类相关性影响可变选择性贝叶斯分类器

程玉虎,仝瑶瑶,王雪松
(中国矿业大学信息与电气工程学院,江苏徐州 221116)

摘 要: 在最大相关最小冗余(mRMR)属性选择方法的基础上,通过设置一个调节因子来改变类别相关性在属性选择中的影响程度,解决mRMR方法易于引入冗余属性的问题,提出一种类相关性影响可变选择性贝叶斯分类器(CCRI SBC).为克服人为指定属性个数易于导致的分类结果随意性,采用贝叶斯信息准则来自动确定最优属性个数.为使CCRI SBC能够处理含有连续变量的数据集,提出等频类别依赖最大化离散化方法,具有分类准确率高和离散化时间短的优点.UCI数据集的实验结果表明,本文方法能够有效处理离散和连续高维数据的分类问题.

关键词: 选择性贝叶斯分类器; 属性选择; 最大相关最小冗余; 贝叶斯信息准则; 离散化
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2011) 07-1628-06

A Selective Bayesian Classifier Based on Change of Class Relevance Influence

CHENG Yu-hu, TONG Yao-yao, WANG Xue-song
(*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*)

Abstract: A selective Bayesian classifier based on change of class relevance influence (CCRI SBC) was proposed by introducing a regulator factor into an attribute selection method, namely maximum relevance and minimum redundancy (mRMR). The regulator factor was used to change the influence degree of class relevance on the attribute selection, which can avoid the existence of redundant attributes in mRMR. In addition, a Bayesian information criterion was used to determine the optimal number of attributes automatically, which can overcome the randomness of classification results that easily caused by the setting number of attributes manually. In order to further make the CCRI SBC is applicable for continuous data, a discretization method, i.e., equal frequency class attribute interdependent maximization was proposed, which has advantages of high classification correct rate and short discretization time. Experimental results on UCI datasets show that the proposed method can deal with the classification problem for discrete or continuous and high-dimensional data effectively.

Key words: selective Bayesian classifier; attribute selection; maximum relevance and minimum redundancy; Bayesian information criterion; discretization

1 引言

分类是机器学习和数据挖掘中一个重要的研究内容,分类就是生成一个分类模型或函数,然后用该模型把数据库中的数据项映射到某一给定类别中,从而预测出未知数据所属类别.朴素贝叶斯分类器(Naive Bayesian Classifier, NBC)是一种非常简单、有效的分类器,其性能可以与神经网络、决策树相媲美[1].但是,朴素贝叶斯分类器具有很强的限制条件,即条件独立性假设,这一假设同现实世界不相符合,限制了其在实际问题中的进一步应用.为此,国内外学者提出了各种改进方法.

选择性贝叶斯分类器(Selective Bayesian Classifier, SBC)是一种有效的改进方法,它通过属性选择的方法,选择出部分具有较好独立性关系的属性结点,然后在选择出的属性子集上构建朴素贝叶斯分类器,从而改善分类效果.属性选择的方法主要分为包装法和过滤法.Langly和Sage提出了基于包装法的选择性朴素贝叶斯分类器[2],利用包装法选择出部分具有较好属性独立关系的子集.但是,包装法依赖于某种或多种机器学习算法,虽然能够有效提高分类准确率,但是计算复杂性大、效率低,不能很好地应用在高维数据上[3].过滤法独立于机器学习算法,分类效果一般,但是计算代价小、效率高,能够很好地处理高维数据,比包装法有着更广泛的

收稿日期:2010-04-21;修回日期:2010-08-10
基金项目:国家自然科学基金(No.60804022,No.60974050,No.61072094);教育部新世纪优秀人才支持计划(No.NCET-08-0836);霍英东教育基金会青年教师基金(No.121066);江苏省自然科学基金(No.BK2008126)

