一个新的多分类器组合模型

蒋林波[1],蔡立军[1,2],易叶青[2]
JIANG Lin-bo[1],CAI Li-jun[1,2],YI Ye-qing[2]

1.湖南大学 计算机与通信学院,长沙 410082
2.湖南大学 软件学院,长沙 410082
1.College of Computer and Communication,Hunan University,Changsha 410082,China
2.Software School,Hunan University,Changsha 410082,China
E-mail:jxfgt_005@163.com

JIANG Lin-bo,CAI Li-jun,YI Ye-qing.New model of combining multiple classifiers.Computer Engineering and Applications,2008,44(17):131-134.

Abstract: Classification is a very important part in the domain of data mining,however,single classifiers have many defects,such as very finite applicability and low accuracy.Combining multiple classifiers can overcome the defects.The existent combination rule,which is the pivotal conception during the process of combination,including product rule,sum rule,median rule,voting rule and so on,but these are not steady enough.In this paper,the authors develop a new model of combining multiple classifiers based on nerve net,and it is proved that it can improve not only the accuracy of classification but also its applicability.
Key words: data mining;classification;nerve net;combining multiple classifiers

摘 要:分类在数据挖掘中扮演着很重要的角色,然而单个分类器有很多缺点,包括适用范围十分有限和分类准确度不高等。把多个单分类器的分类结果融合起来是克服这些缺点的有效途径,因此存在很高的研究价值。组合多分类器的一个核心内容是融合规则,现存的融合规则有积规则、和规则、中值规则与投票规则等,但这些规则性能还不够稳定。提出了一个新的基于神经网络的融合规则,并依此建立一个新的多分类器组合模型,实验表明它能提高分类准确度和稳定性。
关键词:数据挖掘;分类;神经网络;组合多分类器
DOI:10.3778/j.issn.1002-8331.2008.17.039 **文章编号:**1002-8331(2008)17-0131-04 **文献标识码:**A **中图分类号:**TP311

1 引言

数据挖掘的主要任务有分类分析、聚类分析、关联分析、序列模式分析等,其中的分类分析由于其特殊地位,一直是数据挖掘研究的热点。一个具体样本的形式为:(v_1,v_2,…,v_i,…,v_n;c),其中 v_i 表示字段(属性)值,c 表示类别。数据挖掘分类的任务就是训练一个分类器,分析输入的样本集合,通过在训练集中的数据表现出来的特性,为每一个类找到一种准确的描述或者模型,这种描述常常用谓词表示。由此生成的类描述用来对未来的测试数据进行分类,尽管这些未来的测试数据的类标签是未知的,仍可以由此预测这些新数据所属的类[3]。

许多分类方法和技术可以用于构造分类模型,例如决策树、决策表、神经网络、k-最近邻、遗传算法、贝叶斯方法以及支持向量机等。然而,这些单一的分类技术在应用中常常会受到一定条件的限制,因此寻求能广义上提高分类性能的方法成为分类算法的一个研究方向,构造一个好的组合分类器已成为当前分类数据挖掘的研究热点和难点之一。

组合多分类器就是通过某种组合技术,将多个分类器的预测结果进行融合,从而产生一个新的分类器,并用新分类器对样本进行分类。如果融合得当,组合分类器的性能比任何单个分类器都优越[1]。最近 10 年,组合多分类器在现实研究各领域中取得了重大成果,比如字迹和文本识别[2],银行借贷风险预测[4],生物种群分类[5]等等。分类器融合方法一般分为并联和串联两种方式,这篇文章中,将提出一种新的基于 BP 神经网络的并联融合算法,这个算法利用各分类器的度量级输出信息和模型的决策误差自动调整各单分类器与类别间的权值。

2 组合多分类器概述

尽管一个实际的单分类器的最终输出是一个单值 j(第 j 个类别),但实质上很多分类器可以提供更多信息,只是这些信息被遗弃罢了。比如贝叶斯算法能提供每个样本 X 属于某类别的概率,而在新样本上利用它所取得的结果,其实就是它计算出来的具有最大概率值相对应的那个类标签。

基金项目:湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.06JJ20049,No.07JJ5085)。
作者简介:蒋林波(1985-),男,硕士研究生,主要研究方向:机器学习与数据挖掘;蔡立军(1964-),男,博士,教授,主要研究方向:机器学习、计算机网络、基因分类等。
收稿日期:2007-09-14 **修回日期:**2007-12-03

