第34卷 第24期 计 算 机 工 程 2008年12月
Vol.34 No.24 Computer Engineering December 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)24—0218—03 文献标识码：A 中图分类号：TP391.4

集成学习的多分类器动态组合方法

陈 冰，张化祥
(山东师范大学信息科学与工程学院，济南 250014)

摘 要：为了提高数据的分类性能，提出一种集成学习的多分类器动态组合方法(DEA)。该方法在多个UCI标准数据集上进行测试，并与文中使用的基于Adaboost算法训练出的各个成员分类器的分类效果进行比较，证明了DEA的有效性。
关键词：多分类器；聚类；动态分类器组合；Adaboost算法

Dynamic Combinatorial Method of Multiple Classifiers on Ensemble Learning

CHEN Bing, ZHANG Hua-xiang
(College of Information Science and Engineering, Shandong Normal University, Jinan 250014)

【Abstract】 In order to improve the classification performance of dataset, a dynamic combinatorial method of multiple classifiers on ensemble learning DEA is proposed in the paper. DEA is tested on the UCI benchmark data sets, and is compared with several member classifiers trained based on the algorithm of Adaboost. In this way, the utility of DEA can be proved.
【Key words】 multiple classifiers; clustering; dynamic classifier ensemble; Adaboost algorithm

1 概述

近年来，多分类器组合(DEA)技术在各个领域已经得到了广泛的应用，如模式识别中的人脸识别、网络安全、语言学中的词义消歧[1]等。

关于多分类器系统的研究越来越多，大量的理论和实验结果表明，通过多分类器组合不但可以提高分类的正确率，而且能够提高模式识别系统的效率和鲁棒性。尽管在各个方面提出了不同的分类器组合方法，但这些方法都或多或少地存在某些缺陷，它们或者先利用聚类对数据集进行处理，再直接用同种类型的分类器来分类[2]；或者采用不同类型的分类器，而不对数据集做任何处理[1]；更多的是利用不同的融合算法来训练生成同种类型的分类器，再利用它们对数据分类。另外，通常所使用的分类方法如决策树、K-近邻、Bayes等都是有导师信息的机器学习过程。但实际中存在着大量的数据没有标记样本类别，如果再运用这些分类方法，其操作性就比较差了。而聚类等非监督学习能自适应地处理大量的未知类别的样本。基于监督学习与非监督学习的优势互补，将两者结合起来各取所长，一定能够收到很好的效果。另外值得注意的一点：目标识别中利用不同的分类器可以得到不同的分类识别结果，而且结果之间具备相当的互补性，从而可以提高分类的效果，克服单分类器存在的问题。

2 多分类器动态组合流程

图1是DEA方法一次随机取样的流程。这里，小样本集1,2,…,k是对训练数据集按照类别标号得到的k个小集合；分类器组合 1,2,…,k 表示的是由训练数据集训练出的分类器对每个小样本集合分类根据分类错误率得到的k组性能较好(错误率较低)的分类器组合。其中，总的分类器是在 Adaboost 基础上每次随机地生成以决策树、贝叶斯、k-近邻中的一个作为基分类器，直到生成50个为止。接下来利用这k组分类器去分类类别标号相对应的测试数据中的聚类集合(为了表示的方便，图中假设小样本集与聚类集合是一一对应的)。最后用每个聚类集中被错误分类的样本数之和除以测试数据总数，即得一次采样的错误率。

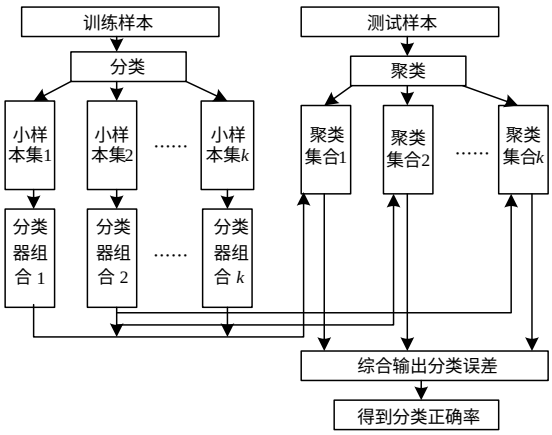

图1 多分类器动态组合流程

3 多分类器动态组合

3.1 集成学习

集成学习[3]方法是根据样本训练多分类器来完成分类任务的方法，这些分类器具有一定的互补功能，在减少分类误

基金项目：山东省科技攻关计划基金资助项目(2005GG4210002)；山东省青年科学家科研奖励基金资助项目(2006BS01020)；山东省教育厅科技计划基金资助项目(J07YJ04)；山东省自然科学基金资助项目(Y2007G16)
作者简介：陈 冰(1981－)，女，硕士研究生，主研方向：数据挖掘，机器学习；张化祥，教授、博士
收稿日期：2008-04-14 **E-mail：**zyxcscb@163.com