适用性[3].在过滤方法中,基于互信息的属性选择是一种应用较为广泛的降维方法,已被人们成功地应用在各种分类问题中.在互信息的基础上,Yang 等提出了互信息最大化(max Mutual Information, maxMI)的属性选择方法[4].但是 maxMI 不考虑已选属性间的相关性,所以属性间的冗余性较大,难以获得较高的分类精度.为此,Peng 等提出了最大相关最小冗余(Maximum Relevance and Minimum Redundancy, mRMR)的属性选择方法[5],用二维互信息近似估计高维互信息,应用于贝叶斯分类器上有很好的分类效果.但是,由于 mRMR 方法是根据类别互信息和属性间互信息的差值来进行属性选择的,因此,通常会引入冗余属性.为此,在最大相关最小冗余属性选择方法的基础上,通过设置一个调节因子来改变类别相关性大小在属性选择中的影响程度,从而消除冗余属性以选择出更加符合朴素贝叶斯条件独立性假设的属性结点.对于以上属性选择方法无法确定最优属性个数的问题,提出一种基于贝叶斯信息准则的最优属性个数自动确定方法.另外,为使贝叶斯分类器能够处理连续数据,将类别属性依赖性最大化和等频离散化方法相结合,提出一种等频类别依赖最大化的方法对连续数据进行离散化处理.

2　选择性贝叶斯分类器

2.1　朴素贝叶斯分类器

朴素贝叶斯分类器建立在一个类条件独立性假设(朴素假设)基础之上:给定类结点(变量)后,各属性结点之间相互独立[6].朴素贝叶斯分类器的分类过程如下:设 c 表示类结点,c 有 q 个不同取值 $c_1,\cdots,c_k,\cdots,c_q$,即 q 个不同类别;$x=\{x_1',\cdots,x_l',\cdots,x_n'\}$ 是属性变量 $X=\{x_1,\cdots,x_l,\cdots,x_n\}$ 的具体取值.假设类结点的先验概率为 $p(c_k)$,由朴素假设,概率 $p(x\mid c_k)$ 的计算可以简化为:

$$p(x\mid c_k)=p(x_1',x_2',\cdots,x_n'\mid c_k)=\prod_{l=1}^{n}p(x_l'\mid c_k)\quad(1)$$

由贝叶斯定理,类结点的后验概率为:

$$p(c_k\mid x)=\frac{p(x\mid c_k)p(c_k)}{p(x)}\quad(2)$$

分类的主要目标是求在给定待分实例的条件下类别的后验概率,即选择后验概率最大的类别作为该样本的类别.$p(x)$ 对于所有的类为常数,$p(c_k)$ 为类的先验概率,于是计算的主要目标是求 $p(x\mid c_k)$,$p(x\mid c_k)$ 可由式(1)求出.

尽管朴素贝叶斯分类器结构简单,分类效果良好,但是在面对一些具有冗余属性的高维数据集时,其分类效果明显降低.选择性贝叶斯分类器可以通过属性选择的方法,选择出类相关性较大,冗余性较小的属性,从而降低数据维数,提高分类准确率.

2.2　互信息最大化选择性贝叶斯分类器

互信息最大化选择性贝叶斯分类器(maxMI SBC),使用 maxMI 的方法进行属性选择,即从原始的属性变量 $X=\{x_1,\cdots,x_l,\cdots,x_n\}$ 集合中选择出它的一个子集 $S=\{x_1,\cdots,x_i,\cdots,x_m\}$ 来构成一个新的属性子空间,其中 $m<n$.x_l 是 X 中的某一属性结点,I 用来表示结点间的互信息,则原始属性空间的互信息为:

$$I(X;c)=\sum_{l=1}^{n}I(x_l;c)\quad(3)$$

互信息最大化属性选择方法的基本思想是:选择出的属性子集应该尽可能多地提供关于类别的信息,即使 $I(S;c)$ 最大化[4].因此,maxMI 准则可表示为:

$$\max\ D(S,c)=\frac{1}{|S|}\sum_{x_i\in S}I(x_i;c)\quad(4)$$

式中,x_i 表示属性子集 S 中某一属性结点,$|S|$ 表示选择出的属性子集 S 中属性结点的个数,D 表示 m 个属性结点同类结点间互信息的均值.