一般把分类器 E 的输出信息分为三个等级[6]：

(1)抽象级：E 仅输出类标签 j，或可能的类标签子集 J(类别集的真子集)；

(2)排列级：E 把类别集或 J 中的类标签根据一个内部规则排列起来，排在第一位的即为 X 所属样本类别的首选；

(3)度量级：E 输出在给定样本 X 各属性值时，X 属于每一类别的概率 $P(i/X)$(i=1,2,…,M)或与概率相对应的某些值。

很显然，分类器所输出的信息级别越高，它所包含的信息越全面。如果能充分利用这些被遗弃的信息，无疑会使分类准确度大大提高。然而，单个分类器由于缺少信息融合机制，无法把这些信息综合起来对样本进行类别预测，它们只能在这些信息里做出简单的选择。因此，组合多分类器的概念应运而生，有关它的研究和发展也得到许多学者的重视。

多分类器组合方式最突出的优点是模型可以综合不同分类器所得到的分类信息，避免单一分类器可能存在的片面性，以达到更好的分类效果。一般化的组合多分类器模型如图 1 所示。

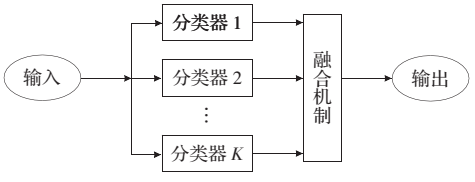

图 1 一般化的多分类器组合模型

如果把组合多分类器看作一个完整系统，则它由系统输入、单分类器设计、组合结构和融合规则四部分组成[7]。系统输入指输入的表示方式及单个分类器输入的确定，在一般化的模型中，主要是针对单分类器的输入，因为一般情况下，在接受一个输入时，每个单分类器都要得到它的独立结果；单分类器设计是指各个分类器学习算法的构造和相关参数的定义，它涉及到单分类器的协作状况，选择的单分类器一般要尽量差异化，以便各分类器的优缺点互补；组合机构是各单分类器的组合方式，它有并联和串联两种类型，对一个新样本，并联总是把所有单分类器的结果都并行融合起来，这样几乎总是能得到此样本属于某类别的相对概率，然后输出可能性最高的那个类别，而串联方式是把一系列分类器前后相接，后面的分类器注意力集中到它前面分类器所发生的预测错误上，通过训练使之成为一个有效整体；融合规则是各单分类器输出信息的组合方式，它是整个模型的核心，根据融合规则所采用不同的各单分类器输出信息的级别，融合规则也相应的分为抽象级、排列级和度量级，一般来讲，利用度量级所得到的分类结果要优于其他级别。一旦上面四个部分被确定，那么一个完整的组合分类器系统也就确定了。

3 度量级信息计算方法

分类器输出度量级信息对融合多分类器最有利，但并不是所有的分类器都像贝叶斯那样能直接输出度量级信息的，因此，有必要寻找其他的计算和表示方法。本文使用相似度和混淆距阵的概念，以便把不能直接输出度量级信息的分类器所产生分类结果度量级化。相似度和后验概率都给出了在给定单分类器条件下样本属于某类别的相对概率。

3.1 相似度

相似度是针对基于距离的分类器的，即根据某种距离(如欧氏距离，见式(3)对未知样本 X 进行分类的算法。设 f 表示距离分类器 K-Means，$d(X,i)$表示由模式的特征确定的 X 与类别 C_i 的中心的距离(i=1,2,…,M)，那么 X 与类别 C_i 之间的相似度 $s(X,i)$定义如下：

$$s(X,i)=\begin{cases}1 & d(X,i_0)=0 \text{ 且 } d(X,i)=0\\ 0 & d(X,i_0)=0 \text{ 且 } d(X,i)\neq 0\\ \dfrac{d(X,i_0)}{d(X,i)} & d(X,i_0)\neq 0\end{cases} \tag{1}$$

其中 $d(X,i_0)$是所有 $d(X,i)$的最小值。相似度 $s(X,i)$反映了未知样本 X 在 K-Means 算法基础上属于 C_i 类的可能性，此值越大，X 属于 C_i 的概率也就越大[6]。

