基于覆盖算法决策界的特征选择算法

万 忠[1],张燕平[1,2],张 铃[1,2],陈 洁[1],张晨希[1],张迎春[1]
(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 人工智能研究所,安徽 合肥 230039)

摘 要:特征选择是模式识别系统的分类器设计之前一个重要而困难的一个课题。在目前现有的方法中,基于决策界的特征选择是其中一类方法。文中将覆盖算法应用于特征提取,提出了基于覆盖算法决策界的特征选择算法(Feature Selection Algorithm based on the Decision Boundary of Covering Algorithm,简称FSACA法),然后将该算法应用于一个字符识别的实例并与其他算法比较。实验结果证明了FSACA法的可行性和有效性。

关键词:特征选择;覆盖算法;特征选择算法;决策界
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)04-0084-04

Feature Selection Algorithm Based on Decision Boundary of Covering Algorithm

WAN Zhong[1],ZHANG Yan-ping[1,2],ZHANG Ling[1,2],
CHEN Jie[1],ZHANG Chen-xi[1],ZHANG Ying-chun[1]
(1.Ministry of Education Key Lab. of Intelligent Computing & Signal Processing
at Anhui University,Hefei 230039,China;
2.Institute of Artificial Intelligence,Anhui University, Hefei 230039,China)

Abstract:Feature selection is one of the important and difficult subjects before the design of the classifier in a pattern recognition system. Among the existed methods, one kind of these methods is based on the decision boundary. This paper applies covering algorithm into the field of feature selection, and puts forward the Feature Selection Algorithm based on Covering Algorithm(for short FSACA), and then a character recognition experiment is done to compare the proposed algorithm with others. The results of the experiment demonstrate the feasibility and the validity of FSACA.

Key words:feature selection;covering algorithm;feature selection algorithm;decision boundary

0 引 言

所谓模式识别,是指利用机器自动识别与分类。一个典型的模式识别系统由数据获取、预处理、特征提取与选择、分类器设计及分类决策五部分组成[1]。其中特征的提取选择是一个重要环节,直接影响着分类器的设计与性能。

覆盖算法是一种构造性学习算法[2],也是一种有效的分类器设计方法之一。文中将基于覆盖算法,提出描述覆盖算法决策界的点对分量分析,进而提出了基于覆盖算法决策界的特征选择算法(以下简称FSACA法)。然后将FSACA法与Rough Set属性约简的特征选择方法应用于图像识别的实例,实验结果证明了FSACA法的可行性与有效性。

收稿日期:2005-08-05
基金项目:"九七三"计划(国家重点基础研究)(2004CB318108);国家自然科学基金资助项目(60475017;60135010)
作者简介:万 忠(1980-),男,安徽合肥人,硕士研究生,研究方向为人工智能、神经网络和智能计算技术以及其在图像处理和智能交通工程的应用;张 铃,教授,博士生导师,从事人工智能理论、机器学习理论和方法、智能计算技术、神经网络技术的研究。

1 相关的概念介绍

在模式识别系统中,特征提取选择在不同应用环境下有不同的具体内容。一般说来它包括将所获取的原始量测数据转换成反映事物本质、并将其最有效分类的特征表示,这就必须要对测量空间(数据获取、预处理后的数据形成的空间)中的特征向量进行变换和选择来得到更有效的特征组合[1,3]。

1.1 特征提取

所谓特征提取是指通过映射或者变换降低描述对象样本的向量维数,一般可表示为变换 A:Y→X,其中 Y 为

测量空间,X 为特征空间。文中将不讨论这一步的优化问题。

1.2 特征选择

所谓特征选择是指在不降低分类能力的情况下从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的。当人们用高维特征进行分类器设计,无论从计算的复杂程度还是分类器性能来看都是不适宜的。因此研究如何从高维特征空间选择出最有效的低维特征以便有效地设计分类器就成为一个重要的课题。

1.3 类别可分离性判距

特征选择的任务是求出一组对分类最有效的特征,显然就需要一个定量的准测(或者叫判距)来衡量特征对分类的有效性,这个准则就是类别可分离性判距。一般常用的判距有基于类内类间距离的、基于概率分布的、基于熵函数的、基于分类界的等[3],另外还有一些基于各种搜索算法的特征选择方法,如模拟退火、Tabu 搜索、遗传算法等[3]。