由式(3)可以看出,要使 $I(S;c)$ 达到最大,只要选择 $I(x_l;c)$ 的值最大的前 m 个属性来构成集合 S 即可.

2.3　最大相关最小冗余选择性贝叶斯分类器

互信息最大化属性选择方法只考虑了属性与类结点间的相关性,而没有考虑属性间的关联,所以得到的属性结点中会包含大量的冗余结点.当两个属性结点间相关性很大时,我们称其中的一个属性结点为冗余结点,此时去掉这个冗余结点可以降低计算复杂度,同时对分类效果不会有太大的影响[5].基于 mRMR 方法的选择性贝叶斯分类器(mRMR SBC)将类别相关性最大同属性间冗余性最小这两种属性选择方法相结合,既考虑了类别的作用,也去掉了部分冗余属性.最大相关准则表示法同式(4),最小冗余准则表示方法如式(5).

$$\min\ R(S)=\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i,x_j)\quad(5)$$

式中,x_i 和 x_j 表示 S 中任意两个属性,R 表示选择出的属性间的独立性.将式(4)和(5)相结合,即可得到 mRMR 准则:

$$\max\ \varphi(D,R)=D-R\quad(6)$$

式中,φ 用来增量式地选择同类结点相关性较大,而同其他已选属性结点相关性较小的属性.

利用 mRMR 准则进行属性选择,通常采用递增方式逐步获取,假设已经获得 $m-1$ 个属性结点,则可以通过式(7)选择第 m 个属性结点.

$$\max_{x_j\in X-S_{m-1}}\left[I(x_j;c)-\frac{1}{m-1}\sum_{x_i\in S_{m-1}}I(x_j;x_i)\right]\quad(7)$$

式中,S_{m-1} 表示已经选择好的 $m-1$ 个属性结点集合.

3 类相关性影响可变选择性贝叶斯分类器

3.1 类相关性影响可变的属性选择方法

mRMR SBC 能够得到同朴素贝叶斯分类器的类条件独立性假设比较符合的属性结点.但是,在进行属性选择的过程中,mRMR 方法选择出的是类互信息同属性间互信息的差值最大的属性结点.这种方法会引入部分无关和冗余结点,即差值很大、类别相关性很小或属性间冗余性很大的结点.由于类结点同属性结点的相关性远大于属性结点之间的相关性,所以无关属性出现的可能性较小,主要存在的噪声结点是冗余结点.为此,本文在 mRMR 的基础上,通过加入调节因子来调节类相关性互信息的影响程度,提出一种类相关性影响可变(Change of Class Relevance Influence,CCRI)选择性贝叶斯分类器.利用 CCRI 方法来选择属性的计算方法如下:

$$\max_{x_j\in X-S_{m-1}}\left[\alpha I(x_j;c)-\frac{1}{m-1}\sum_{x_i\in S_{m-1}}I(x_j;x_i)\right] \quad (8)$$

式中,α 为调节因子.为了消除差值引入的冗余结点,一般设置 α 的取值范围为[0,1],相当于在差值不变的条件下降低类别相关性的影响程度.调节因子 α 可通过交叉验证的方法计算得到:以分类准确率为评价标准,在[0,1]内搜索得到全局最优参数.同时为了降低交叉验证方法的计算复杂度,在进行参数选择的过程中,α 在[0,1]取值范围内每隔 0.05 取一个值.在 α 值不同的情况下,如果选择出相同的属性,那么将得到相同的分类准确率,从而设置交叉验证的结果相同.