对于 K-NN 分类器，本文对以上定义稍作修改，令 $d(X,i)$表示 X 与它 K 个近邻中属于类别 C_i 的所有样本的平均距离，即 $d(X,i)=\sum_{K_i} d(X,X_j)/K_i$($j$=1,2,…,$K_i$；$X_j$ 是属于类别 C_i 的 X 的近邻；K_i 是这些近邻的个数)，$d(X,i_0)$同样表示所有 $d(X,i)$的最小值，那么有：

$$s(X,i)=\begin{cases}1 & d(X,i_0)=0 \text{ 且 } d(X,i)=0\\ 0 & d(X,i_0)=0 \text{ 且 } d(X,i)\neq 0,\text{或者 } K_i=0\\ \dfrac{d(X,i_0)\cdot K_i}{d(X,i)\cdot K} & d(X,i_0)\neq 0 \text{ 且 } K_i\neq 0\end{cases} \tag{2}$$

3.2 混淆距阵及后验概率

$E(X)=j$ 表示分类器 E 将来自 X 的模式分到类 C_j 中，为了了解分类器的识别情况，可以先对分类器进行训练，然后利用测试样本集来计算此分类器在具体数据集上的混淆距阵 $\boldsymbol{CE}_{M\times N}$($M$ 为类别数，由于存在拒识的情况，所以 N 比 M 大 1)，混淆距阵反应了此单分类器对样本集的识别情况，CE_{ij} 表示分类器 E 将类 C_i 中的样本识别成 C_j 的数量。若 $i=j$，则 CE_{ij} 为 E 正确识别 C_i 类中样本的数量，否则为错误识别的数量。那么对 E 而言，所有样本被识别的结果为 $j=E(X)$的总数为 $N_j=\sum_i CE_{ij}$。因此，在 E 的识别结果为 j 的条件下，样本来自 C_i 类的概率可以用条件概率表示为：

$$P_{ij}=P(X\in C_i|E(X)=j)=\frac{CE_{ij}}{N_j} \tag{3}$$

对某个输入样本，单分类器对其有一个结果输出 j，而利用式(3)结合以前的经验，总是可以得到这个样本属于某个类别的条件概率。因此，进而可以在测试集的基础上建立一个与此分类器相关的后验概率距阵 $\boldsymbol{W}_{N\times M}$ (M 为类别数，N=M+1)，$\boldsymbol{W}_{ij}$ 表示在单分类器识别样本为类别 C_i 的前提下，样本实质上属于 C_j 的概率。这样，当分类器判别出新样本的类别后，可以用输出新样本属于各类别的后验概率的方式来模拟它真实属于某类别的概率，这是用得最广泛的获得度量级信息的一种方式。

4 新的多分类器组合模型

组合多分类器之所以比单分类器具有更好的性能，是因为它融合各单分类器判别信息的同时，实现了各分类器之间优缺点的互补。既然如此，在组合多分类器的过程中，要求研究者使各单分类器尽可能差异化、融合规则尽可能全面化和精确化。

在以前的很多研究中,人们使用相同类型的分类算法作为单分类器,而选取不同的参数、不同的样本或特征子集去满足单分类器差异化的需求,这虽然对分类性能有所提高,但提高幅度不大,因为这种方式在一定程度上是以降低其中一些单分类器的分类效果为代价的。也有研究者使用不同类型的分类器来进行组合,但他们对单分类器输出度量级信息的计算方式比较单一(最常用的是混淆距阵),有的甚至直接使用各单分类器的抽象级信息,并且在融合规则的选择上使用相对比较简单的投票规则等,在很大程度上影响了系统分类的准确度和稳定性。

本文提出一种新的基于神经网络的多分类器组合模型(CMCBOB)。本文不采用传统的特征子集训练方式,而是使用不同的训练集去训练不同类型的分类器,然后把这些不同的训练集整合起来对融合规则进行训练,这样不但有利于统一各单分类器的输入模式,简化系统的实现和使用,而且在不降低单分类器在数据集上的分类性能的前提下尽量加大了各单分类器的差异,有利于各单分类器的互补,扩大了组合模型的适用范围。与此同时,在计算单分类器的输出上,对每一样本使用不同的计算方式:能直接输出度量级信息的(贝叶斯)就直接使用其输出信息;基于距离的单分类器(例如 K-Means、K-NN)则输出样本对各类别的相似度;而其他的单分类器则输出由混淆距阵得出的样本的后验概率。最后,我们不用传统的积规则、和规则等去融合各单分类器的输出信息,而是在神经网络的基础上利用系统的决策误差调整各单分类器与类别间的权值,实现了融合规则的自动调整和优化。很显然,以上几点说明本文所提出的组合分类器模型从理论上讲具有大大提高分类准确度和稳定性的潜能。

