文章编号:1001 - 2486(2004)05 - 0063 - 07

基于搜索编码的简单贝叶斯分类方法*

蒋艳凰,杨学军
(国防科技大学计算机学院,湖南 长沙 410073)

摘 要:简单贝叶斯法性能稳定,分类精度难以提高。通过分析搜索编码法产生的纠错输出码的性质,提出基于搜索编码的简单贝叶斯算法 SCNB,并详细阐述了 SCNB 算法的应用流程。实验结果表明,采用搜索编码法能够有效提高简单贝叶斯分类器的泛化能力。

关键词:监督分类;简单贝叶斯算法;纠错输出码;搜索编码法
中图分类号:TP181 文献标识码:A

A Bayesian Learning Algorithm Based on Search-Coding Method

JIANG Yan-huang, YANG Xue-jun
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Naïve-Bayes algorithm is a stable supervised learning method, and it is difficult to improve its predicting accuracy. This paper analyzes the properties of the error-correcting output codes generated by search-coding method at first, then presents a search coding based on vaïve Bayes algorithm (SCNB), and describes the flow chart of SCNB in detail. Experimental results show that search-coding method is an efficient approach to improve the generalization for Bayesian classifiers.

Key words: supervised; classification; Naïve-Bayes algorithm; error-correcting output code (ECOC); search-coding method

简单贝叶斯法[1]是监督分类中最为常用的学习算法,对于大多数的应用问题,即使属性之间不满足独立性,仍然能够取得较为理想的分类结果。提高学习算法的泛化能力(generalization)一直是监督分类的重要研究内容,高精度的预测结果也是应用的不断需求[2]。由于简单贝叶斯法性能稳定,训练样本集小的变动对学习结果的影响很小,因此很难通过分类器组合等方法来提高其预测精度[3]。

纠错输出编码技术[4,5]不仅可用于将多类分类问题简化为多个两类问题来处理,扩展学习算法的应用范围,而且利用输出码具有纠错能力这一特性,可以提高监督分类器的泛化能力。搜索编码法对含任意类别数的监督分类问题,均能产生满足要求的纠错输出码。

在下面的讨论中,将样本 $\underline{X}$ 表示为属性向量的形式,即 $\underline{X}=(x_1,x_2,\cdots,x_l)$,元素 x_j 为样本 $\underline{X}$ 的在第 j 个属性上的值,l 为属性的个数,各属性可以为离散或连续属性;并令 $CS=\{c_1,c_2,\cdots,c_m\}$ 为类别的集合,m 为类别的个数;$|LS|$ 为训练样本集 LS 中的元素个数。

1 简单贝叶斯算法

贝叶斯统计分类法的思想是将未知类别的样本 $\underline{X}=(x_1,x_2,\cdots,x_l)$ 分类给其最可能属于的类别 $\hat{y}$,即有:

$$\hat{y}=\arg\max_{j}\max_{c_j\in CS}P(c_j\mid x_1,x_2,\cdots,x_l)$$

根据贝叶斯理论,可将上式写为:

$$\hat{y}=\arg\max_{k}\max_{c_k\in CS}\frac{P(x_1,x_2,\cdots,x_l\mid c_k)\cdot P(c_k)}{P(x_1,x_2,\cdots,x_l)}$$

* 收稿日期:2004 - 08 - 06
基金项目:国家杰出青年科学基金资助项目(69825104)
作者简介:蒋艳凰(1976—),女,博士生。

$$= \arg\max_{k} \max_{c_k \in CS} P(x_1, x_2, \cdots, x_l \mid c_k) \cdot P(c_k) \tag{1}$$

显然需要根据训练样本集,估计式(1)中各概率项的值。一般取 $P(c_k)$ 的值等于训练样本集中属于类别的样本出现的频率,但是 $P(x_1, x_2, \cdots, x_l \mid c_k)$ 的值难以估计。

