多分类器选择集成方法

郭红玲,程显毅
GUO Hong-ling,CHENG Xian-yi

江苏大学 计算机科学与通信工程学院,江苏 镇江 212013
School of Computer Science and Telecommunications Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China
E-mail:guohongling2006@126.com

GUO Hong-ling,CHENG Xian-yi.Method of selective multiple classifiers ensemble.Computer Engineering and Applications,2009,45(13):186-187.

Abstract: Because of the high request to classifies performance of people and the implementation complexity of multiple classifiers ensemble approach,this paper proposes an new method of selective multiple classifiers ensemble which considers of the accuracy of individual classifier and diversity among individual classifiers.This algorithm first chooses the more accuracy classifies from the production base,then chooses more different ones using diversity measure before integration.The result of the UCI database experiment demonstrate that the method is better than the Bagging method,and it is very good and useful for classification.

Key words: multiple classifiers ensemble;diversity measure;classifiers selection

摘 要:针对目前人们对分类性能的高要求和多分类器集成实现的复杂性,从基分类器准确率和基分类器间差异性两方面出发,提出了一种新的多分类器选择集成算法。该算法首先从生成的基分类器中选择出分类准确率较高的,然后利用分类器差异性度量来选择差异性大的高性能基分类器,在分类器集成之前先对分类器集进行选择获得新的分类器集。在 UCI 数据库上的实验结果证明,该方法优于 bagging 方法,取得了很好的分类识别效果。

关键词:多分类器集成;差异性度量;基分类器选择
DOI:10.3778/j.issn.1002-8331.2009.13.054 **文章编号**:1002-8331(2009)13-0186-02 **文献标识码**:A **中图分类号**:TP311

近年来,随着计算机技术的发展,信息融合技术成为一种新兴的数据处理技术,并已经取得可喜的进展。在模式识别领域,以多分类器集成为代表的融合技术也受到越来越多的关注,在许多方面都得到了广泛的应用,如手写体数字和文字的识别、人脸识别、身份识别、语音识别、医疗诊断、地震预测闹、遥感图像识别、军事目标识别等[1]。

多分类器集成就是通过使用大量的基学习器来获得更好的识别性能。事实上,在多分类器集成学习的文献中,可以容易地找到很多工作使用了成百上千个分类器。但这一做法有一些负面影响,一方面,使用更多的分类器将导致更大的计算和存储开销,另一方面,当基分类器数目增加之后,分类器之间的差异有可能会变小。因此,并不是集成的分类器数目越多越好,zhou 等人已经验证了“Many Could Be Better Than All”,从已有的个体学习器中进行选择之后再集成,就可以获得更好的性能[2]。本文从参与集成的基分类器的分类准确性和差异性两方面考虑,将差异性度量方法-不一致度量法应用到分类器集成中来,从而使得多分类器系统性能显著性提高。

1 分类器的差异性度量

到目前为止,对多分类器集成的研究成果已经很多了,但是面对这么多的分类器如何进行选择就成了一个棘手的问题。现在研究者们希望找到某种分类器的关联度量方式来对集成的多分类器系统的构造提供依据。

通常认为,集成多个完全一致的分类器(输入输出均完全一样)是不会对性能有任何帮助的。如果存在完美的分类器,则集成又是没有必要的。既然分类器不是完美的,那么参与集成的分类器必须是存在差异的,也就是说,至少其中一些分类器要对其中一些多分类器判断错误的样本作出正确的决策。这种性质被称作分类器的差异性。衡量这种差异性的方法被称为差异性度量方法(Diversity Measure)。目前对差异性度量的研究主要集中在两个方面:如何寻找合适的度量方法以及如果找到了合适的度量方法,如何利用它来对集成的多分类器系统进行改造从而达到提高分类性能的目的[3]。

对于两个分类器 D_i 和 D_j,常用的定义它们之间差异性的数据有:N^{11} 与 N^{00} 代表两分类器均预测正确与均预测错误,即两分类器均作出正确预测或错误预测的训练样本占总训练样本的比例。N^{10} 为 D_i 预测正确而在 D_j 中预测错误,而 N_{01} 为 D_i 预测错误而在 D_j 中预测正确。

常用的分类器差异性度量有以下几种度量方法:Q 统计法、不一致度度量法、熵度量法等[4]。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60702056)。
作者简介:郭红玲(1982-),女,硕士研究生,主要研究方向模式识别;程显毅(1956-),男,教授,博士生导师,主要研究方向模式识别,机器学习。
收稿日期:2008-03-10 **修回日期**:2008-05-27

(1)Q 统计

Q 统计方法对两个分类器 D_i 和 D_j 之间的差异性定义如下：

$$Q_{i,j}=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}} \tag{1}$$

当分类器相互独立时，$Q_{i,j}$ 的值为 0，$Q_{i,j}$ 值变化的范围从-1 到 1 之间，当两分类器倾向于同时将同一个目标分类正确时，$Q_{i,j}$ 值会是正值，相反 $Q_{i,j}$ 的绝对值越大，差异性越小。

$$Q_{av}=\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}Q_{i,j} \tag{2}$$