差上表现比较好。其中，Adaboost算法[4]就是一个比较成功的集成学习算法。

Adaboost 算法的形式化描述如下：

假设一个特征空间 X，二分标签空间 Y={-1,+1}和一系列的训练样例 $S=\{(x_i,y_i)\mid x_i\in X,y_i\in Y,i=1,2,\cdots,n\}$。

Adaboost 首先从训练数据集 S 中，以一个均衡的权向量 $\boldsymbol{w}_1(1)=\boldsymbol{w}_2(1)=\cdots=\boldsymbol{w}_1(n)=1/n$ 作为概率，取 n 个训练样本组成训练集合，用决策树桩作为分类器基本学习算法，训练出第一个成员分类器 h_1。

在 t 时刻，用创建的基于权向量 $\boldsymbol{w}_t$ 的分类器 $h_t()$ 对原数据集 S 进行分类并且考虑权向量训练错误 ε_t ：

$$\varepsilon_t=\sum_{i=1}^{n}w_t(i)I(y_i\neq h_t(x_i)) \tag{1}$$

其中，当事件 E 出现时，$I(E)=1$，否则等于 0。

再设

$$\beta_t=\frac{1}{2}\log_e\frac{1-\varepsilon_t}{\varepsilon_t} \tag{2}$$

然后用下面规则更新权向量：

$$\boldsymbol{w}_{t+1}(i)=\boldsymbol{w}_t(i)\exp\{-\beta_t y_i h_t(x_i)\}/Z_t \tag{3}$$

其中，Z_t 是一个归一化常数。

这样实现的效果是，对每一个样例，如果前一次的分类器将它错分了 $(y_i h_t(x_i)<0)$，就增加该样例的权值，反之则减少它的权值。

上面的过程被重复 T 个循环，最后用

$$\alpha_t=\beta_t\Big/\sum_{t=1}^{T}\beta_t \tag{4}$$

作为成员分类器的投票权值得到联合分类器。显然，那些有低训练错误的带权分类器在联合中将拥有高的投票权向量。

基于 Adaboost 算法，DEA 采用不同的分类器算法训练出不同类型的成员分类器，而且分类器的生成是随机的，这样能进一步增加分类器的差异性，由于各个分类器对不同的数据集有不同的偏重，因此利用这种差异的互补性可以提高分类的性能。

3.2 经典聚类——k-均值(k-means)

k-均值[5]算法的基本思想是：给定类的个数k，将n个对象分到k个类中，使得类内对象之间的相似性最大，而类之间的相似性最小。相似度的计算根据一个聚类中对象的平均值(被看作聚类的中心)来进行，即每个簇用该簇中对象的平均值来表示。

DEA 利用 k-均值对测试数据聚类，并利用欧氏距离找出测试数据聚类集与训练数据小集合之间的对应关系。

3.3 最小个体错误(MIE)

在针对训练样本的局部区域(小样本集)选择具有最好局部分类准确率的分类器组合时，采用了最小个体错误(MIE)[6]的思想，即对每一个小区域，利用训练生成的N(50)个分类器产生的分类错误率进行排序，选出错误率最小的前n(10)个分类器，用于对测试数据集中对应的小聚类集进行测试。

3.4 分类器动态组合方法(DEA)

3.4.1 DEA 的思想

训练数据按类别分成 k 个小集合(k 表示类别数)，再用 k-means 方法对测试数据聚类，把测试数据也聚类成 k 个聚类集。然后对这 k 个聚类集与训练数据分成的 k 个小集合找出相互之间的对应关系。方法如下：

假设训练数据的 k 个小集合表示为 T_i (i=1,2,⋯,k)，测试数据的 k 个聚类集表示为 t_j (j=1,2,⋯,k)。

对每一个 t_j 的聚类中心 tc_j (j=1,2,⋯,k)

begin

对每一个 T_i 的聚类中心 Tc_i (i=1,2,⋯,k)

$$dis_{ji}=\min(d(tc_j,Tc_i)) \tag{5}$$

end

其中，$d(tc_j,Tc_i)$表示测试数据的第 j 个聚类集与训练数据的第 i 个小集合的聚类中心之间的距离。t_j 的类别标号即为 i。

最终找到测试数据的 k 个聚类集与训练数据的 k 个小集合的对应关系。

对于由训练数据随机训练得到的 N 个不同类型的分类器，由于训练数据分成了 k 个小集合，因此每个小集合可以得到 N 个分类错误率。根据 MIE 取出前 n 个错误率较低的分类器，每个小集合都将得到对应的 n 个分类器，然后利用它们去分类类别标号对应的测试数据的每个聚类集，得到错误率。