简单贝叶斯法(Naïve-Bayes, NB)是各种贝叶斯方法中理论简单、使用广泛、效果好的一种方法。简单贝叶斯法基于一个简单的假设:在给定任一类别的条件下,各属性的值相互独立,即 $P(x_1, x_2, \cdots, x_l \mid c_k) = \prod_{i=1}^{l} P(x_i \mid c_k)$。因此简单贝叶斯法可描述为:

$$y_{NB} = \arg\max_{k} \max_{c_k \in CS} P(c_k) \cdot \prod_{i=1}^{l} P(x_i \mid c_k) \tag{2}$$

式(2)中,$P(c_k)$ 与 $P(x_i \mid c_k)$ 的估计值分别为它们在训练样本集中出现的频率。

上述关于贝叶斯方法的讨论是针对离散属性而言,若第 i 个属性为连续属性,显然不能用计数法确定 $P(x_i \mid c_k)$,可采用如下两种方法:一是假定属性 i 关于第 k 个类别的条件概率密度函数(用函数 $f_{i \mid k}$ 表示)的形式已知,如服从正态分布,利用极大似然估计获得这些概率密度函数中的未知参数,然后预测未知样本 $\underline{X}$ 的类别:

$$y_{NB} = \arg\max_{k} \max_{c_k \in CS} P(c_k) \cdot \prod_{i=1}^{l_1} f_{i \mid k}(x_i) \cdot \prod_{j=1}^{l_2} P(x_j \mid c_k) \tag{3}$$

式(3)中,l_1 与 l_2 分别表示连续属性与离散属性的数目,满足 $l_1 + l_2 = l$。第二种方法是将连续属性的取值范围离散化成若干区间,然后将该连续属性作为一个离散属性处理,每个区间对应离散属性的一个取值,并将落入某区间内的属性值转化为该区间对应的离散值,最后统计每个区间内的样本数,并利用式(2)计算预测结果。由于大多数应用领域中概率分布形式难以用公式表示,因此将连续属性离散化是一种更有效的方法[6,7]。

与决策树、BP 神经网络等方法不同,简单贝叶斯法是一种性能稳定的学习算法,由于算法本身利用的是训练样本的统计信息,因此对训练样本集小的变动基本不会影响贝叶斯分类器的预测精度。对于性能稳定的监督学习算法,其分类精度难以提高。如何提高简单贝叶斯法的泛化能力值得研究者的关注[3,8]。

2 搜索输出编码

2.1 搜索编码算法

令 m 个长度为 n 的码字组成的集合为 CM,CM 可表示成大小为 $m \times n$ 的矩阵形式,矩阵的每一行对应一个码字,我们称 CM 为码矩阵。在后面的讨论中,将码字的集合用码矩阵表示,并令 $CM[i]$ 表示码矩阵的第 i 个码字(即第 i 行),$CM[*, j]$ 表示码矩阵的第 j 列,$CM[i, j]$ 则表示 CM 中第 i 行第 j 列对应位的值。

纠错输出码将编码理论中纠错码的思想用于监督分类。基于纠错输出码的监督分类过程可描述为:首先利用类别数 m 构造一个具有纠错能力的码矩阵 $CM^{m \times n}$(称 CM 为纠错输出码),每个类别对应着 CM 中一个长度为 n 的码字,这些码字的每一列对应一个两类分类问题。令第 i 列的理想二值函数为 f_i,样本 $\underline{X}$ 的真实类别的编号为 $Class(\underline{X})$,则有:

$$f_i(\underline{X}) = \begin{cases} 1 & \text{if } CM[Class(\underline{X}), i] = 1 \\ 0 & \text{else} \end{cases}$$

然后利用训练样本对各列的二值函数进行学习,获得 n 个二分器;在分类阶段,各二分器的输出形成一个输出向量,再利用决策方法判定该输出向量与 CM 中各码字的相似度,预测样本所属的类别。纠错输出码一方面将一个 $m(m \geqslant 2)$ 类问题转化为 n 个两类问题,另一方面,利用输出码本身具有的纠错能力,可以纠正某些二分器的错误输出,从而提高分类器的泛化能力。

