究领域为机器学习、数据挖掘;薛小平,博士生,副教授,主要研究领域为RFID网络、传感器。

2 RBC 算法

RBC(Robust Bayes Classifier)[2]是一种从不完整数据构建的贝叶斯分类器。RBC 是朴素贝叶斯分类器的扩展。RBC 的训练过程是先在所给的不完整数据集上计算有关的不完整实例的频数,然后利用这些频数将各个属性变量的类条件概率分布以及类变量的边缘分布进行区间界定。分类过程是利用上述区间求出在给定新实例的条件下,类变量后验概率所属的区间,并通过给区间打分,将新实例分到最高分值关联的类中。这种方法具有很高的分类效率。

与朴素贝叶斯分类器相似,RBC 也是假定在给定类变量时,各个属性变量之间是相互独立的。

3 用于不完整数据的选择性分类器

利用 RBC 就可以构建用于不完整数据的选择性贝叶斯分类器 Selective Bayes Classifier for Incomplete Data,SBCID)。在此过程中,采用了搜索效果好而复杂度相对较低的最优优先(best first search)前向搜索方法[3]对属性空间进行搜索。

记 $A=\{a_1,a_2,\cdots,a_N\}$为整个属性集合,N为 A 中属性的个数。Q 为一个队列,用来存放曾经是最优的属性子集及其对应的分类精度。S_b 为当前最优属性子集。$f(S)$表示 RBC 在属性子集 S 上的分类精度。阈值 T 为用来控制搜索过程是否停止的参数,即如果连续 T 次对 Q 的头结点进行扩展都没有使当前最高分类精度改善,则搜索过程结束。

算法 SBCID 可描述如下:

(1)设置 T,令整数 t=0,令属性 $a_s=\arg\max_{1\leqslant i\leqslant N}\{f(\{a_i\})\}$,当前最高分类精度 $f_{\max}=f(\{a_s\})$,将属性子集$\{a_s\}$作为一个结点加入到队列 Q 中。

(2)当 $t<T$ 时执行步骤(3)~(5),否则,执行步骤(6)。

(3)取出 Q 的头结点 S_h(为一属性子集),令 *added*=false(*added* 用来标志在对 Q 的头结点的扩展中,是否向 Q 中加入新的结点)。对每一属性 $a\in A-S_h$,如果 $S_h\cup\{a\}$没有被评价过,而且 $f(S_h\cup\{a\})>f_{\max}$,那么,令 *added*=true,$S_b=S_h\cup\{a\}$,$f_{\max}=f(S_h\cup\{a\})$,以及 t=0,并且将 S_b 作为一个新结点加入到队列 Q 中。

(4)如果 *added*=false,那么 $t\leftarrow t+1$。

(5)转到步骤(2)继续执行。

(6)在最终的属性子集 S_b 上构建 RBC 分类器。

为什么会用 RBC 分类器来构造基于特征选择的分类器呢?有两个方面的原因:

首先,在主要的用于不完整数据的分类算法中,象 EM 算法、Gibs 抽样等,虽然有时能得到较好的分类效果,但它们的计算复杂度一般都很高,难以用于构建基于包装法的选择性分类器。另外,朴素贝叶斯分类器也可对不完整数据进行分类,而且分类的效率很高。它对不完整数据主要采取两种简单处理方法:删除包含不完整数据的实例和给缺失数据指定一个虚拟值。然而这两种方法都可能会引起大的估计偏差。RBC 分类器不仅效率高,而且分类准确率也比朴素贝叶斯方法的要高,完全克服了上述方法的不足。因此,对构建基于包装法的选择性分类器来说,RBC 是一个理想的选择。

其次,前面已经提及,与朴素贝叶斯分类器相似,RBC 也是假定在给定类变量时,各个属性变量之间是相互独立的,而这一假定在实际中多数情况下不成立,这往往会使分类准确率降低。因此,通过特征选择,从原特征集合中去除冗余的特征,会提高 RBC 的分类准确率。在下面一章中的实验结果也证实了这一点。

4 实验结果及分析

为了验证所提出的算法的有效性,在 12 个不完整数据集上进行了实验。这 12 个数据集均来自 UCI 机器学习知识库[4]。表 1 对这 12 个数据集进行了描述,从上到下按照数据集合的实例个数从大到小顺序依次排列。数据集实例个数从最多 8 124 到最少 32 个,属性个数(在表 2 中列出)从最多 279 个到最少 10 个,分别分布在一个很宽的范围内。

表 1 对实验中的 12 个不完整数据集的描述

数据集	大小	类数	数据集	大小	类数
Mushroom	8 124	2	Vote	435	2
Annealing	798	5	Horse-colic	368	2
B.cancer	699	2	Audiology	200	2
Credit	690	2	Echocardiogram	132	2
Cylinder	512	2	Bridges	108	6
Arrhythmia	452	16	L.cancer	32	3

实验目的是,在每一个数据集上比较 RBC 与 SBCID 的分类准确率。同时考察通过属性选择使属性个数减少的程度。

整个实验是在 weka 系统[5]环境下,在内存为 1 GB,主频为 2.93 GHz 的 Pentium IV PC 机上运行的。在整个实验过程中,令 α=1(α 在 RBC 算法中用来确定先验信息,详情参见文献[2])。在属性选择过程中,参数 T 取 weka 系统中的默认值 T=5。通过 5 重交叉验证来评价每个候选的属性子集。

在每一个数据集上对 RBC 的分类准确率与 SBCID 的分类准确率进行比较时,分别进行 20 次 10 重交叉验证。表 2 列出了比较结果。其中分类准确率是 20 次的平均准确率。在每一个数据集上的较高的准确率,以粗体表示。

表 2 RBC 与 SBCID 的分类性能比较

数据集	属性数	被选属性数	SBCID	RBC
Mushroom	22	3	**99.67±0.05**	95.96±0.02
Annealing	38	8	91.62±0.10	**96.00±0.31**
B.cancer	10	9	**97.32±0.10**	97.14±0.15
Credit	15	7	**87.05±0.34**	86.12±0.44
Cylinder	39	8	**76.16±0.51**	71.32±0.56
Arrhythmia	279	11	**75.49±0.74**	72.85±0.75
Vote	16	3	**96.31±0.00**	90.31±0.19
Horse-colic	27	5	**87.98±0.38**	85.12±0.57
Audiology	70	12	**76.68±0.54**	67.81±0.68
Echocardiogram	12	3	97.26±0.43	**98.36±0.82**
Bridges	12	6	**65.90±1.11**	61.19±2.02
L.cancer	56	5	**79.52±3.11**	56.61±2.60

从表 2 可以看出,在 12 个数据集中,有 10 个数据集,SRBCID 的分类准确率明显高于 RBC 的分类准确率。尤其是在数据集 L.cancer 上,SRBCID 的分类准确率比 RBC 的分类准确率高出 22.91%。

为什么在数据集 L.cancer 上分类准确率会提高这么多呢?除了算法 SBCID 本身的作用外也与数据集 L.cancer 本身的特点有关。L.cancer 总共有 32 个实例,而属性个数却有 56 个之多。一般情况下,当实例个数相对于属性个数较少时,对各个属性变量的类条件概率估计以及对类变量的概率估计都会变得

(下转 38 页)