类相关性影响可变选择性贝叶斯分类器(CCRI SBC)的算法步骤:

步骤 1 初始化:将属性子集的初始值设为空集 $\varnothing$,即 $S\leftarrow\varnothing$,并确定 m 的取值;

步骤 2 确定 α 的取值:设置 $\alpha=[0,1]$,每隔 0.05 取一个值.使用交叉验证的方法得到最优参数 α;

步骤 3 确定最优属性子集 S:确定同类结点相关性最大的结点为 S 中第一个属性,使用式(8)依次选择出剩下的 $m-1$ 个属性;

步骤 4 在选择好的属性子集 S 上构建朴素贝叶斯分类器.

3.2 基于贝叶斯信息准则的属性个数确定

CCRI、maxMI 和 mRMR 方法能够进行指定数目的属性选择,但是这种人为指定属性个数的方式将导致最终的分类结果具有随意性.另外,通过多次实验尝试以确定属性个数的方法不仅费时耗力,而且最终得到的分类结果也不一定是最优的.

贝叶斯信息准则(Bayesian Information Criterion,BIC)通常被用来评估贝叶斯网络的结构模型,能够在很好的拟合数据和模型的简洁性之间达到某种折中.这里,给出一种利用贝叶斯信息准则来确定最优属性个数的方法:使用 BIC 测度判断是否在类结点和属性结点之间添加弧,如果添加弧可以增加 BIC 评分,则表示这个属性是有效的;反之,则删除该属性.Schwarz 给出了 BIC 测度的具体计算公式[7]:

$$Q_{\mathrm{BIC}}(B,T)=LL(B/T)-\frac{1}{2}\log N*\mathrm{Dim}(B) \quad (9)$$

式中,T 表示数据集,B 表示贝叶斯网络结构,$LL(B/T)$表示基于概率分布描述 T 所需要的比特数的度量,$\mathrm{Dim}(B)$是贝叶斯网络的维度,N 表示数据集中样本个数,$\frac{1}{2}\log N$ 表示每一个参数使用的比特数.

基于 BIC 的类相关性影响可变选择性贝叶斯分类器的算法步骤:

步骤 1 初始化:$S\leftarrow\mathrm{CCRI}(T,m)$,即使用 CCRI 方法选择出 m 个属性结点;

步骤 2 用 BIC 测度从初始子集 S 中选择出使 BIC 评分增加的属性结点,最终确定属性子集 S';

步骤 3 在选择出的属性子集 S' 上构建朴素贝叶斯分类器.

3.3 连续属性离散化

使用 CCRI 方法进行属性选择只能处理离散数据集,对于含有连续变量的数据集需要进行离散化处理.另外,由于 CCRI 方法需要处理高维数据集,所以在离散化过程中要求离散化后的属性值的种类尽可能少,同时也要保持信息丢失尽量小.

为了简化分类任务和提高分类精度,Kurgan 等提出类别属性依赖最大化(Class-Attribute Interdependent Maximization,CAIM)离散化方法[8].它主要使用 CAIM 准则来测试离散化后的属性同类别结点间的互信息,CAIM 值越大,说明类别依赖性越大,分类效果越好.表 1 描述了经过离散化后的属性 x_l.

表 1 离散化后的属性 x_l 取值表

类别	区间					类总数
	$[d_0,d_1]$	…	$(d_{r-1},d_r]$	…	$(d_{e-1},d_e]$	
c_1	w_{11}	…	w_{1r}	…	w_{1e}	M_{1+}
⋮	⋮	…	⋮	…	⋮	⋮
c_k	w_{k1}	…	w_{kr}	…	w_{ke}	M_{k+}
⋮	⋮	…	⋮	…	⋮	⋮
c_q	w_{q1}	…	w_{qr}	…	w_{qe}	M_{q+}
区间总数	M_{+1}	…	M_{+r}	…	M_{+e}	M

CAIM 值的计算方法如下:

$$\mathrm{CAIM}(c,(d,x_l))=\frac{\sum_{r=1}^{e}\frac{\max_r^2}{M_{+r}}}{e} \quad (10)$$

式中,d 表示对应于属性 x_l 的离散化变量;e 表示离散化后属性取值个数;w_{kr}是属于第 k 类,在区间$(d_{r-1}, d_r]$内的样本总个数;M_{k+} 是属于第 k 类的样本总个数;M_{+r}是属于第$(d_{r-1}, d_r]$区间内的样本个数;$\max_r$ 是在第 r 列中所有 w_{kr}中的最大值.

