一种搜索编码法及其在监督分类中的应用*

蒋艳凰+, 赵强利, 杨学军

(国防科学技术大学 计算机学院 软件研究所,湖南 长沙 410073)

A Search Coding Method and Its Application in Supervised Classification

JIANG Yan-Huang+, ZHAO Qiang-Li, YANG Xue-Jun

(Institute of Software, School of Computer Science, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: Phn: +86-731-4575810, E-mail: jiang-yh@163.com, http://www.nudt.edu.cn

Received 2003-11-10; Accepted 2005-01-07

Jiang YH, Zhao QL, Yang XJ. A search coding method and its application in supervised classification. *Journal of Software*, 2005,16(6):1081−1089. DOI: 10.1360/jos161081

Abstract: Supervised classification based on error-correcting output codes (ECOC) is a new research direction to improve the generalization of classifiers, yet there is no general method to construct ECOC for any number of classes. This paper analyzes the properties of ECOC and presents a search coding method which corresponds to codewords with integers and gets a satisfied output code through searching an integer range in sequence. It then describes the supervised classification technique based on the search coding method. By applying the search coding method to naïve-Bayes algorithm and BP neural networks, experimental results show that the method is an effective and general coding method to construct error-correcting output codes.

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

摘 要: 纠错输出码作为监督分类领域中的一个新的研究方向,是提高分类器泛化能力的一种有效方法,但目前还没有通用的确定性编码方法.分析了现有纠错输出码的性质,提出一种搜索编码法,该方法通过对整数空间的顺序搜索,获得满足任意类别数目与最小汉明距离要求的输出码;然后探讨了基于搜索编码的监督分类技术.对简单贝叶斯与 BP 神经网络算法进行实验,结果表明,搜索编码法可作为一种通用的编码方法用于提高监督分类器的泛化能力.

关键词: 监督分类;纠错输出码(ECOC);搜索编码法;简单贝叶斯算法;BP 神经网络

中图法分类号: TN911 **文献标识码**: A

监督分类是机器学习领域中的重要研究内容.我们将样本 $\underline{X}$ 表示为属性向量的形式,即 $\underline{X}=(x_1,x_2,\ldots,x_l)$,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 $CS=\{c_1,c_2,\ldots,c_m\}$ 为类别

* Supported by the National Natural Science Foundation of China under Grant No.69825104 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2002AA1Z2101(国家高技术研究发展计划(863))

作者简介: 蒋艳凰(1976−),女,湖南邵阳人,博士,助理研究员,主要研究领域为机器学习,图像处理,并行计算;赵强利(1973−),男,高工,主要研究领域为信息安全,人工智能;杨学军(1963−),男,教授,博士生导师,主要研究领域为并行处理,计算机体系结构.

的集合,m 为类别的个数;则监督分类问题可描述为:给定一个已知类别的样本集合 $LS=\{(\underline{X}_1,y_1),(\underline{X}_2,y_2),\ldots,(\underline{X}_{|LS|},y_{|LS|})\}$(其中$|LS|$为集合 LS 中的元素个数,$y_i\in CS$ 为样本 $\underline{X}_i$ 的类别),该样本集合确定了属性向量 $\underline{X}$ 与类别 y 之间的某种未知函数关系 $y=f(\underline{X})$,首先利用学习算法 $\Im$ 对已知样本集 LS 进行学习,得到分类器 $\tilde{f}(LS)$ 以近似未知函数 f,然后利用 $\tilde{f}(LS)$ 对未知类别的样本进行类别预测.

许多监督学习算法是针对两类分类问题设计的,如支持向量机[1]、二分决策树[2]等,而实际应用中的众多问题为多类分类问题,将多类分类问题转化为多个两类问题,一方面拓宽了学习算法的应用领域,另一方面,对于具体的分类问题,可选择的算法更多.目前多类问题两类化的方法主要有:

① 按类输出法(one-per-class):即根据类别数 m 将原问题转化为 m 个两类问题,对于第 i 个两类问题,其理想二值函数对属于第 i 种类别的样本输出为 1,对属于其余类别的样本输出为 0.

② 两两比较法(all pairs):每两种类别进行相互比较,共形成 C_m^2 个二分问题[3].