搜索编码法[9]通过对整数空间的顺序搜索,可以获得满足任意类别数与最小汉明距离要求的纠错

输出码。该方法将非负整数与二进制位串对应起来,输入类别数 m 与期望的最小汉明距离 d,利用计算机自动搜索出满足要求的 m 个码字。图 1(a)和图 1(b)分别给出了搜索编码法中创建码数表项和生成纠错输出码的伪代码。

```
Create TableItem (d,n)
1. If n < d Then Return 0;
2. Initialization: CI = {0};
3. For Each Integer x in [1,2^n - 1]
    3.1 Tag = True;
    3.2 For Each Integer y in CI
        If DiffBit(Bin(x,n),Bin(y,n)) < d
            Then Tag = False;
    3.3 If Tag = True Then CI = {x} ∪ CI;
4. Return |CI|.
```

(a)创建码数表项伪代码

(a) Pesudo code of creating CodeTable item

```
SearchCode (d,m)
1. Initialization: i = 0, CI = {0}, x = 1;
2. n = FindCodeLen(CodeTable,d,m);
3. While |CI| < m and x < 2^n do
    3.1 Tag = True;
    3.2 For Each Integer y in CI
        If DiffBit(Bin(x,n),Bin(y,n)) < d
            Then Tag = False;
    3.3 If Tag = True Then CI = {x} ∪ CI;
    3.4 x = x + 1;
4. For Each Element y in CI;
    CM[i] = Bin(y,n), i = i + 1;
    5. Return CM.
```

(b)生成纠错输出码伪代码

(b) Pesudo code of generate error-correcting output codes

图 1 搜索编码算法伪代码

Fig.1 Pesudo code of search coding method

在进行搜索编码之前,需要生成码数表 $CodeTable$。该表的每一项 $item(d,n)$ 记录了最小汉明距离为 $d(d\geqslant 1)$,码长为 $n(n\geqslant 1)$ 的情况下,满足要求的所有码字数目。码数表可作为永久信息保存起来。编码时,首先根据类别数 m、期望的最小汉明距离 d 以及创建好的码数表,确定纠错输出码的码长,若有 $item(d,n-1)<m\leqslant item(d,n)$,则码长为 n,然后利用计算机搜索产生 m 个码长为 n 的码字,形成纠错输出码。图 1(b)中的函数 $SearchCode(d,m)$ 用于获取含有 m 个码字的码矩阵 CM,且满足最小汉明距离为 d。在搜索过程中,首先利用 $FindCodeLen(CodeTable,d,m)$ 确定码长 n,然后利用与函数 $CreateTableItem(d,n)$ 相似的处理方式确定一个含 m 个整数的集合,再将这 m 个整数转化为 m 个长度为 n 的二进制位串,形成纠错输出码 CM。显然,在搜索编码法中,$Bin(0,n)$ 为输出码中缺省的码字。

2.2 搜索输出码的特性分析

我们用 $\Phi(d,n)$ 表示在给定 d,n 的情况下,函数 $CreateTableItem(d,n)$ 中获得的整数集合 CI,且集合中的元素按从小到大的顺序排列;$\lambda(\Phi(d,n),i,n)$ 表示将集合 $\Phi(d,n)$ 中的前 i 个整数转化为 i 个码长为 n 的码字组成的码矩阵,显然 $\lambda(\Phi(d,n),m,n)$ 即为搜索编码法获得的码矩阵。由于编码过程均是从零码开始搜索,因此有 $item(d,n-1)\leqslant item(d,n)$,且 $\Phi(d,n-1)\subseteq\Phi(d,n)$。