CAIM 离散化方法将每个属性的样本从小到大排列,合并取值相同的样本,取每两个相邻样本值的中点作为待测试的划分点[8]. CAIM 离散化方法存在的主要问题是离散化的时间太长,无法处理样本量过大的数据集.

等频离散化方法(Equal Frequency Discretization, EFD)是一种比较简单的离散化方法,其离散化思想是:按等区间大小的原则(即每个区间上的样本个数相等)进行离散化[9]. EFD 离散化方法对朴素贝叶斯分类器有较好的分类效果,这主要是由于一般假设数值型数据具有 dirichlet 分布,而 dirichlet 分布完美的聚合性使得使用离散化的朴素贝叶斯恰当地估计了属性的分布[10].但是,它需要人为地设定离散化后区间的个数,具有任意性,无法确定产生多少个区间是最好的.

结合 EFD 和 CAIM 离散化方法的优点,本文提出等频类别依赖最大化(Equal Frequency Class Attribute Interdependent Maximization, EFCAIM)离散化方法. EFCAIM 离散化的主要思想是:首先,用 EFD 方法确定待选的划分点;然后,使用 CAIM 的评分准则从待选划分点中挑选出一定数量的最优划分点.

EFCAIM 离散化方法的计算步骤如下:

步骤 1　初始化:设置每个区间的样本个数为 u,最大划分点个数为 t;

步骤 2　将属性 x_l 的样本按从小到大排列,按每个区间样本个数 u 来确定每个区间;

步骤 3　将取值相等的样本合并到同一区间,设置第二个区间以后的每个区间的最小值为一个划分点,得到待选的划分点集合 $H = \{d_1, \cdots, d_r, \cdots, d_e\}$;

步骤 4　从 H 中选择 CAIM 准则评分最大的划分点 $d_{\max}$,放入最优划分点集合 H_{best}中,$H_{\text{best}} \leftarrow d_{\max}$;

步骤 5　重复步骤 3 至划分点个数等于 t;

步骤 6　确定最终的划分点集合 H_{best},对属性 x_l 的样本进行离散化处理.

4　实验研究

4.1　数据集概况

UCI 上很多人工数据集和实验数据集被世界各地相关领域的学者和专家用来进行机器学习、知识发现、人工智能和概率统计以及其它领域算法的实验研究和比较.选取 6 个具有代表性的 UCI 标准数据集来测试贝叶斯分类器的分类性能,它们分别为: Chess、Splice、Semetion、Musk、Arr 和 Hdr,其中包含 2 分类和多分类,连续数据和离散数据,低维数据和高维数据.数据集的概况见表 2.

表 2　数据集概要

数据集	类别数	属性个数	属性类型	样本数
Chess	2	36	离散	3196
Splice	3	60	离散	3190
Semetion	10	256	离散	1593
Musk	2	166	连续	476
Arr	2	279	连续	452
Hdr	10	649	连续	2000

4.2　选择性贝叶斯分类器对比实验

实验分为两部分,首先测试调节因子 α 同分类准确率之间的关系,说明类别相关性的影响程度对分类准确率的影响. 对 Chess、Semetion、Arr 和 Hdr 数据集进行调节因子 α 的测试,选择属性结点个数均为 20 个,实验结果见图 1. 图 1 中,黑色实心方块标记表明:与最优参数 α 相对应的分类准确率,即 CCRI SBC 得到的分类准确率;$\alpha = 1$ 时对应的纵坐标的值即为 mRMR SBC 得到的分类准确率. 以 Chess 数据集为例, CCRI SBC 和 mRMR SBC 的分类准确率分别为 0.9308 和 0.8848.