4.1 模型说明

神经网络本身也是一个分类器,它是一组连接的输入/输出单元,其中每个连接都与一个权值相联。在学习阶段,通过调整神经网络的权来预测样本的正确类标号。它虽然需要大量靠经验确定的参数,解释性也比较差,但它对噪声有很高的承受能力,分类能力也相当好,所以到了广泛应用。本文并不把神经网络仅仅当作一个简单的分类器,而是充分利用它对噪声的高承受能力的优点,把它当作一个模型架构来对融合规则进行训练。

本文的模型建立在三层神经网络基础之上,它由输入层、中间层和输出层组成,它与单个神经网络分类器的区别在于:(1)输入层到中间层的数据不需要经过加权操作,因为它作为一个统一输入模式,在各单分类器中会得到相应的处理;(2)隐含层(即中间层)不是一般意义上的结点,而是一个个独立的单分类器,因此,他们的输出也不同于普通结点的单一输出,而是一个向量;(3)同一个中间层结点对输出层的不同结点净输出也不同。新的组合多分类器模型如图 2。

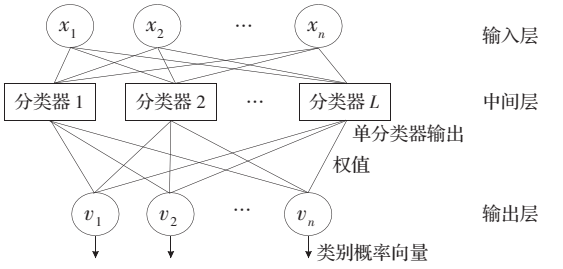

图 2 新的多分类器组合模型

输入层的 n 个结点对应样本集的 n 个特征,它把原始值直接输入到各分类器;中间层为 L 个不同类型的单分类器,每个分类器接受系统的统一输入模式,但对不同的输出层结点的输出是不同的,即它的输出是一个长度与类别数目相等的向量,分量大小为此单分类器判定样本属于对应类别的概率(各分类器的度量级输出),这些概率值在预先设定规则下与权值进行运算,所得结果作为输出层的输入;输出层每个结点的输出是一个在[0,1]上的小数,它由这层的输入经过一个赋活函数计算所得,表示系统判定输入样本属于某类别的概率,它是组合分类器的最终输出,其中最接近 1 的那个分量对应着样本被预测的类别。

本文把数据集分为训练集和测试集两部分,其中训练集又分为 L 部分,分别用来训练 L 个单分类器,然后整个训练集用来训练组合分类器模型中中间层到输出层的权值,权值由输出层的实际输出组成的向量与期望输出向量(大小为 M,有且仅有一个分量为 1,其他为 0)之间的差别来矫正。在训练组合模型的时候采用在线学习的方式,为了得到更精确的输出结果,可以进行许多个训练周期。

4.2 模型实现

前文已经确定了系统输入、组织结构和融合规则的计算方式,为了加强单分类器之间的差异,同时充分利用不同的度量级信息计算方法,选用朴素贝叶斯(NB)、K-中心点(K-Means)、K 近邻(K-NN)和支持向量机(SVM)作为单分类器。