③ 语义编码法(semantic coding):该方法根据语义对各类别进行编码,码字的每一位都具有特定的含义[4].

④ 纠错输出编码法(error-correcting output codes,简称 ECOC):每个类别对应一个码字,且这些码字形成的输出码具有一定的纠错能力[5].Allwein[6]将各种多类问题两类化的方法统一用编码矩阵表示,并认为两类化后各二分器的组合性能关键取决于输出码的设计,即如何编码的问题.

纠错输出码不仅可用于多类问题两类化,而且利用输出码具有纠错能力这一特性,可以提高分类器的泛化能力(generalization).关于纠错输出码的研究已成为监督分类领域的一个新方向[7−9].但是,目前没有一个通用的编码方法能够对任意类别数 m 均生成合适的纠错输出码.

本文通过分析纠错输出码的性质,针对现有编码方法的缺陷,提出一种搜索编码法,并详细分析了利用搜索编码法获得的输出码的特性;然后探讨了搜索编码法在监督分类中的应用.实验结果表明,搜索编码法可作为一种通用的编码方法,且能有效地提高分类器的泛化能力.

1 纠错输出编码

令 m 个长度为 n 的码字组成的集合为 CM,CM 可表示为大小为 $m\times n$ 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,我们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以描述为: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 CM 中的一个长度为 n 的码字,这些码字的每一列对应一个两类问题,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 $m(m\geq 2)$ 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.

一个有用的纠错输出码应具有如下特性:

性质 1. 具有一定的纠错能力.

根据编码理论,对于最小汉明距离为 d 的纠错码,可以纠正 $[(d-1)/2]$ 位错误.因此具有纠错能力的输出码其各码字之间的最小汉明距离应≥3.毫无纠错能力的编码不能称为纠错输出编码.

性质 2. 码矩阵中无全 0 列、无全 1 列.

若码矩阵 CM 中存在某列 $j(1\leq j\leq n)$,对于任意 $i(1\leq i\leq m)$,均有 $CM[i,j]=0$(或 $CM[i,j]=1$),则函数 f_j 对属于任何类别的样本输出均为 0(或 1),即 f_j 为单值函数,对分类毫无作用,应将该列从码矩阵中删除.

性质 3. 码矩阵中无相同列,无互补列.

假设码矩阵的第 $i,j(i\neq j)$两列完全相同,则它们所对应的二值函数的理想输出完全相同,即理论上有 $f_i=f_j$.对于不含随机因素的学习过程,对这两个函数进行学习得到的二分器也完全相同,它们的作用等同于一个二分器,因此,删除这两列中的一列对输出码的纠错能力毫无影响.对于互补的两列,它们所对应的二值函数仅是输出结果互换,如果学习过程不含随机因素,学习后得到的两个二分器的预测结果完全相关,也就是说,这两列的作用仍等同于一列,需将其中的一列删除.

显然,m 种类别最多能够形成 2^m 种互不相同的列,删除全 0 列、全 1 列以及互补列,实际上生成的码矩阵中最多可有 $2^{m-1}-1$ 列.只有当类别数 $m\geq 5$ 时,才能生成满足上述 3 个性质的纠错输出码.

Dietterich[5]列出了 4 种编码方法,包括列举编码法、列选择法、随机爬山法和 BCH 编码法[10].然而这几种方法均存在自己的缺陷:列举编码法生成的输出码虽然纠错能力强,但码长随着类别数目呈指数增长,导致转化后的二值函数急剧增多,学习过程复杂,因此对于 $m\geq 8$ 的情况,一般不再采用列举编码法,列选择法和随机爬山法属于非确定性算法,其编码过程复杂,生成的输出码矩阵具有随机性,采用 BCH 编码法获得的码矩阵中码字的个数均为 2 的幂次方,若类别数 m 不是 2 的幂次方,则需要利用一些启发式方法缩短码长,减少码字数,应用起来很不方便.Crammer 与 Singer[11]提出连续码的概念,并提出一种编码方法用于设计与具体问题相关的输出码,但编码过程复杂,通用性差.目前还没有通用的确定性编码算法能对任意类别数 m 均获得适用的纠错输出码.