由图 1 可以看出,并不是在 $\alpha = 1$ 的情况下分类效果最好,这是由于利用 mRMR 方法选择出的属性子集中仍然包含有冗余信息. 在 CCRI SBC 中,由于调节因子的调节作用可以得到最佳的属性子集,从而提高分类准确率. 分类准确率一般在 α 取值小于 1 的条件下得到最优值,这说明在相对降低类相关性影响程度的条件下,能够得到较好的分类效果.

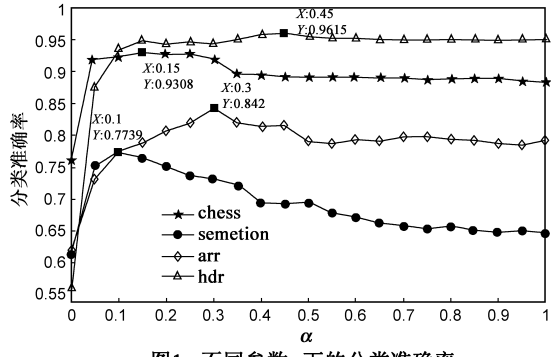

图1　不同参数α下的分类准确率

分别采用 maxMI SBC、mRMR SBC 和 CCRI SBC 对这 6 组数据集进行分类,属性结点个数分别取 5、10、15、20、25 和 30 共 6 种情况,表 3 ~ 表 5 给出了这 3 种选择性贝叶斯分类器的分类准确率. 在实验中,连续数据集采用 EFCAIM 方法进行离散化处理.

表 3 maxMI SBC 的分类准确率

属性个数	Chess	Splice	Semetion	Musk	Arr	Hdr
5	0.8670	0.9116	0.2705	0.7063	0.7380	0.8435
10	0.8836	0.9433	0.4422	0.7953	0.7632	0.9070
15	0.8858	0.9567	0.5216	0.7986	0.7546	0.9275
20	0.8817	0.9611	0.5396	0.8246	0.7518	0.9445
25	0.8783	0.9611	0.6071	0.8105	0.7713	0.9470
30	0.8792	0.9602	0.6308	0.8317	0.7672	0.9465
均值	0.8793	0.9490	0.5020	0.7945	0.7577	0.9193

表 4 mRMR SBC 的分类准确率

属性个数	Chess	Splice	Semetion	Musk	Arr	Hdr
5	0.9215	0.9116	0.4488	0.7851	0.7789	0.8510
10	0.8993	0.9445	0.4380	0.8290	0.7985	0.9395
15	0.8992	0.9567	0.5951	0.8480	0.7922	0.9465
20	0.8848	0.9618	0.6139	0.8741	0.7945	0.9510
25	0.8808	0.9611	0.6616	0.8947	0.8058	0.9550
30	0.8738	0.9611	0.7069	0.8833	0.7963	0.9505
均值	0.8932	0.9495	0.5774	0.8524	0.7944	0.9323

表 5 CCRI SBC 的分类准确率

属性个数	Chess	Splice	Semetion	Musk	Arr	Hdr
5	0.9409	0.9116	0.5022	0.8051	0.7709	0.9030
10	0.9224	0.9495	0.6559	0.8539	0.8274	0.9455
15	0.9221	0.9580	0.7268	0.8545	0.8416	0.9540
20	0.9308	0.9627	0.7740	0.8751	0.8421	0.9615
25	0.9274	0.9617	0.7847	0.8947	0.8221	0.9560
30	0.9099	0.9614	0.8035	0.8887	0.8284	0.9570
均值	0.9256	0.9508	0.7079	0.8620	0.8221	0.9462

从表 3 到表 5 可以看出,互信息最大化的方法虽然简单,但是分类准确率不够高.这说明仅仅选择同类结点相关性最大的属性结点是不充分的,冗余结点的存在降低了分类准确率.mRMR 方法相比较 maxMI 方法有着较好的分类准确率,但是仍然存在部分属性结点冗余的问题,因此分类效果也不是最优的.CCRI 方法有着最好的分类准确率,可以最大程度地减少属性之间的冗余性,同时也兼顾到了类别结点相关性的影响.