4.2.1 NB

设每个数据样本用一个 n 维特征向量来描述 n 个属性的值,即:$X=\{x_1,x_2,\cdots,x_n\}$,有 m 个类,分别用 $C_1,C_2,\cdots,C_m$ 表示。给定一个未知的数据样本 X(即没有类标号),若朴素贝叶斯分类法将未知的样本 X 分配给类 C_i,则一定有 $P(C_i|X)>P(C_j|X)(1\leqslant j\leqslant m,i\neq j)$,即在给定样本数据的情况下,NB 虽然有能力得出一组度量级信息,但它总是给出后验概率最大的那个类别标签。根据贝叶斯定理:

$$P(C_i|X)=P(X|C_i)\cdot P(C_i)/P(X) \tag{4}$$

由于 $P(X)$对于所有类都为常数,最大化后验概率 $P(C_i|X)$ 的工作可转化为最大化先验概率(即式(4)中的分子部分)。另外,在给定一定量的训练集的情况下,$P(C_i)$也是一个常数,所以我们只需要计算每个 $P(X|C_i)$,但是在一般情况下训练数据集有许多属性和元组,计算 $P(X|C_i)$的开销可能非常大,为此,通常假设各属性的取值互相独立,这样就有:

$$P(X|C_i)=\prod_j P(x_j|C_i)\quad j=1,2,\cdots,n \tag{5}$$

根据此方法,对任意未知类别的样本 X,就可以计算出 X 属于每一个类别 C_i 的概率。

4.2.2 K-Means

K-平均算法(K-Means)将研究对象按照相似性准则划分到若干个子集中,使得相同子集中各元素间差别最小,而不同子集中各元素间差别最大。通常的空间聚类算法是建立在各种距离基础上的,其中最常用的就是欧几里德距离:

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2} \tag{6}$$

其中,$i=(x_{i1},x_{i2},\cdots,x_{in})$和 $j=(x_{j1},x_{j2},\cdots,x_{jn})$是两个 n 维的数据对象,$d(i,j)$则表示了两个样本数据之间的欧几里德距离。本文

的策略是:先用 K-Means 对样本数据进行聚类,得到与类别数量相等的 M 个簇,然后把各簇里的奇异点去掉,分别计算各簇所剩样本的平均值,作为类别中心。对任意未知类别的样本 X,用式(1)求得它属于某类别的概率。

4.2.3 K-NN

给定一个未知样本 X,K-NN 分类算法搜索模式空间,找出最接近 X 的 K 个训练样本,这 K 个样本即为 X 的 K 个近邻(临近性用距离公式(6)定义)。然后,根据定义的式(2)计算 X 属于各类别的概率。

4.2.4 SVM

支持向量机(Support Vector Machine,SVM)的核心思想是:对于输入空间中非线形可分的情形,选择一个适当的非线形映射,将输入空间中的样本点映射到一个高维的特征空间,使得对应的样本在该特征空间中是线形可分的。

设训练样本为(x_i,y_i),$i=1,\cdots,n$,$x\in R^d$,$y\in\{+1,-1\}$为样本所属的类别标号,d 为样本特征空间的维数。最优分类面就是满足式(4),且分类间隔 $2/\|w\|$ 最大,即 $\|w\|$ 最小。

$$y_i[(w\times x_i)+b]-1\geqslant 0,i=1,2,\cdots,n \tag{7}$$

本文利用它在训练集上的分类结果产生混淆距阵,进而得到 SVM 在数据集上的后验概率距阵,从而使得 SVM 能输出新样本 X 属于各类别的概率。

4.3 算法描述

模型输入为样本对象 $X=\{x_1,x_2,\cdots,x_n\}$;o_{ij} 表示第 i 个分类器判断输入对象属于第 j 类的概率,也是第 i 个中间层结点对第 j 个输出层结点的输出值;w_{ij} 表示第 i 个中间层结点与第 j 个输出层结点连线的权值;输出层的净输入由式(5)决定:

$$I_j=\sum_i w_{ij}\cdot o_{ij}+u_j \tag{8}$$

其中 u_j 是单元 j 的偏值。当 I_j 进入输出层结点后,通过一个赋值函数来求得输出层结点的输出,赋活函数取平均值函数,即 I_j/L(L 为中间节点个数)。另外,令 $V(v_1,v_2,\cdots,v_m)$为模型的实际输出向量,它的分量 $v_i(i=1,2,\cdots,M)$即为组合分类器最终判定输入样本属于类别 C_i 的概率。

为了更新权和偏值,需要计算模型预测的误差,使用式(6)来计算输出层第 j 个结点的误差:

$$Err_j=v_j(1-v_j)(t_j-v_j) \tag{9}$$

其中 t_j 是模型期望的输出,当训练样本的真实类标号为 j 时,则它取值为 1,否则为 0。以 $\Delta w_{ij}=k\cdot Err_j\cdot o_{ij}$ 来更新权值 w_{ij},$\Delta u_j=k\cdot Err_j$ 来更新 u_j,其中 k 为学习率,通常取 0 和 1 之间的一个常数,一个经验规则是取训练周期的倒数。