2 搜索编码算法

按照纠错输出码的 3 个性质,对 $m<5$ 的情况没有可用的纠错输出码,这在一定程度上限定了纠错输出码的应用范围.另外,由于没有通用的编码算法,对于 m 较大的情况,很难设计出合适的输出码.本节针对上述两个问题,提出一种通用的编码方法——搜索编码法,该方法适用于任意类别数 $m(m\geq 2)$.

2.1 搜索编码算法

搜索编码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.

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 $d(d\geq 1)$,码长为 $n(n\geq 1)$ 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.算法 1 给出了创建码数表项的伪代码,函数 $CreateTableItem(d,n)$ 用于确定码数表中 $item(d,n)$ 项的值.整数集合 CI 的初始状态仅含有元素 0,从整数 1 开始,按递增的顺序检查$[1,2^n-1]$区间内的所有整数,若某整数对应的二进制位串与集合 CI 中所有整数对应的二进制位串的汉明距离均大于等于 d ,则将该整数加入集合 CI 中.当搜索过程结束,集合 CI 中所有元素的个数即为码数表项 $item(d,n)$ 的值.显然,若存在满足要求的码字,必有 $n\geq d$.函数 $Bin(x,n)$ 表示将十进制整数 x 表示成长度为 n 的二进制位串的形式,二进制位串的第 $j(0\leq j\leq n)$位为$[x/2^j]$ mod 2.函数 $DiffBit(G,H)$ 表示两个二进制位串 G,H 中相同分量上的值不同的位数.表 1 为在 $d\leq 9,n\leq 16$ 的情况下的码数表,其中省略了 $n=1,2$ 两列.

算法 1. 创建码数表项的伪代码.

```
CreateTableItem(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|.
```

Table 1 CodeTable with $d \le 9$ and $3 \le n \le 16$

表 1 码数表($d \le 9, 3 \le n \le 16$)

$d(CC)$	Length of codewords													
	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768	65536
2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768
3	2	2	4	8	16	16	32	64	128	256	512	1024	2048	2048
4	0	2	2	4	8	16	16	32	64	128	256	512	1024	2048
5	0	0	2	2	2	4	4	8	16	16	32	64	128	256
6	0	0	0	2	2	2	4	4	8	16	16	32	64	128
7	0	0	0	0	2	2	2	2	4	4	8	16	32	32
8	0	0	0	0	0	2	2	2	2	4	4	8	16	32
9	0	0	0	0	0	0	2	2	2	2	2	4	4	4

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,确定输出码的码长,若有 $item(d,n-1) < m \le item(d,n)$,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,形成满足要求的搜索输出码.算法 2 为生成输出码的伪代码.函数 $SearchCode(d,m)$ 通过搜索,获得一个含有 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包含 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.

算法 2. 生成输出码的伪代码.

```
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.
```

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 $\lambda(\Phi(d,n),m,n)$ 即为由搜索编码法获得的码矩阵.由于我们的编码过程均是从零码开始搜索,因此有 $item(d,n-1) \le item(d,n)$,且 $\Phi(d,n-1) \subseteq \Phi(d,n)$.下面我们证明如下定理.

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 $item(d,n-1) < m \le item(d,n)$.令 $\Phi(d,n)$ 中前 m 个整数为 $s_1,s_2,...,s_m$,由于搜索编码法中整数 0 对应的码字为缺省码,且从 1 开始按递增顺序对整数区间[1,2^n−1]进行搜索,因此有 $s_1 = 0$, $s_i\ (2 \le i \le m)$ 为继 s_{i-1} 满足最小汉明距离条件的最小整数.我们令 $CM = \lambda(\Phi(d,n),m,n)$,并对 3 种情况分别加以证明:

(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 $m \le item(d,n-1)$.这与 $m > item(d,n-1)$ 矛盾,因此全 0 列不会出现在最高位.

② 若 $CM[*,i]\ (1 \le i < n)$ 为全 0 列,令 CM 中的第 k 行为其首行存在某 $j\ (j \ge i)$ 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 CM',令 CM' 中各码字对应的整数依次为 $t_1,t_2,...,t_m$,则有 $s_j=t_j (1 \le j < k), s_j > t_j (k \le j \le m)$.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} 之后满足要求的最小整数相矛盾,故第 $i(i<n)$ 列不可能为全 0 列.