4.3 基于 BIC 的选择性贝叶斯分类器对比实验

将基于贝叶斯信息准则的属性个数确定方法分别应用于 maxMI SBC、mRMR SBC 和 CCRI SBC 这 3 种选择性贝叶斯分类器,即首先使用 maxMI、mRMR 或 CCRI 方法来确定初始子集 S,然后使用 BIC 测度进行属性结点的确定.通过分类准确率和属性结点选择的个数两方面来验证 BIC 测度进行属性结点个数确定的有效性.表 6 给出了基于 BIC 的选择性贝叶斯分类器的分类结果,其中分类准确率为十重交叉验证方法得到的平均值,属性结点个数为十次实验的平均值.

由表 6 可以看出,使用 BIC 测度进行属性结点个数确定是有效的.它在原有属性子集的基础上选择出了同数据集拟合最好的属性,大部分数据集都可以得到比较好甚至更优的分类准确率.CCRI SBC 能够在选择较少属性结点的条件下,依然有着较高的分类准确率,说明这种方法得到的初始属性结点集合比较优秀.maxMI 和 mRMR 两种方法得到的属性个数较多,但分类精度不高,这说明了 BIC 评分不能消除冗余属性,所以不能单独使用 BIC 测度进行属性个数的确定.

表 6 基于 BIC 的选择性贝叶斯分类器的分类结果

选择性贝叶斯分类器	Chess		Splice		Semetion	
	分类准确率	属性个数	分类准确率	属性个数	分类准确率	属性个数
BIC-Based maxMI SBC	0.8814	21	0.9602	30	0.6416	30
BIC-Based mRMR SBC	0.8902	18	0.9618	29	0.6836	30
BIC-Based CCRI SBC	0.9293	13	0.9624	28	0.8047	29
选择性贝叶斯分类器	Arr		Musk		Hdr	
	分类准确率	属性个数	分类准确率	属性个数	分类准确率	属性个数
BIC-Based maxMI SBC	0.7695	30	0.8344	30	0.9465	30
BIC-Based mRMR SBC	0.7963	30	0.8887	26	0.9505	30
BIC-Based CCRI SBC	0.8461	11	0.8675	20	0.9560	30

4.4 离散化方法对比实验

对 Musk、Arr 和 Hdr 三个连续数据集分别用 EFD、CAIM 和 EFCAIM 三种方法进行离散化处理,然后采用 CCRI SBC 对离散化后的数据集进行分类.比较这三种离散化方法对分类准确率和离散化时间的影响,表 7 给出了实验结果.

表 7 不同离散化方法分类实验

离散化方法	Musk		Arr		Hdr	
	分类准确率	离散化时间(s)	分类准确率	离散化时间(s)	分类准确率	离散化时间(s)
EFD	0.7404	0.88	0.7363	0.77	0.9345	4.59
CAIM	0.8422	96.11	0.8138	49.38	0.9450	7484.00
EFCAIM	0.8620	24.14	0.8221	29.70	0.9462	269.06

由表 7 可以看出,由于 EFCAIM 离散化方法结合了 EFD 和 CAIM 两种方法的优点,具有分类准确率高,离散化时间短的优点.EFD 方法虽然实现起来比较简单且花费时间较短,但是分类的效果不好,不能成为一个有效的离散化方法.CAIM 离散化方法主要存在离散化时间长的缺点,这主要是因为待测划分点的个数比较多,所以需要花费很长的时间进行最优划分点的选择.对于 CCRI 选择性贝叶斯分类器,EFCAIM 离散化方法有着较好的分类效果.