1.4 覆盖算法

张铃教授于 1997 年就给出了 M－P 神经元模型的几何意义[2],指出用三层神经网络构造分类器,等价于求出一组领域,这组领域能将不同类的点分隔开来,并进一步给出覆盖设计算法[4~6]。

定义 1:覆盖 C 是指 n 维欧氏空间的一个球形领域(以 a 为中心,以 r 为半径的超球体)。

定义 2:给定样本集 S 分为 k 类,表示为集合 $S=\{S^1,S^2,\cdots,S^k\}$,$S^i(i=1,\cdots,k)$ 为属于第 i 类的样本集。如果覆盖集 $C=\{C^1,C^2,\cdots,C^p\}$,$C^j(j=1,\cdots,p)$ 均为覆盖,满足 $C^i\cap C^j$ 为空集$(i\neq j)$,并且每个 C^j 只和一个 S^i 相交以及 C^j 的并覆盖整个 S,则称 C 为 S 的划分覆盖集。

该算法的主要思路是:先求一个领域 C^1,它只覆盖一类中的点,而不覆盖其它类的点;对余下的点求二类覆盖领域 C^2,它只覆盖二类中的点而不覆盖其它类的点 …… 如此交叉进行覆盖,直到样本集中的点均被领域覆盖了为止。

覆盖算法的实质就是用求出的覆盖领域作为三层网络的隐含层,输入层为测试集,输出层为测试集的分类结果。构造覆盖算法的三层前向网络 FP 学习算法具体神经网络各层的设计算法[6] 如下:

设给定样本集为 $K=\{x^1,x^2,\cdots,x^k\}$(K 为 n 维欧氏空间的点集)。设 K 分为 s 个子集 $K^1=\{x^1,x^2,\cdots,x^{m(1)}\},\cdots,K^s=\{x^{m(s-1)+1},x^{m(s-2)+2},\cdots,x^k\}$,通过三层网络后,属于 K^i 的点的输出均为"y^i",其中 $y^i=(0,\cdots,1,0,\cdots,0)$(只有第 i 个分量为 1),$i=1,2,\cdots,s$。

第 1 层输入层直接输入测试样本特征向量;

第 2 层隐藏层,设计 p(p 为覆盖个数)个神经元 $A^1,\cdots,A^p$ 分别对应于覆盖 C^i,$i=1,2,\cdots,p$;

第 3 层输出层,取一个输入 x^j,其与 A^i(设以 a^i 为中心,r^i 为半径)神经元的权和阈值($W^i=(\omega^i),\theta=(\theta_i)$),则 $\theta_i=r_i$,$W=(a^i)$,$\theta=(\theta_i)$,$r(x)=<a^i,x>$。如 $r(x)\geqslant r_i$,则对应的 y^j 的第 i 个分量为 1,否则为 0。这样的三层前向神经网络就构成了分类器,功能是将 K 输入样本自动分为 s 个类别。

2 FSACA 法

2.1 决策界与覆盖领域的分类界

上文中提到,特征选择需要一种类别可分离性判距,而决策界可以作为判距之一。决策界是分类器在特征空间中所划分的不同类别空间领域的边界,如线性分类器的决策面是直线或者超平面,非线性分类器的决策面是曲线或者超曲面等[1,3]。基于决策界的特征选择是通过决策界的描述来对每一维特征的分类有效性进行定量分析,从而达到不降低识别能力的条件下缩减特征向量的维数。

覆盖算法是在识别对象的特征空间中求出若干覆盖,这些覆盖领域在高维特征空间中是一个超球体,覆盖领域的决策界就是超球体的超球面[7]。文中利用点对分量分析来描述覆盖领域的决策界[8]。