定理 1 已知类别数为 m,期望的最小汉明距离为 d,码数表为 $CodeTable$,令 $n=FindCodeLen(CodeTable,d,m)$,则由搜索法获得的码矩阵 $\lambda(\Phi(d,n),m,n)$ 中无全 0 列、无全 1 列、无互补列。

证明 根据已知条件,有 $item(d,n-1)<m\leqslant item(d,n)$。令 $\Phi(d,n)$ 中前 m 个整数为 $s_1,s_2,\cdots,s_m$,由于搜索编码法中整数 0 对应的码字为缺省码,且从 1 开始按递增的顺序对整数区间 $[1,2^n-1]$ 进行搜索,因此有 $s_1=0$,$s_i(2\leqslant i\leqslant m)$ 为继 s_{i-1} 之后满足最小汉明距离条件的最小整数。令 $CM=\lambda(\Phi(d,n),m,n)$,并对三种情况分别证明:

(1) 若 $\lambda(\Phi(d,n),m,n)$ 中存在全 0 列

① 若 $CM[*,n]$ 为全 0 列,则直接删除这一列可以获得新的码矩阵 $CM'=\lambda(\Phi(d,n-1),m,n-1)$,该码矩阵不仅满足最小汉明距离为 d 的条件,而且各码字所对应的整数也不变,由此可知 $\{s_1,s_2,$

…,s_m}$\subseteq\Phi(d,n-1)$,即 $m\leqslant item(d,n-1)$。这与 $m>item(d,n-1)$矛盾,因此全0列不会出现在最高位。

② 若 $CM[*,i](1\leqslant i<n)$为全0列,令 $\boldsymbol{CM}$ 中的第 k 行为其首行存在某 $j(j\geqslant i)$位为1的一行,且该行对应的整数为 s_k。由①知 $CM[*,n]$不是全0列,因此满足条件的 k 必然存在。删除 CM 中的第 i 列得到码矩阵 $\boldsymbol{CM}'$,令 $\boldsymbol{CM}'$ 中各码字对应的整数依次为 $t_1,t_2,\cdots,t_m$,则有 $s_j=t_j(1\leqslant j<k)$,$s_j>t_j(k\leqslant j\leqslant m)$。因此有 $t_k<s_k$,且 t_k 与前面各行的汉明距离均大于等于 d,这与 s_k 为继 s_{k-1} 满足要求的最小整数矛盾,故第 $i(i<n)$列不可能为全0列。

由①、②可知,$\lambda(\Phi(d,n),m,n)$中无全0列。

(2) 若 $\lambda(\Phi(d,n),m,n)$中存在全1列

假设 $CM[*,i](1\leqslant i\leqslant n)$为全1列,则必有 $CM[1,i]=1$,这与 $s_1=0$ 矛盾,故 $\lambda(\Phi(d,n),m,n)$中无全1列。

(3) 若 $\lambda(\Phi(d,n),m,n)$中存在互补列

假设 $CM[*,i](1\leqslant i\leqslant n)$与 $CM[*,j](1\leqslant j\leqslant n,j\neq i)$两列互补,则 $CM[1,i]$与 $CM[1,j]$中必有一位为1,这与 $s_1=0$ 矛盾,故 $\lambda(\Phi(d,n),m,n)$中无互补列。

定理1得证。

定理1描述了搜索输出码的性质。表1给出了码矩阵 $\lambda(\Phi(5,10),8,10)$,可见,码矩阵 $\lambda(\Phi(5,10),8,10)$中第四列与第五列,第七列与第八列,第九与第十列分别相同,因此利用搜索编码法获得的输出码矩阵中可能存在相同列。

表1 搜索法所得的码矩阵 $\lambda(\Phi(5,10),item(5,10),n)$

Tab.1 Code matrix $\lambda(\Phi(5,10),item(5,10),n)$

行	列									
	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0
3	1	1	0	0	0	1	1	1	0	0
4	0	0	1	1	1	1	1	1	0	0
5	1	0	1	0	0	1	0	0	1	1
6	0	1	0	1	1	1	0	0	1	1
7	0	1	1	0	0	0	1	1	1	1
8	1	0	0	1	1	0	1	1	1	1