5　结论

选择性贝叶斯分类器通过属性选择方法,如最大相关最小冗余 mRMR 法,来选择出具有较好属性独立性关系的属性来参与分类模型的学习.但是,由于 mRMR 方法是根据类别互信息和属性间互信息的差值来进行属性选择的,因此,通常会引入冗余的属性.为此,通过给 mRMR 引入一个用于改变类别相关性在属性选择中影响程度的调节因子,提出一种类相关性影响可变选择性贝叶斯分类器 CCRI SBC,同时采用贝叶斯信息准则来自动确定最优属性个数,从而可以有效减少冗余结点的个数和提高分类准确率.进一步,为使 CCRI SBC 能够处理含有连续变量的数据分类问题,将属性依赖最大化和等频离散化方法相结合,提出一种等频类别依赖最大化离散化方法,具有分类准确率高和离散化时间短的优点.典型 UCI 数据集的实验结果验证了所提方法的可行性和有效性.值得指出的是,交叉验证法确定调节因子需要花费较多的时间,因此,在今后的工作中将研究如何更加高效地选择合适的调节因子以进一步提高 CCRI SBC 的分类性能.

参考文献

[1] 蔺志青,郭军.贝叶斯分类器在手写汉字识别中的应用[J].电子学报,2002,30(12):1 – 4.
Lin Zhi-Qing, Guo Jun. An application of bayesian classifier in the recognition of handwritten chinese character[J]. Acta Electronica Sinica, 2002, 30(12): 1 – 4. (in Chinese)

[2] P Langley, S Sage. Induction of selective bayesian classifiers [A]. Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence[C]. San Fransisco: Morgan Kaufmann Publishers Inc, 1994. 339 – 406.

[3] 赵军阳,张志利.基于最大互信息最大相关熵的特征选择方法[J].计算机应用研究,2009,26(1):233 – 235.
Zhao Jun-Yang, Zhang Zhi-Li. Feature subset selection based on max mutual information and max correlaton entropy[J]. Application Research of Computers, 2009, 26(1): 233 – 235. (in Chinese)

[4] Y Yang, J O Pedersen. A comparative study on feature selection in text categorization[A]. Proceedings of the 14th International Conference on Machine Learning[C]. New York: ACM Press, 1997. 412 – 420.

[5] H Peng, F Long, C Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1227 – 1238.

[6] N Friedman, D geiger, M Goldszmidt. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2 – 3): 131 – 163.

[7] G Schwarz. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6(2): 461 – 464.

[8] L A Kurgan, K J Cios. CAIM discretization algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 145 – 153.

[9] Y Yang, G I Webb. Weighted proportional k-interval discretization for naive-bayes classifiers[J]. Lecture Notes in Computer Science, 2003, 2637: 501 – 512.

[10] Y Yang, G I Webb. On why discretization works for naive bayesian classifiers[J]. Lectures Notes in Artificial Intelligence, 2003, 2903: 309 – 406.

作者简介

程玉虎　男,1973 年生于安徽淮南,2005 年获中国科学院自动化研究所控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院教授、博士生导师.主要研究方向为机器学习和智能系统等.
E-mail: chengyuhu@163.com

仝瑶瑶　女,1987 年生于江苏徐州,2009 年获中国矿业大学学士学位,现为中国矿业大学控制理论与控制工程专业硕士研究生,研究方向为贝叶斯分类器.
E-mail: lcxtynf@163.com

王雪松　女,1974 年生于安徽泗县,2002 年获中国矿业大学控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院教授,博士生导师.主要研究方向为机器学习、复杂系统优化与控制、生物信息学等.
E-mail: wangxuesongcumt@163.com