具体步骤如下:

(1)把样本总体分为训练集和测试集两部分,对训练集进一步细分为 L 组分别用来训练 L 个单分类器;确定 K-NN 中的 K 值和 K-Means 中各聚类中心,并计算出 SVM 的混淆距阵和后验概率距阵;

(2)指定学习率 k,初始化各权值 w_{ij} 和偏值 u_j;

(3)对训练集中的每个样本 X:

(4)根据相应的度量级信息计算方法计算分类器 i 在输入 X 的情况下对输出层结点 j 的输出 O_{ij},并计算输出层结点 j 的净输入 $I_j=\sum_i w_{ij}\cdot O_{ij}+u_j$ 和输出 $v_j=I_j/L$;

(5)计算输出层结点 j 的误差 $Err_j=v_j(1-v_j)(t_j-v_j)$;

(6)修正权值 $w_{ij}=w_{ij}+\Delta w_{ij}$;

(7)修正偏值 $u_j=u_j+\Delta u_j$,转(3);

(8)当周期数达到 $1/k$,结束。

此模型中,w_{ij} 可以看作是系统中分类器 i 对类别 j 的敏感程度,它越大,说明分类器 i 对类别 j 的识别能力越强。u_j 是模型对类别 j 的偏见度,它一般与具体的特定领域数据有关。这里,可以把 w_{ij} 一律初始化为 1,把 u_j 初始化为类别 j 在训练集中所占的比例,并指定 k 为 0.02,即模型将学习 50 个周期。

5 实验分析

标准 UCI 数据集是 UCI(University of California Irvine)整理的一个机器学习试验专用数据集,公布在 http://www.ics.uci.edu/~mlearn/MLRepository.html,数据集样本分为三类。文献[9]的思想是在标准 UCI 数据集上计算三个 SVM 的混淆距阵,然后利用这些混淆距阵去训练各个 SVM 的权值,以此来构建组合分类器(MASWOD)。文章把各个单分类器在各个类别中的预测准确度列出,并把组合分类器的准确度与传统的贝叶斯分类算法做了比较,如表 1。

表 1 文献[9]的试验结果

类别	算法				
	SVM1	SVM2	SVM3	MASWOD	Bayesian
	准确率				
类别 1	70.2	68.6	59.4	75.3	69.8
类别 2	60.4	56.9	62.3	73.6	69.5
类别 3	92.8	89.6	78.3	94.3	93.6

本文把这个标准 UCI 数据集分成两大部分,其中第一部分(U1)用于训练,第二部分(U2)用于测试。UCI 共有 8 080 个样本,不妨令$|U1|=7\,000$,$|U2|=1\,080$。$U1$ 的 7 000 个样本随意分为四个部分$|L1|=3\,000$,$|L2|=2\,000$,$|L3|=1\,000$,$|L4|=1\,000$ 分别用来训练 NB、SVM、K-MEANS、K-NN 四个单分类器,并在子测试集上获得各单分类的预测准确率。然后依照模型训练权值矩阵和偏值向量,获得本文组合模型(CMCBOB)的预测能力,比较结果如表 2。

表 2 本文的实验结果

类别	算法					
	NB	K-Means	K-NN	SVM	CMCBOB	MASWOD
	准确率					
类别 1	70.4	65.2	64.3	69.3	84.6	75.3
类别 2	69.5	68.4	67.2	58.6	83.8	73.6
类别 3	92.8	80.5	78.3	80.5	94.3	94.3

由表 1 和表 2 可以看出,组合多分类器系统的分类准确度比任意一个构成它的单分类器的分类准确度都要高出很多,这体现了组合分类器的明显优势。同时也可以看出,提出的组合多分类器模型 CMCBOB 要比文献[9]的 MASWOD 分类准确度更高,类别的偏见度更小,即系统对每个类别分类的准确度之间的差异变小了,这充分说明了 CMCBOB 高的准确性和稳定性。

6 结论

本文提出了一个新的基于神经网络的多分类器组合模型(CMCBOB),在神经网络的结构上引进用相似度和混淆距阵计

(下转 147 页)