搜索编码法具有两个优点:一是对于给定的 m 与 d,搜索编码法得到的输出码的码长较其它编码法短,因此其生成的二值函数较少,这能在一定程度上降低学习的复杂度,尤其是对于 m 较大的情况效果明显;二是搜索法可获得满足任意类别数 $m(m\geqslant 2)$、任意最小汉明距离 $d(d\geqslant 3)$的纠错输出码(为了通用性,搜索编码法对 $d=1,2$ 的情形也予以考虑)。因此,搜索编码法可作为一种通用的构造纠错输出码的方法。由于搜索法获得的输出码矩阵中可能存在相同列,而相同列所表示的二值函数完全相同,因此如何对相同列进行处理,从而保持纠错输出码的纠错能力,是将搜索编码法用于监督分类的一个关键问题。

3 基于搜索编码的简单贝叶斯算法 SCNB

3.1 算法流程

我们提出的基于搜索编码的简单贝叶斯算法(SCNB)是利用简单贝叶斯法对码矩阵中各列所对应的二值函数进行学习。图2给出了SCNB算法的流程。从图2可知,将SCNB算法用于监督分类,其处理流程分为三个部分:编码、学习、分类。编码过程是利用搜索编码法获得合适的纠错输出码;学习阶段包括对连续属性离散化的预处理部分与对 n 个二值函数进行学习的学习部分;分类阶段需采用合适的决策方法对输出向量进行评估。

3.2 学习策略

利用纠错输出码将原来的分类问题转化为若干个两类分类问题后,即使原分类问题中各类别在连续属性上概率分布形式已知,经过编码转化后,所形成的两类分类问题中的条件概率密度函数也难以确定,因此我们在学习之前采用区间离散化的方法处理连续属性。

在连续属性离散化过程中,我们设定两个参数 $LIMIT_R$ 和 $LIMIT_S$。$LIMIT_R$ 是离散化后的最大区间数,用于保证划分的充分性;$LIMIT_S$ 是落入每个区间内的最少样本数,用于保证各区间的概率估计信息的可靠性。令连续属性 A 的最大值为 $\max_A$,最小值为 $\min_A$,首先将属性 A 的值等分成 $LIMIT_R$ 个区间,每个区间的长度为 $(\max_A - \min_A)/LIMIT_R$,然后统计落入每个区间内的样本数目,若某区间内的样本数目少于 $LIMIT_S$,将其合并到下一区间,若最后一个区间的样本数过少,则将其合并到前一个区间。为了使每个区间的概率估计能近似描述相应的概率分布信息,参数 $LIMIT_R$ 不能太小,$LIMIT_R$ 大一些倒无妨,因为 $LIMIT_S$ 可保证概率估计的可靠性,$LIMIT_S$ 越大,概率估计值越可靠,但导致划分不充分;$LIMIT_S$ 越小,影响相反。

离散化过程结束后,则利用简单贝叶斯法依次对搜索输出码所确定的 n 个二值函数进行学习。由于搜索编码法获得的输出码矩阵中可能存在相同列,若对它们的二值函数采用同一训练样本集进行学习,获得的分类器也相同,这将降低纠错输出码的纠错能力。我们采用放回式随机抽样的方法解决这个问题:对每个二值函数 f_i 进行学习之前,均对 LS 采用一次放回式随机抽样,生成新的训练集 LS_i,使 $|LS_i| = |LS|$,然后利用 LS_i 对 f_i 进行学习。