由①②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 $CM[*,i](1\le i\le n)$ 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 $CM[*,i](1\le i\le n)$ 与 $CM[*,j](1\le j\le n, j\ne i)$ 两列互补,则 $CM[1,i]$ 与 $CM[1,j]$ 中必有一位为 1,这与 $s_1=0$ 相矛盾,故 $\lambda(\Phi(d,n),m,n)$ 中无互补列.

定理 1 得证. □

定理 2. 已知类别数为 m,期望的最小汉明距离为 d,码数表为 CodeTable,且有 n=*FindCodeLen*(*CodeTable*,d,m),若 $\lambda(\Phi(d,n),m,n)$ 中存在完全相同的 k 列,则它们必为连续的 k 列.

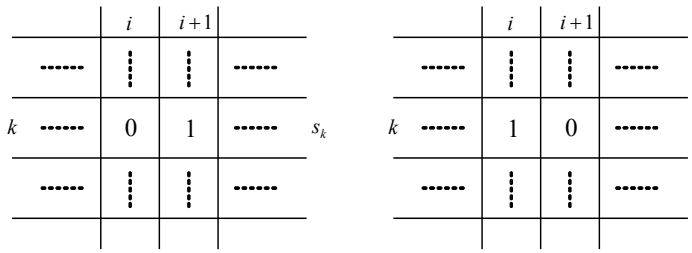
证明:令 $CM=\lambda(\Phi(d,n),m,n)$.首先,我们求得码矩阵 CM 中各行与各列的二进制位串对应的整数.对于每行,码矩阵的左边为低位,右边为高位,令 $CM[i]\ (1\le i\le m)$ 对应的整数为 s_i,有 $s_i=CM[i,k]\sum_{k=1}^{n}2^{CM[i,k]\cdot(k-1)}$;对于每列,码矩阵的第 1 行为最高位,最后一行为最低位,令 $CM[*,j]\ (1\le j\le n)$ 对应的整数为 t_j,则有 $t_j=CM[k,j]\sum_{k=1}^{m}2^{CM[k,j]\cdot(m-k)}$.表 2 给出了码矩阵 $\lambda(\Phi(5,10),8,10)$,以及 $s_i(1\le i\le 8)$ 与 $t_j(1\le j\le 10)$ 的值.显然,码矩阵 $\lambda(\Phi(5,10),8,10)$ 中第 4 列与第 5 列,第 7 列与第 8 列,第 9 列与第 10 列分别相同,由此可知,利用搜索编码法获得的输出码矩阵可能存在相同列.

Table 2 Code matrix $\lambda(\Phi(5,10),item(5,10),n)$ obtained by the search coding method

表 2 搜索编码法获得的码矩阵 $\lambda(\Phi(5,10),item(5,10),n)$

Row	Column										
	1	2	3	4	5	6	7	8	9	10	s_i
1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0	31
3	1	1	0	0	0	1	1	1	0	0	227
4	0	0	1	1	1	1	1	1	0	0	252
5	1	0	1	0	0	1	0	0	1	1	805
6	0	1	0	1	1	1	0	0	1	1	826
7	0	1	1	0	0	0	1	1	1	1	966
8	1	0	0	1	1	0	1	1	1	1	985
t_j	105	102	90	85	85	60	51	51	15	15	

下面,我们证明不等式 $t_1\ge t_2\ge\ldots\ge t_n$ 成立.假设存在 $i\ (1\le i<n)$,满足 $t_i<t_{i+1}$,则码矩阵的第 i 列与第 $i+1$ 列必不相同,从第 1 行开始,依次比较这两列中相同位之值,假设第 k 行为首行出现这两列的值不相同的行,该行对应的整数为 s_k,由于 $t_i<t_{i+1}$,必有 $CM[k,i]=0$, $CM[k,i+1]=1$,如图 1(a)所示.