多分类器组成的集合的差异性可记为子分类器两两之间差异性的均值[5]。

(2)不一致度量

不一致度量方法对两个分类器 D_i 和 D_j 之间的差异性定义如下：

$$D_{i,j}=\frac{N^{01}+N^{10}}{N^{11}+N^{10}+N^{01}+N^{00}} \tag{3}$$

$$D_{av}=\frac{2}{N(N-1)}\sum_{i\neq j}D_{i,j} \tag{4}$$

不一致度量值 $D_{i,j}$ 变化的范围从 0 到 1 之间。当两个基分类器同时将每一个目标分类正确或错误时，度量值会是 0，当两个分类器预测不同，且有一个预测是正确时，度量值会是 1。D_{av} 越大，分类器间差异性越大。

(3)熵度量方法，即首先度量各分类器在一个样本上分类结果的离散度，然后得到所有样本离散度的均值，其定义如下：

$$\bar{E}=\frac{1}{m}\sum_{x=1}^{m}\sum_{k=1}^{c}-p_k^x\log(p_k^x) \tag{5}$$

其中，p_k^x 表示样本 x 被分到类 k 的概率，m 为样本个数，C 为类别个数。

Kuncheva 和 Shipp 等对各种差异性度量方法进行了分析和实验[6]，发现各种度量方法间相关度很大，和多分类器准确率的关系也近似，在后续的算法分析中运用了广为采用的不一致度量方法。

2　多分类器选择集成算法

分类器间差异性对多分类器系统的性能非常重要，同时分类器的平均准确率也很重要，而差异性和平均准确率又存在一定的矛盾，当分类器的个数增多时，分类器间的差异性常常会变低。本文试图设计一种算法，该算法能取得两者间较好的平衡。即此方法期望挑选出分类精度高且差异度较大的分类器来构建集成。

为了使分类器集的识别性能更高，从以下两方面来着手，一从已有分类器集中选择出识别率较高的分类器；二从这部分分类器中选择出差异性较大的分类器来实现最后的集合。其具体算法如下：

输入：K 个样本（其中 X 个训练样本，Y 个集成训练样本，其余的为测试样本）

步骤 1　训练出 m 个分类精度较高的基分类器

(1)用训练样本训练生成 L 个分类器 C_i。

(2)用每个分类器识别集成训练样本，根据识别结果，计算出每个分类器的识别准确率 R_i。

(3)根据分类器的识别率的大小，选择出 m 个识别效果较好的分类器，$n<m<L$。

步骤 2　选择出 n 个差异性较大的分类器进行集成

(1)for $i=1,\cdots,(m-1)$

{ for $j=i+1,\cdots,m$

计算出分类器 C_i 和 C_j 间的差异性 $Q_{i,j}$

}

(2)从所有分类器中任选出 n 个，求这个新的分类器集的差异 Q_{av}^i，$i=1,\cdots,n$。

(3)比较所有的分类器集的差异 Q_{av}^i，则最大 Q_{av}^i 对应的分类器集就是最后用来集成的分类器器集。

3　实验仿真和分析

本文采用不一致度量法来度量选择组成分类器集的基分类器。实验中，采用 bagging 算法[7]来训练产生基分类器，即重复地随机选择样本子集作为训练集；采用的基分类器是最近邻分类器。

使用 UCI 标准数据集“letter”进行测试。Letter 数据集包括 A~Z 26 个大写字母，共 26 类。每个样本提取 16 个特征，共 20 000 个样本，每类样本数大约为 800 个。将样本集划分为前 13 000 个样本作为分类器训练集；其后 3 000 个样本作为集成训练集；最后 4 000 个样本作为集成测试集。

利用 bagging 算法从训练样本中随机选取一些样本分组成 25、20、15、10 个样本子集，并用它们生成了由 25、20、15、10 个基分类器（最近邻分类器）组成分类器集，并分别计算它们的正确率，然后用其对集成训练样本集进行识别，根据识别准确率来去除识别效果最差的 3 个，最后利用差异性度量法选出差异性最大的基分类器来组成最后的分类器集（去除 2 个冗余的分类器），并计算它们各自的正确率。

分别用贝叶斯法和投票法进行实验，表 1 列出了直接由 bagging 方法构造的不同大小的分类器集的正确率，表 2 为对基分类器进行选择后的正确率，图 1 为分类器选择前后的正确率比较。

表 1　bagging **方法构造的分类器集的分类正确率**　(%)

集成方法	分类器数量			
	bagging 25	bagging 20	bagging 15	bagging 10
贝叶斯	88.5	89.1	90.0	88.1
投票法	88.0	88.0	89.0	87.5

表 2　**经过选择组成的分类器集的分类准确率**　(%)

集成方法	分类器数量			
	bagging 25	bagging 20	bagging 15	bagging 10
贝叶斯	92.0	93.5	91.7	91.0
投票法	91.5	93.0	91.5	90.5

从表 1 和表 2 可以看出，构造多分类器系统的时候，分类器数量并非越多越好，也并非越少越好。

注：图 1 中选择后分类器系统识别准确率中，对应的分类器数目是选择前的数目，而非选择后的数目。

从图 1 可以看出，相同数目的基分类器组成的分类器集，经过选择的分类性能明显的比未经选择的好得多，系统效率得到了提高。

这个实验结果说明，从基分类器的分类准确率和基分类器间差异性两方面考虑多分类器系统的设计和改进是非常有前景的。

（下转 190 页）