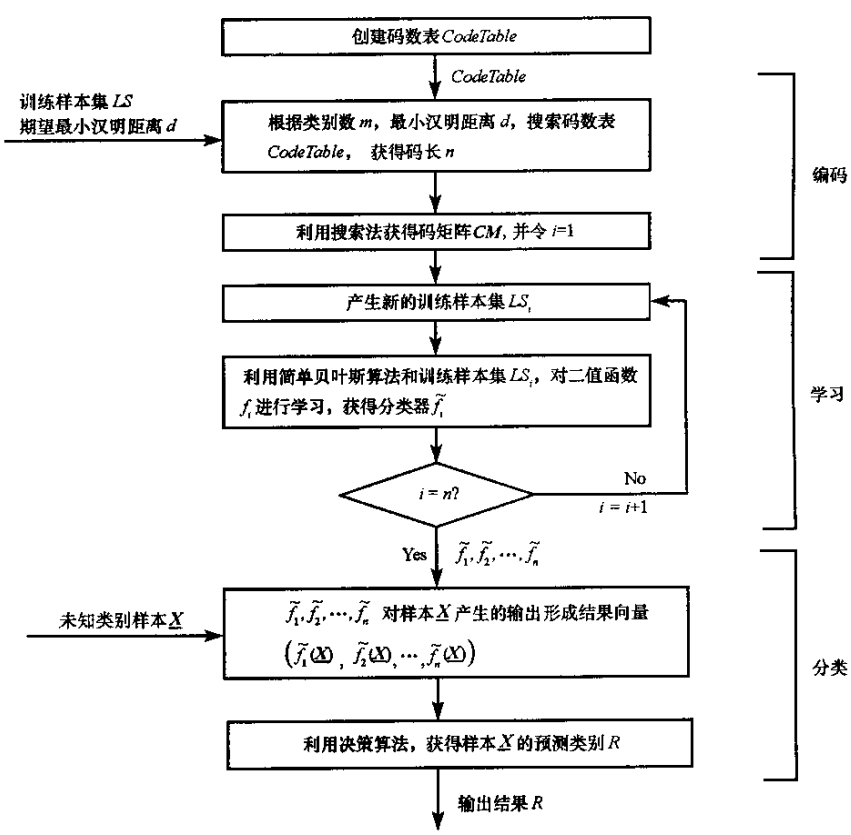

图 2 SCNB 算法流程

Fig. 2 Flow chart of SCNB algorithm

3.3 决策方法

学习结束后,获得 n 个二值函数的分类器近似表示。在分类阶段,令这 n 个分类器对样本 $\underline{X}$ 的输出结果形成结果向量 $\tilde{f}(\underline{X})=(\tilde{f}_1(\underline{X}),\tilde{f}_2(\underline{X}),\cdots,\tilde{f}_n(\underline{X}))$,如何通过结果向量判断输入样本属于何种类别?我们提出如下两种决策方法:

(1) 汉明距离法

该方法是利用阈值向量 $(h_1,h_2,\cdots,h_n)$,将结果向量 $\tilde{f}(\underline{X})$ 转化为输出码字 $\boldsymbol{OC}(\underline{X})=(y_1,y_2,\cdots,y_n)$,其中

$$y_i=\begin{cases}1 & \text{if} \quad \tilde{f}_i(\underline{X})-h_i\geqslant 0\\ 0 & \text{else}\end{cases}$$

缺省情况下,各分类器的阈值均为 0.5,即 $h_i=0.5(1\leqslant i\leqslant n)$。经过阈值转换后,计算 $\boldsymbol{OC}(\underline{X})$ 与码矩阵 $\boldsymbol{CM}$ 中各码字之间的汉明距离,令

$$R=\min\{j\mid d_H(\boldsymbol{OC}(\underline{X}),\boldsymbol{CM}[j])=\min_{1\leqslant k\leqslant m}(d_H(OC(\underline{X}),CM[k]))\}$$

即从 $\boldsymbol{CM}$ 中选择与输出码字的汉明距离最小的码字所对应的类别为输入样本的类别。当 $\boldsymbol{OC}(\underline{X})$ 与 $\boldsymbol{CM}$ 中的多个码字的距离都是最小距离时,则选择类别标识最小的作为输入样本的类别。汉明距离法利用了汉明纠错码的思想,决策方法简单易懂,但它忽略了每个分类器输出值的大小,而这些值可能是非常有用的决策信息,为此我们提出绝对距离法。