2.2 点对分量分析

定义 3:样本点是指样本集中的一个样本表示为一个特征向量,正好对应于特征空间的一个点,所以样本点是样本在特征空间的一种表示。

定义 4:点对是两个最邻近的(覆盖领域的边界之间的距离最短)、分别包含不同类别样本点的覆盖领域中彼此最靠近的一对异类样本点(文中的距离均为欧氏距离)。

定义 5:点对分量分析是找出特征空间中所有覆盖领域的点对进行以特征维为单位的定量计算分析。

下面通过例子给出点对分量分析的基本思路。设样本集是一个两个类别,两维特征向量集 $K=\{x^1,x^2,x^3,x^4,x^5,x^6,x^7,x^8,x^9,x^{10},x^{11},x^{12},x^{13}\}$,样本数为 13 个,$\{x^1,x^2,x^3,x^4,x^5,x^6\}$ 和 $\{x^7,x^8,x^9,x^{10},x^{11},x^{12},x^{13}\}$ 分别为一个类别。样本各特征向量值为 $x^1=(2,2)$,$x^2=(4,3)$,$x^3=(3,4)$,$x^4=(11,3)$,$x^5=(3,8)$,$x^6=(4,9)$,$x^7=(9,3)$,$x^8=(10,2)$,$x^9=(11,7)$,$x^{10}=(2,9)$,$x^{11}=(8,7)$,$x^{12}=(2,8)$,$x^{13}=(9,5)$,其样本空间分布如图 1 所示($*$ 与 $\circ$ 分别代表两个类别)。

用覆盖算法对 K 求覆盖,得到 9 个覆盖领域 A,B,C,D,E,F,G,H,I,如图 1 所示,其中 A,E,H,I 是同类覆盖,B,C,D,F,G 是同类覆盖。

由定义 3 可知,每个覆盖只有一个点对,如图 1 下面求 A 的点对,与 A 包含不同类别点的覆盖有 B,C,D,F,G,其中 B 是距离 A 的决策界最近的覆盖,这里两个覆盖的距离是决策界的最段距离,例如图 1 中 A 与 B 的距离可以由 A、B 中心(即圆心)的距离减去 A、B 的覆盖半径之和得到,这样 A 的点对为(2,13),因为 2 和 13 是 A 和 B 中的最靠近的一对异类样本点。由此可求出所有 9 个点对为(2,13),(13,2),(7,4),(8,4),(4,8),(10,5),(12,5),(5,12),(6,10),其中(2,13),(8,4),(12,5)重复出现,这样图

1 的所有覆盖的点对集为 $Q = \{(2,13),(7,4),(8,4),(10,5),(12,5),(6,10)\}$。

显然，点对是那些类别不同、彼此最靠近、互相纠缠的样本点对，也是距离每个覆盖的决策界最邻近的一对异类样本点，显然它们之间差别最能体现各维特征对于区分不同类别样本的可分性判据。点对分量分析，就是以特征向量的每一维为分析单位，计算每个点对在这个特征维上的差的绝对值，然后累计所有绝对值，以此绝对值的大小来作为该维特征对区分不同类别样本的可分性判据。

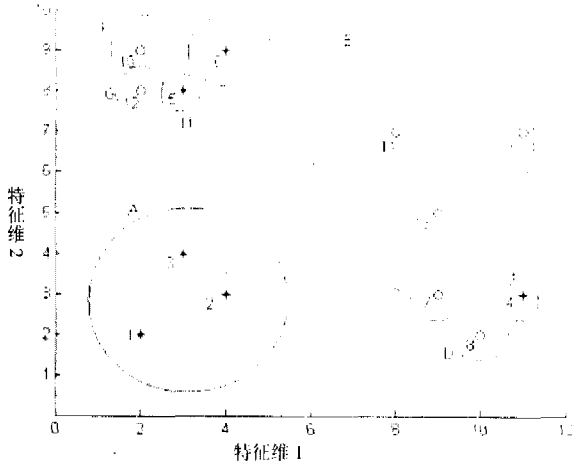

图 1 K 及其覆盖领域在二维平面上的分布

2.3 FSACA 法

基于点对分量分析，进一步提出了基于覆盖算法决策界的特征选择算法(FSACA 法)。

FSACA 法步骤：

①对给定样本集 X 用覆盖算法求出所有分类的覆盖领域；

② 求出覆盖领域中的所有对应的点对。给定一个 ε，对点对之间的距离 $\leqslant \varepsilon$ 的覆盖，转入 ③；

③ 对所有满足点对中两点距离 $\leqslant \varepsilon$ 的点对集合 Q 按分量分别统计绝对值之和，以这个和来表示该特征分量的可分性衡量，即若 $Q = \{(x^1,y^1),(x^2,y^2),\cdots,(x^t,y^t)\}$ 共 t 个点对，那么原特征向量第 j 个特征分量的可分性判据可有下式求出：

$$P_j = \sum_{i=1}^{t} | x_j^i - y_j^i |$$

删除 s 个分量差的绝对值最小的分量，即形成新的特征空间 X'。

④ 按新选的特征测试识别的错误率，若错误率下降，则返回 ①，继续特征选择；否则，若错误率上升，则恢复 ③ 中删除的 s 个特征，停止。

需要说明的是：FSACA 法中第 ② 步中引入 ε 的目的是剔除那些距离过大的点对。

3 实验与分析

3.1 字符图像识别的应用

文中所做的实验是将 FSACA 法应用到图像识别的特征选择中。这里以一组 512 维、共 457 个样本的训练集(512 * 457)和 320 个向量为测试集(512 * 320)的实例来说明。