(a) (b)

Fig.1 Exchange two corresponding bits

图 1 交换相应位示意图

现在我们交换 $CM[k,i]$ 与 $CM[k,i+1]$ 之值(如图 1(b)所示),显然这不改变第 k 行与前面各行之间的汉明距离,令此时第 k 行对应的整数为 s'_k,则有 $s'_k<s_k$,且该行与前面各行之间的汉明距离大于等于 d,这与搜索编码法所得到的 s_k 为继 s_{k-1} 之后满足要求的最小整数相矛盾,因此假设不成立.故不等式 $t_1\ge t_2\ge\ldots\ge t_n$ 成立.

若码矩阵 CM 中存在完全相同的 k 列,则这 k 列对应同一个整数相同,由于 $t_1\ge t_2\ge\ldots\ge t_n$,可知它们必为相邻的 k 列.

故定理 2 得证. □

定理 3. 已知类别数为 m,期望的最小汉明距离为 d,码数表为 CodeTable,且有 n=*FindCodeLen*(*CodeTable*, d,m),若 $\lambda(\Phi(d,n),m,n)$ 中存在完全相同的 k 列,则必有 $k \le \lceil d/2 \rceil$.

证明:令 $CM=\lambda(\Phi(d,n),m,n)$.若 CM 中存在相同的 k 列,由定理 2 可知,这 k 列必然相邻.令这 k 列为第 $r+1,...,r+k$ 列,并假设 $k>\lceil d/2 \rceil$.由于第 1 行为零码,因此 CM 中必存在大小为 $q\times k$ $(2\le q\le m)$的一个子块(如图 2(a)所示),使得 $CM[i,j]=0$ 且 $CM[q,j]=1$ $(1\le i\le q-1, r+1\le j\le r+k)$,由搜索编码的过程可知,$s_q>s_i(1\le i<q)$.假设存在 u,v $(1\le u\le q-1, v>r+k)$ 使得 $CM[u,v]=1$,则将 CM 中 $CM[u]$ 这一码字进行修改,得到新的码矩阵 CM',使得 $CM'[u,r+1]=1$,$CM'[u,v]=0$,CM'中其余位与 CM 相同,显然有 $s'_u<s_u$,且 $CM'[u]$ 与其前面各码字的汉明距离均≥d,这与 s_u 为继 s_{u-1} 之后的最小整数相矛盾.故有 $CM[i,j]=0$ $(1\le i\le q-1, j>r+k)$.

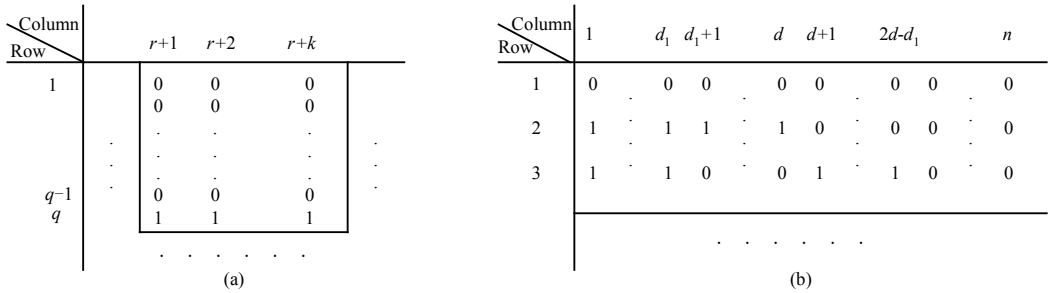


Fig.2 Several columns with the same bit values

图 2 相同列示意图

若 $r<d$,根据搜索法,第 2 个码字先是 d 个 1,后面是 $n-d$ 个 0,因此必有 $q=2$.若 $CM[3]$的前 d 位与前面两个码字的最小汉明距离为 d_1,则后面的 $n-d$ 位中必有 $d-d_1$ 个 1.由搜索法可知,s_3 为继 s_2 之后满足要求的最小整数,因此,d_1 必取最大值,即 $d_1=\lfloor d/2 \rfloor$,因此 $CM[3]$的前面 d 位应先是 d_1 位 1,后为 $d-d_1$ 位 0.而且,$CM[3]$从 $d+1$ 到 $2d-d_1$ 位的值为 1,最后面的 $n-2d+d_1$ 位为 0,这种情况如图 2(b)所示.由此可知,若 $r<d$,则必有 $k\le d-\lfloor d/2 \rfloor$,即 $k\le\lceil d/2 \rceil$.