3.4.2 DEA 的性能评价

对测试数据的 k 个聚类集，假设如下：

第 1 个聚类集被错误分类的样本数量为 m_1 ；

第 2 个聚类集被错误分类的样本数量为 m_2 ；

……

第 k 个聚类集被错误分类的样本数量为 m_k 。

且假设测试数据总数为 M，则 DEA 的错误率为

$$error_{DEA}=\sum_{i=1}^{k}m_t\Big/M\ (i=1,2,\cdots,k) \tag{6}$$

3.5 平均错误(ME)

DEA[6]采用 4 折交叉验证，对每次得到的错误率 $error_{\mathrm{DEA}}$，计算获得最终的错误率，由于最后比较的是正确率，用 1.0 减去所得的错误率即可，从而获得DEA的分类性能。

4 实验及结果分析

对 DEA 算法以及在 Adaboost 基础上以决策树、贝叶斯、k-近邻为基分类器的 3 种算法在 UCI 标准数据集上所测得的正确率进行比较，实验结果如表 1 所示，由于考虑到运行速度的问题，数据集 letter1 是原 letter 中前 5 000 个实例。

表 1 数据集的正确率比较

数据集	正确率/(%)			
	DEA	Adaboost(J48)	Adaboost(NB)	Adaboost(IBk)
audiology	79.798 1	84.513 3	79.203 5	78.318 6
automobile	60.988 6	83.414 6	60.000 0	74.146 3
breastcancer-w	99.714 3	96.137 3	95.851 2	95.994 3
credit-g	95.050 0	75.400 0	75.400 0	72.400 0
glass	77.921 9	73.364 5	46.729 0	67.289 7
hayes-roth	53.787 9	74.242 4	75.757 6	61.363 6
heart-statlog	97.886 0	80.000 0	81.481 5	75.555 6
hepatitis	91.599 2	83.870 0	78.709 7	81.935 5
machine	95.709 4	87.081 3	79.904 3	87.559 8
labor	92.142 9	91.228 1	89.473 7	85.964 9
letter1	89.520 0	91.360 0	63.220 0	89.980 0
sonar	86.538 5	83.173 1	85.096 2	86.057 7
vehicle	71.750 1	78.487 0	45.981 1	69.267 1
artificial	78.185 7	60.246 6	30.182 0	56.821 3
clean1	65.336 1	92.016 8	83.193 3	85.504 2

实验中采用 Adaboost 算法作为集成学习方法，决策树、贝叶斯、k-近邻作为基分类器，使用的学习算法分别为 J48, NaiveBayes 和 IBk。利用随机数生成器在整个训练数据集上随机地生成 50 个不同类型的分类器，然后针对每个类别标号相同的小样本集根据误差率选择出 10 个较好的，用它们去分

类对应测试集中的聚类集。由于要对测试集进行聚类，考虑到小数据集的情况，DEA 采用 4 折交叉验证来生成随机的数据集。另外，由于分类器的生成是随机的，因此 DEA 选用 50 次循环求平均的方法来最终求得正确率。

4.1 实验数据分析

对于数据集的选择，在实验中选用了 15 个UCI标准数据集，这些数据集是从不同方面考虑[7]选择出来的，如：从数据集的大小考虑，较大的数据集如artificial, letter1，较小的数据集如labor；从数据集的属性个数来考虑，属性数量较多的数据集如clean1, sonar，属性数量较少的数据集如hayes-roth；从类数量的多少来考虑，类的数量较多的如letter1, audiology，类的数量较少的数据集如breastcancer-w, credit-g。

4.2 实验结果分析

由表 1 中的实验数据对 DEA 做如下分析：

(1)在这 15 个 UCI 数据集中，DEA 算法正确率较高的有 9 个，其他 3 种方法只有 6 个，且 DEA 算法获得的结果明显比其他 3 种好得多。由此可以非常容易看出 DEA 更适合于处理样本数较大、属性个数较多的数据集，也适合于处理含有较多数值性属性并且样本数量不是很大的数据集。

(2)DEA 的特点：对测试样本进行聚类；训练一定数量的不同类型的基分类；从训练出的多个分类器中选出少量性能较好的用于测试。综合 DEA 的这 3 个优点，使得它的优越性更加突出。

(3)DEA 采用多种不同类型的基分类器提高了分类的正确率，说明不同分类器之间的差异是互补的。但是从实验数据中也可以看出，基分类器的分类效果对 DEA 影响很大，如正确率差距很小的数据集 breastcancer-w, creat-g 等，DEA 的正确率较高；而分类效果差距很大的数据集如 glass, letter1, vehicle 等，DEA 的正确率就相对较低。因此，在选用基分类器时需要仔细考虑。