应用 FSACA 法直接将所有的点对求对应的各维向量差的绝对值之和，对最小的 S 个分量删除，即：

① 对给定样本集 X 用覆盖算法求出所有分类的覆盖领域。

② 求出每一覆盖领域对应的点对。

③ 统计所有点对各分量差的绝对值之和，删除 s 个分量差的绝对值最小的分量，即形成新的 X。

④ 按新选的特征向量测试，若 s 识别错误率下降，则返回 ①，继续特征选择；否则，若错误率上升，则恢复 ③ 中删除的 s 个特征，停止用 FSACA 法，每次删除最小的 30 维分量，最后得：

* 以 512 维特征值，(512 * 457)直接覆盖，结果为：

正确：297 错误：9 拒识：14

正确率：297/(297 + 9) = 97.06%，

识别率：306/320 = 95.6%

* 以特征选择后的 212 维特征值，(212 * 457)直接覆盖，结果为：

正确：300 错误：7 拒识：13

正确率：300/(300 + 7) = 97.72%，

识别率：307/320 = 95.9%

FSACA 法特征选择的过程用图 2、图 3、图 4 说明。

上述图形清楚表明，FSACA 法在进行特征的选择时，所删除的特征对分类器的几个重要指标基本保持不变，当特征维数小于某个临界值之后，分类器的性能明显恶化，故此临界点就是最小特征维数。

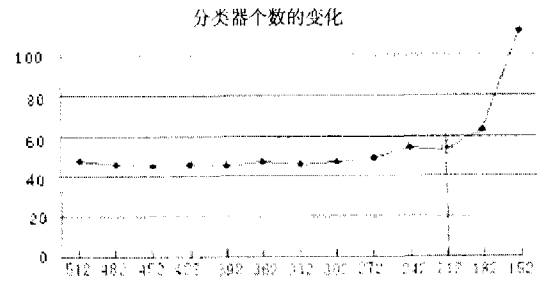

图 2 属性数目 - 分类器个数变化图

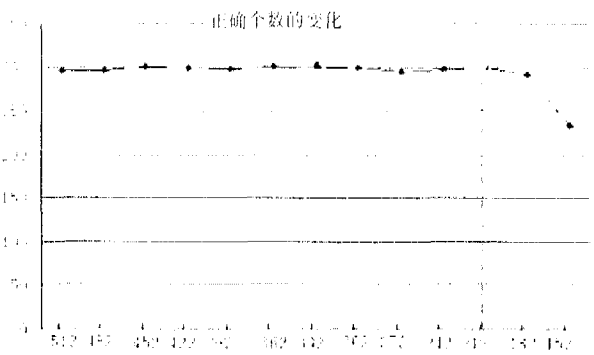

图 3 属性数目 - 正确识别个数变化图

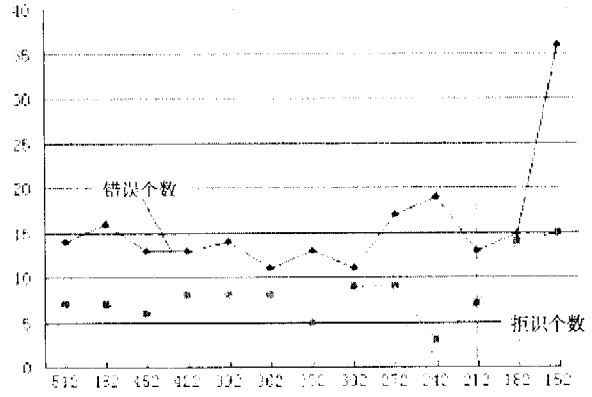

图4　属性数目－错识个数、拒识个数变化图

3.2　其他方法的结果

对同一实例用 Rough Set 属性约简方法进行特征选择，512维特征经过约简后得260维特征组，对新的特征组合再用覆盖算法进行识别，其结果为：

正确：299　错误：6　拒识：15

正确率：299/(299+6)=98%

识别率：305/320=95.3%

与FSACA法相比，主分量维数增加了48维，且运算时间大大延长，是FSACA法的50多倍，计算量十分巨大。未约简的48维是因为直接用一维一维的剔除法不能将有一定关联关系的冗余特征去除，而要在 Rough Set 的属性约简中考虑关联属性的约简，则计算量的增加是几何数量级的。