若 $r\ge d$,我们可将 CM 中 $CM[q]$进行修改得到 CM'',$CM''[q]$的第 1,2,...,$\lfloor d/2 \rfloor$位与第 $r+1,...,r+k-1$ 位均为 1,其余位全为 0,并令 $CM''[q]$对应的整数为 s''_q.由于 $CM[i,j]=0$ $(1\le i\le q-1, j>r+k)$,因此 $CM''[q]$的前 r 位与前面各行相应位的最小汉明距离大于等于$\lfloor d/2 \rfloor$,后面 $n-r$ 位与前面各行相应位的最小汉明距离为 $k-1$.由于 $k>\lceil d/2 \rceil$,即 $k>d-\lfloor d/2 \rfloor$,因此 $CM''[q]$与前面各行的汉明距离均大于等于 d,且有 $s''_q<s_q$,这与 s_q 为继 s_{q-1} 之后的最小整数矛盾.因此当 $r\ge d$ 时,不可能出现 $k>\lceil d/2 \rceil$ 的情况.

故定理 3 得证. □

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他编码法短,因此其生成的二值函数数目少,这能在一定程度上降低学习的复杂性,尤其是对于 m 较大的情况效果更明显;二是搜索法可获得满足任意类别数 m $(m\ge 2)$,任意最小汉明距离 d $(d\ge 1)$的输出码(当 $d\ge 3$ 时,形成的输出码具有纠错能力,能够纠正 $\lfloor (d-1)/2 \rfloor$位错误).因此搜索编码法可作为一种通用的输出码构造方法.由于搜索编码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,如何对相同列进行处理,从而保持搜索输出码的纠错能力,是将其用于监督分类的一个关键问题.

3 基于搜索编码的监督分类技术

3.1 监督分类流程

基于搜索编码的监督分类,即是利用搜索编码法产生的纠错输出码将待解决的分类问题转化为多个两类分类问题,并利用输出码的纠错能力提高分类精度.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,这将降低搜索输出码的纠错能力.我们采用放回式随机抽样的方法解决这个问题:在对每个二值函数 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.

将搜索编码法用于监督分类,其处理流程分为 3 个部分:编码、学习、分类.编码过程是利用搜索编码法获得搜索输出码 *CM*;学习阶段依次对由 *CM* 产生的 *n* 个二值函数进行学习,其学习过程与具体算法相关;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 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 $\tilde{f}(\underline{X})=\left(\tilde{f}_1(\underline{X}),\tilde{f}_2(\underline{X}),...,\tilde{f}_n(\underline{X})\right)$,如何通过结果向量判断输入样本属于何种类别?我们提供如下两种决策方法:

(1) 汉明距离法(HD):该方法是利用阈值向量 $(h_1,h_2,...,h_n)$,将结果向量 $\tilde{f}(\underline{X})$ 转化为输出码字 $OC(\underline{X})=(y_1,y_2,...,y_n)$,其中:

$$y_i=\begin{cases}1, & \text{if } \tilde{f}_i(X)-h_i\ge 0\\ 0, & \text{else}\end{cases}.$$

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

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

即从 *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 $OC(\underline{X})$ 与 *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,但它不能用于最小汉明距离为 1 或 2 的情况,而且在计算过程中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,为此我们提供绝对距离法.

(2) 绝对距离法(AD):该方法直接计算结果向量 $\tilde{f}(\underline{X})$ 与 *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:

$$d_A(f(\underline{X}),CM[k])=\sum_{1\le i\le n}\left|f_i(\underline{X})-CM[k,i]\right|.$$

样本 $\underline{X}$ 的预测类别为

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

4 实验结果与分析

简单贝叶斯与BP神经网络是两种具有代表性的监督学习算法,简单贝叶斯方法性能稳定,训练集的小变动对分类结果几乎没有影响,而 BP 神经网络则相反,是一种不稳定的学习算法.我们将搜索编码法用于简单贝叶斯与 BP 神经网络两种算法中,并利用 UCI 机器学习数据库[12]中的 9 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,我们将未知的属性值取为同类别的样本在该属性上的平均值.两次实验均采用 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 1 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.实验中搜索编码法所取的期望最小汉明距离为 5.