(2) 绝对距离法

该方法直接计算结果向量 $\tilde{f}(\underline{X})$ 与 $\boldsymbol{CM}$ 中各码字之间的绝对距离,选择绝对距离最小的码字所对应的类别作为预测类别,距离计算方法如下:

$$d_A(\boldsymbol{f}(\underline{X}),\boldsymbol{CM}[k])=\sum_{1\leqslant i\leqslant n}|\boldsymbol{f}_i(\underline{X})-\boldsymbol{CM}[k,i]|$$

样本 $\underline{X}$ 的预测类别为:

$$R=\arg\min_{1\leqslant k\leqslant m}d_A(f(\underline{X}),\boldsymbol{CM}[k])$$

4 实验结果与分析

利用 UCI 机器学习数据库[10]中的 11 个数据集进行实验,测试学习性能。由于原 Cancer 与 Cleveland 两个数据集中含有未知的属性值,将未知的属性值取为同类别的样本在该属性上的平均值。为了测试对连续属性的处理,所选数据集的属性均为连续属性。实验中四种简单贝叶斯法分别是:正态分布法利用正态分布来描述条件概率密度函数;直接区间分割法采用区间离散化方法处理连续属性;SCNB(汉明距离)采用搜索编码获得的码矩阵,并利用汉明距离作为决策依据;SCNB(绝对距离)则利用绝对距离作为决策依据。

在学习过程中,根据简单贝叶斯法的计算公式,令类别 c_k 出现的次数 $freq(c_k)=n_k$,属于类别 c_k 的样本中属性 i 上的取值为 x_i 的样本数为 $freq(c_k,x_i)=n_{ik}$,则有 $P(x_i\mid c_k)=n_{ik}/n_k$。如果 $n_{ik}=0$,由于相乘的关系,无论其它属性的作用如何,均有 $P(c_k\mid\underline{X})=0$,这往往是不合理的。为了避免这种现象,我们采用概率的 $m-$估计法,利用 $(n_{ik}+mp)/(n_k+m)$ 作为 $P(x_i\mid c_k)$ 的估计值,其中 P 为 x_i 出现的先验概率,一般采用平均分布的方法,若属性 i 有 r 个取值,则令 $p=1/r$,此处的 m 为一权重,也称相对样本大小,表示新增加 m 个样本,有 pm 个样本在属性 i 的取值为 x_i。实验中取 $m=1$。

实验采用 10 次交叉验证的方法,即将数据集划分成类别分布相似、大小相同的 10 个样本子集,每次取其中的 9 个作为训练集,剩余的一个作为测试集,利用 10 次结果的均值与方差来描述算法的性能。实验中搜索编码时期望的最小汉明距离为 5,连续属性离散化的参数取 $LIMIT_S=8$,$LIMIT_R=20$。