(4)由于 DEA 生成基分类器是随机的，为了能更准确地计算出测量结果，采用了多次测量求平均值的方法，因此不可避免地在时间上会有相当的消耗，但是考虑到正确率，时间复杂度问题应该可以忽略。

5 结束语

多分类器组合技术是近年来在分类器技术取得大量成果的基础上兴起的一种新的模式识别方法，以其优越的泛化能力在各个识别领域得到了广泛的应用。

(1)DEA 使用随机数生成器产生多种不同类型的分类器，并且从训练出的大量分类器中选出部分性能较好的用于测试，体现了分类器多样性的特点。

(2)DEA 采用了有导师学习与无导师学习相结合的思想，使得分类正确率明显提高。

(3)由于 DEA 采用了聚类这种无导师的分类方法，利用有类别标签的训练数据集的每个具有相同标签的小样本集训练产生性能较好的分类器，使用它们去分类测试数据，因此该方法可以用于处理没有类别标签的样本。

因此，在应用中要考虑分类器差异互补的特点，尽量使用多种不同类型的分类器。DEA 使用了常用的 3 种类型，也可以使用其他的基分类器方法，有关这些思想的实施有待于进一步的研究。另外，要考虑使用有导师与无导师的分类方法相结合。因为在实际应用中，要获得大量有类标签的样本比较困难、或者代价很高，所以考虑使用 DEA 方法，将会获得很大收益。

参考文献

[1] 全昌勤, 何婷婷, 姬东鸿, 等. 基于多分类器决策的词义消歧方法[J]. 计算机研究与发展, 2006, 43(5): 933-939.

[2] 刘汝杰, 袁保宗, 唐晓芬. 一种新的基于聚类的多分类器融合算法[J]. 计算机研究与发展, 2001, 38(10): 1236-1241.

[3] 方 敏. 集成学习的多分类器动态融合方法研究[J]. 系统工程与电子技术, 2006, 28(11): 1759-1761, 1769.

[4] Witten I H, Frank E. 数据挖掘实用机器学习技术[M]. 2 版. 北京: 机械工业出版社, 2006.

[5] Mitchell T M. 机器学习[M]. 北京: 机械工业出版社, 2006.

[6] Ruta D, Gabrys B. Classifier Selection for Majority Voting[J]. Information Fusion, 2005, 6(1): 63-81.

[7] Wang Xiao, Wang Han. Classification by Evolutionary Ensembles[J]. Pattern Recognition, 2006, 39(4): 595-607.

(上接第 217 页)

在时间复杂度上，第 1 种调整方法只需遍历一次区分矩阵中的所有项，区分矩阵最多有$|U|(|U|-1)/2$ 项，每项最多包含$|P|$个属性，因此，其时间复杂度的上界是$O(|P||\mathbf{U}|^2)$。第 2 种调整方法的合并项数目最多为$|(|U|-2)$，每项最多包含$|P|$个属性，因此，其时间复杂度的上界是$O(|P|^2|U|)$。所以，整个算法的时间复杂度上界为$O((|P|+|U|)|P||U|)$。可以看出，由于区分矩阵阶数的变化趋势是由大到小的，且区分矩阵每项属性个数的变化趋势是由多到少的，因此改进算法的效率比基于静态区分矩阵的约简算法要高，同时还能保证得到更优的约简。

6 结束语

本文提出一种基于动态区分矩阵的属性约简算法，该算法的特点体现在动态调整的区分矩阵上，它能及时地反映出相对当前的约简。此外，该算法提供了一种改进基于静态区分矩阵的约简算法的思路，能够进一步推广到更多的基于静态区分矩阵的约简算法上。

参考文献

[1] Pawlak Z. Rough Set[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 321-336.

[2] 马廷淮, 赵亚伟. 基于约简剪枝的属性约简算法[J]. 计算机工程, 2007, 33(18): 56-58.

[3] 袁晓峰, 许化龙, 陈淑红. 基于量子遗传算法的粗糙集属性约简新方法[J]. 计算机工程, 2007, 33(15): 184-186.

[4] 刘 山, 张 慧. 基于条件信息量的动态属性约简方法[J]. 计算机工程, 2007, 33(11): 182-183.

[5] Hu Xiaohua. Knowledge Discovery in Database: An Attribute Oriented Rough Set Approach[D]. Regina, Canada: University of Regina, 1995.

[6] 胡可云. 基于概念格和粗糙集的数据挖掘方法研究[D]. 北京: 清华大学, 2001.