4.1 基于搜索编码的简单贝叶斯算法SCNB

简单贝叶斯算法作为一种性能稳定的分类方法,很难利用常用的分类器集成方法,如 Bagging[13],Boosting[14]等来提高其分类精度[15].将搜索编码法用于简单贝叶斯分类,测试搜索编码对稳定算法的预测精度的影响.

SCNB算法是利用搜索编码法获得的输出码将待分类问题转化为多个两类问题,再利用简单贝叶斯法对这些两类问题进行学习.在 SCNB 算法中,即使原分类问题中各类别在连续属性 *i* 上概率分布形式已知,经过编码转化后,所形成的二分问题中的条件概率密度函数也难以确定,因此我们采用区间离散化的方法处理连续属性.

我们首先设定两个参数 $LIMIT_R$ 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的样本数目,若某区间内的样本数少于 $LIMIT_S$,则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.在实验中我们取 $LIMIT_S=8$, $LIMIT_R=20$.

实验中,4 种简单贝叶斯算法分别是:正态分布法(NB-normal),即利用正态分布来描述每一类别在各属性上的分布;直接区间分割法(NB-D)是对连续属性采用区间分割的方式;SCNB(HD)与 SCNB(AD)分别是采用汉明距离和绝对距离作为决策方法的 SCNB 算法.表 3 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(由于 4 种贝叶斯算法学习速度均很快,在学习时间上差别不大,我们未将相应结果列出).

Table 3 Error results of different naïve Bayesian classifiers

表 3 各种简单贝叶斯法的错误率结果

Datasets	NB-normal	NB-D	SCNB (HD)	SCNB (AD)
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**
Iris	4.67±3.22	6.00±5.83	6.00±5.83	**4.00±3.44**
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	23.09	19.94	19.19	**18.58**

从实验结果可以看出,有多个领域的条件概率分布不能简单地用正态分布来描述,因此导致基于正态分布的简单贝叶斯法分类结果不理想.采用直接区间分割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.采用搜索输出码时,两种不同的决策方法中,绝对距离法的结果略优于汉明距离决策法.比较 4 种方法的错误率,采用 SCNB(AD)对 6 个数据集的错误率最低,其平均错误率也最低;采用 SCNB(HD)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能在一定程度上提高简单贝叶斯分类器的泛化能力.

4.2 基于搜索编码的BP神经网络SCBP

BP 神经网络是一种不稳定的学习算法,有很多集成手段可用于提高其泛化能力[16].在此我们检验搜索编码法对其分类精度的影响.

对于 SCBP 算法,根据搜索输出码的码长 n,共学习 n 个 BP 神经网络,每个神经网络用于近似描述一个二值函数.在我们的实验中,对隐含层节点数在 5~30 之间的多个网络结构进行学习,从而确定合适的隐含层节点数目,训练方式采用周期训练法,即对每个样本计算出权重误差导数,直到一个训练周期结束才计算权重的改变量并更新权重.权重的调节公式为

$$\Delta\omega(t+1)=\eta\frac{\partial E}{\partial\omega}+\alpha\Delta\omega(t),$$

其中学习速率 $\eta=0.5$,动量系数 $\alpha=0.9$.学习终止条件是均方误差 $MSE\le 0.02$ 或达到最大训练周期数 10000.表 4 给出了 BPNN 与 SCBP 的实验结果.

从实验结果可以看出,对于 Cancer 数据集,BPNN 与 SCBP 两种方法的分类精度相同,对于其余 8 个数据集,SCBP 的分类精度均优于传统 BPNN.在 SCBP 的两种决策方法中,采用汉明距离法结果要稍好些.从神经网络的训练时间可以看出,SCBP 的学习时间远远长于 BPNN,这是因为 SCBP 需要根据编码矩阵,学习多个神经网络,才能具有一定的纠错能力.这也说明 SCBP 是一种利用时间换取识别精度的方法.

通过对 SCNB 与 SCBP 算法的性能测试,结果表明,我们的搜索编码法是提高监督分类器泛化能力的一种较好的方法.但是对于搜索输出码的两种决策方法,我们的实验结果很难判断哪一种的预测效果更好.