4　结　论

FSACA法是实现特征选择分析的简单而有效的算法。文中引入一个新颖的特征重要性判据，建立了特征选择规则与准确的点对之间距离之间的关系，并说明了FSACA法算法事实是将覆盖算法得到的覆盖领域进一步优化，使得分类空隙最大化。

FSACA法相对其他方法具有判距简单、运算量小、构造性强、直观等特点，适于处理大规模分类或聚类问题中的特征选择。然而，对于筛选点对的涉及到的参数还没有找到好的确定方法，未来工作中将继续对于这方面进行研究与实验。

参考文献：

[1] 边肇祺，张学工. 模式识别(第2版)[M]. 北京：清华大学出版社，1999.176－212.

[2] Zhang Ling, Zhang Bo. A Geometrical Representation of McCulloch－Pitts Neural Model and Its Applications[J]. IEEE Trans on Neural Networks, 1999, 10(4): 925－929.

[3] 孙即祥，王晓华，钟　山，等. 模式识别中的特征提取与计算机视觉不变量[M]. 北京：国防工业出版社，2001.9－103.

[4] 张　铃，张　钹. 多层反馈神经网络的FP学习和综合算法[J]. 软件学报，1994，8(4)：252－258.

[5] 张　铃，张　钹. 多层前向网络的交叉覆盖设计算法[J]. 软件学报，1999，10(7)：737－742.

[6] 张　铃，张　钹. 神经网络的规划学习算法[J]. 计算机学报，1994，17(9)：669－675.

[7] Zhang Ling, Zhang Bo. Relational Between Support Vector Set and Kernel Functions in SVM[J]. Journal of Computer Science & Technology, 2002, 17(5): 549－555.

[8] 张燕平，张　铃，吴　涛. 机器学习中的多侧面递进算法MIDA[J]. 电子学报. 2005，33(2)：327－331.

(上接第83页)

nice=10

disable=no

}

同时确保把 GLOBUS_LOCATION 替换为环境中的实际值。

(3)重启 Inetd/Xinetd。

运行/etc/init. d/xinetd reload，重新启动以便加载 gridftp 服务。

(4)测试。

创建代理证书：%grid－proxy－init －verify －debug，完成安全认证。

新建一个文件/tmp/file1，运行下面命令：

%globus－url－copy gsiftp://localhost/tmp/file file:///tmp/file2

执行成功，就会在/tmp/下找到 file2 文件。

4　GridFTP 的应用前景

由于 GridFTP 的各项特性符合网格数据传输的要求，随着网格应用的发展，GridFTP 也得到了不断的扩展。Globus Toolkit 每个新的版本都将新的功能和特性添加到 GridFTP 中，使其更加完善，从而提供更好的数据传输服务。

参考文献：

[1] 肖　侬. 编织数据网格——实现数据网格的关键技术[N]. 计算机世界报，2002－10－21(9,10)

[2] Allcock B, Bresnahan J, Kettimuthu R, et al. The Globus Striped GridFTP Framework and Server. IEEE/ACM Super Computing Conference 2005[EB/OL]. http://www.globus.org/alliance/publications/papers.php, 2005－11.

[3] Aloisio G, Cafaro M, Epicoco I. Early experiences with the GridFTP protocol using the GRB－GSIFTP library[J]. Future Generation Computer Systems, 2002, 18: 1053－1059.

[4] Allcock B, Bester J, Bresnahan J, et al. Data management and transfer in high－performance computational grid environments[J]. Parallel Computing, 2002, 28: 749－771.

[5] Globus. Toolkit 3.2: Installation Guide, Configuring GridFTP－Basic[EB/OL]. http://www－unix.globus.org/toolkit/docs/3.2/installation/install_config_gridftp.html, 2003.