表 2 各种简单贝叶斯法的实验结果

Tab.2 Experimental results of four naïve-bayes algorithms

Datasets	正态分布法	直接区间分割法	SCNB(汉明距离)	SCNB(绝对距离)
Austra	19.13 ± 4.52	14.20 ± 4.03	13.91 ± 4.49	14.20 ± 4.36
Bupa	43.53 ± 10.26	40.59 ± 6.01	35.88 ± 5.27	35.58 ± 6.05
Cancer	4.06 ± 1.65	2.90 ± 1.93	2.75 ± 1.93	2.60 ± 1.74
Cleveland	43.67 ± 7.45	42.00 ± 5.02	41.00 ± 3.87	40.33 ± 3.67
Glass	50.48 ± 7.84	29.04 ± 7.26	32.38 ± 7.71	30.47 ± 6.83
Heart	15.56 ± 4.88	16.67 ± 6.59	15.18 ± 6.06	14.07 ± 5.53
Ionosphere	10.57 ± 5.23	10.85 ± 5.35	10.28 ± 4.89	10.00 ± 4.51
Iris	4.67 ± 3.22	6.00 ± 5.83	6.00 ± 5.83	4.00 ± 3.44
Page	9.73 ± 1.96	6.14 ± 1.12	6.67 ± 1.15	6.03 ± 0.83
Pima	24.34 ± 5.16	25.13 ± 4.49	23.29 ± 3.62	23.03 ± 3.89
Wine	2.35 ± 3.04	2.94 ± 5.00	2.35 ± 4.11	2.94 ± 5.00
Average	20.74	17.86	17.24	16.66

表 2 给出了各数据集采用不同的简单贝叶斯方法的实验结果,最优的结果用黑体表示。从实验结果可以看出,有多个数据集的条件概率分布不能简单地用正态分布来近似描述,因此导致基于正态分布的简单贝叶斯法分类结果不理想。对连续属性采用区间离散化大大改善了分类的精度,这是因为理论上区间分割法可以用于近似任何形式的概率分布。对于 SCNB 算法,两种不同的决策方法,绝对距离法的结果优于汉明距离决策法,这是因为绝对距离法利用了各二分器预测结果的概率信息,而汉明距离法直接将它们的输出转化为码字的形式,丢掉了一部分有用信息。比较四种方法的错误率,采用 SCNB(绝对距离)法对 8 个数据集的错误率最低,其平均错误率也最低;采用 SCNB(汉明距离)的平均结果也优于正态分布法和直接区间分割法。因此利用搜索编码法可以有效地提高简单贝叶斯分类器的泛化能力。

5 结 论

简单贝叶斯法作为一种性能稳定的分类方法,其分类精度难以提高。通过分析搜索编码法产生的纠错输出码的性质,提出了基于搜索编码的简单贝叶斯算法 SCNB,并对 SCNB 算法的应用流程、学习策略、决策方法等进行了详细阐述。实验结果表明,采用搜索编码法能够有效地提高简单贝叶斯分类器的泛化能力。

参 考 文 献:

[1] Langley P, Iba W, Thompson K. An analysis of Bayesian Classifiers[C]. In Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, CA: AAAI Press, 1992: 223 – 228.

[2] Dietterich T. Machine Learning Research: Four Current Directions[C]. AI Magazine, 1997, 18(4): 97 – 136.

[3] Ting K M, Zheng Z J. Improving the Performance of Boosting for Naïve Bayesian Classification[J]. In Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, Berlin: Springer-Verlag, 1999: 296 – 305.

[4] Dietterich T, Bakiri G. Solving Multiclass Learning Problems Via Error-correcting Output Codes[J]. Journal of Artificial Intelligence Research, 1995, 2: 263 – 286.

[5] Masulli F, Valentini G. Effectiveness of Error Correcting Output Codes in Multiclass Learning Problems[J]. Lecture Notes in Computer Science, 2000, 1857: 107 – 116.

[6] Hsu C N, Huang H J, Wong T T. Why Discretization Works for Naïve Bayesian Classifiers[C]. In Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000: 399 – 406.

[7] Yang Y, Webb G I. Non-disjoint Discretization for Naïve-Bayes Classifiers[C]. In Proceedings of the Nineteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2002: 666 – 673.

[8] Kohavi R. Scaling up the Accuracy of Naïve-Bayes Classifiers: A Decision-tree Hybrid[C]. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, New York, AAAI Press, 1996: 202 – 207.

[9] 蒋艳凰, 周海芳, 杨学军. 基于纠错编码的 CSNN 及其在遥感图像分类中的应用[J]. 计算机研究与发展, 2003, 40(7): 918 – 924.

[10] Bay S D. UCI KDD Archive (http://kdd.ics.uci.edu)[R]. 1999.