Table 4 Experimental results of BPNN and SCBP
表 4 BPNN 与 SCBP 的实验结果

Dataset	BPNN		SCBP		
	Error rate (%)	Training time (s)	Error rate of HD (%)	Error rate of AD (%)	Training time (s)
Austra	15.79±4.29	**23.82**	14.35±3.64	**13.91±4.17**	97.67
Bupa	27.06±4.96	**19.21**	**25.59±5.55**	26.76±5.08	86.14
cancer	3.19±1.50	**11.19**	3.19±1.33	3.19±1.33	23.25
cleveland	46.67±8.16	**9.30**	**42.67±7.50**	43.67±7.10	153.17
Glass	29.52±7.37	**11.75**	**28.57±9.52**	29.05±9.90	38.73
Heart	21.11±7.66	**8.47**	**15.19±6.16**	15.92±7.21	55.52
Iris	4.67±4.50	**0.17**	4.00±4.66	**3.33±4.71**	1.98
Pima	23.15±4.77	**22.19**	**22.50±4.97**	23.28±4.80	102.77
Wine	1.76±2.84	**0.16**	1.76±2.84	**1.18±2.48**	0.60
Average	19.21	**11.81**	**17.53**	17.81	62.20

5 结 论

本文针对现有纠错输出编码的缺陷,提出一种搜索编码法,详细分析了搜索编码法获得的输出码的性质,并对基于搜索编码法的监督分类技术进行讨论.将搜索编码法用于简单贝叶斯与 BP 神经网络,结果表明,搜索编码法不仅通用性强,而且是提高分类器泛化能力的有效方法.

References:

[1] Cortes C, Vapnik V. Support-Vector networks. Machine Learning, 1995,20(3):273−297.

[2] Jiang YH, Yang XJ, Zhao QL. Constructing decision tree with high intelligibility. Journal of Software, 2003,14(12):1996−2005 (in Chinese with English abstract) http://www.jos.org.cn/1000-9825/14/1996.htm

[3] Hastie T, Tibshirani R. Classification by pairwise coupling. The Annals of Statistics, 1998,26(2):451−471.

[4] Sejnowski TJ, Rosenberg CR. Parallel networks that learn to pronounce english text. Journal of Complex Systems, 1987,1(1): 145−168.

[5] Dietterich T, Bakiri G. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research, 1995,2:263−286.

[6] Allwein E, Schapire RE, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifier. Journal of Machine Learning Research, 2000,1:113−141.

[7] Masulli F, Valentini G. Effectiveness of error correcting output codes in multiclass learning problems. Lecture Notes in Computer Science 1857, 2000. 107−116.

[8] Kuncheva LI. Using diversity measures for generating error-correcting output codes in classifier ensembles. Pattern Recognition Letters, 2005,26:83−90.

[9] Hauger SRB. Ensemble learned neural networks using error-correcting output codes and boosting [MS. Dissertation]. Department of Electronic Engineering, School of Electronics and Physical Science, University of Surrey, 2003.

[10] Bose RC, Ray-Chaudhuri DK. On a class of error-correcting binary group codes. Information and Control, 1960,3:68−79.

[11] Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems. In: Proc. of the 13th Annual Conf. on Computational Learning Theory. 2000. 35−46.

[12] Bay SD. UCI KDD Archive, 1999. http://kdd.ics.uci.edu

[13] Breiman L. Bagging predictors. Machine Learning, 1996,24(2):123−140.

[14] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Saitta L, ed. Proc. of the 13th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufman Publishers. 1996. 148−156.

[15] Ting KM, Zheng ZJ. Improving the performance of boosting for naïve Bayesian classification. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Berlin: Springer-Verlag. 1999. 296−305.

[16] Zhou ZH, Chen SF. Neural network ensemble. Chinese Journal of Computers, 2002,25(1):1−8 (in Chinese with English abstract).

附中文参考文献:

[2] 蒋艳凰,杨学军.具有高可理解性的二分决策树生成算法研究.软件学报,2003,14(12):1996−2005. http://www.jos.org.cn/1000-9825/14/1996.htm

[16] 周志华,陈世福.神经网络集成.计算机学报,2002,25(1):1−8.