一种基于信息熵的多分类器动态组合方法

陈 冰,张化祥
CHEN Bing,ZHANG Hua-xiang

山东师范大学 信息科学与工程学院,济南 250014
College of Information Science and Engineering,Shandong Normal University,Jinan 250014,China
E-mail:zyxcscb@163.com

CHEN Bing,ZHANG Hua-xiang.Method of dynamic ensemble of multiple classifiers based on information entropy.Computer Engineering and Applications,2008,44(22):146-148.

Abstract: A method of dynamic ensemble of multiple classifiers based on information entropy(EMDA) is proposed in the paper, in order to improve the classification performance of dataset.The algorithm is tested on the UCI benchmark data sets,and comparative classification efficiency with several member classfiers trained based on ensemble learning algorithm—Adaboost.In the end,the utility of EMDA algorithm can be proved in the paper.

Key words: multiple classifiers;information entropy;clustering;classifier ensemble;Adaboost

摘 要:为提高数据分类的性能,提出了一种基于信息熵[1]的多分类器动态组合方法(EMDA)。此方法在多个 UCI 标准数据集上进行了测试,并与由集成学习算法—AdaBoost,训练出的各个基分类器的分类效果进行比较,证明了该算法的有效性。

关键词:多分类器;信息熵;聚类;分类器组合;Adaboost
DOI:10.3778/j.issn.1002-8331.2008.22.043 **文章编号**:1002-8331(2008)22-0146-03 **文献标识码**:A **中图分类号**:TP391.4

1 引言

目前,关于多分类器系统的研究越来越多,并且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。多分类器得到如此的重视,其主要原因是多分类器组合技术在各个领域已经得到了广泛的应用。同时也在不同的应用领域提出了很多种分类器组合方法,但是这些方法并不很理想,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[3];更多的情况是利用不同的融合算法来训练生成同种类型的分类器,再利用它们来对数据分类。另外,通常所使用的分类方法如:决策树、贝叶斯等都是有导师信息的机器学习过程,即所有的学习过程都是在已知类别标签的样本集上进行。但如果利用它们去分类没有类别标记的样本,其效果就比较差了。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,当然由于缺乏导师信息,使其结果具有不确定性,并且聚类初始中心的选取对聚类的结果影响很大。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。还有使用多分类器组合值得注意的一点:考虑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。

基于以上考虑,本文提出了一种有效的 EMDA 算法,该方法首先根据类别标号将训练数据划分成一个个小集合,并在训练数据类别数的指导下对测试数据聚类,依据欧氏距离找出聚类集与训练数据的小集合之间的对应关系。在 Adaboost 基础上采用不同的分类器算法,在训练数据的每个小集合中训练出不同类型的成员分类器,并使用信息熵选择出可靠性较强的分类器去分类测试数据中相对应的聚类集,从而获得 EMDA 的分类性能。并在多个 UCI 标准数据集上进行测试,同时与在 Adaboost 基础上采用不同分类器算法训练出的成员分类器(决策树、神经网络、K-近邻)的分类效果进行比较,证明其有较好的分类性能和较强的泛化能力。

2 多分类器动态组合流程

EMDA 方法一次对样本随机取样的流程,如图 1。图中样本集 $1\cdots k$ 是对训练样本集按照类别标号得到的 k 个小集合;分类器组合 $1\cdots k$ 表示的是由训练样本集中的每个样本集根据信息熵训练得到的 k 组可靠性较强的分类器组合。再利用这 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(为表

基金项目:山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Y2007G16);山东省科技公关计划(the Key Technologies R&D Program of Shandong Province,China under Grant No.2005GG4210002);山东省青年科学家科研奖励基金(the Young Scientist Scientific Research Premium Foundation of Shandong Province under Grant No.2006BS01020);山东省教育厅科技计划项目(the Science and Technology Plan Project of Shandong Province Education Department under Grant No.J07YJ04)。

作者简介:陈冰(1981-),女,硕士,主要研究方向:数据挖掘、机器学习;张化祥,博士,教授,主要研究方向:机器学习、人工智能及 Web 挖掘。

收稿日期:2007-10-10 **修回日期**:2008-01-21

示的方便假设样本集与聚类集合是一一对应的)。

最后用每个聚类集中错误分类的样本数之和除以测试数据总数,得一次采样的错误率。

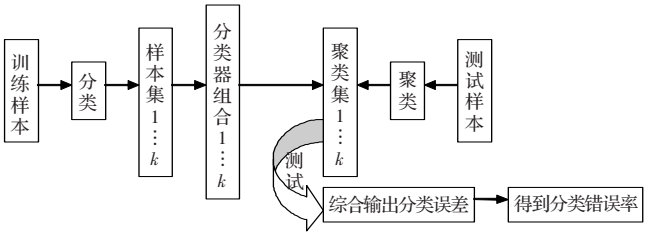

图1　多分类器动态组合流程

3　多分类器动态组合

3.1　经典聚类——k-均值[4](K-means)

k-均值算法的基本思想是: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。

在 k 均值中,用到的欧氏距离,可给出如下表示:

假定所有的实例对应于 n 维空间 R_n 中的点,把任意的实例 x 表示为如下的特征向量:$<a_1(x),a_2(x),\cdots,a_n(x)>$其中,$a_r(x)$ 表示实例 x 的第 r 个属性值,那么两个实例 x_i 和 x_j 间的距离定义为:

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2} \tag{1}$$

EMDA 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。

3.2　集成学习[5]

集成学习方法是根据样本训练多分类器来完成分类任务的方法,训练出的分类器具有一定的互补功能,能较好地减少分类误差。其中,Adaboost 算法[6]就是一个比较成功的集成学习算法。

Adaboost 算法的形式化描述:

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 $S=\{(x_i,y_i)|x_i\in X,y_i\in Y,i=1,\cdots,n\}$。Adaboost 首先从训练数据集 S 中,以一个均衡的权向量 $\boldsymbol{w}_1(1)=\cdots=\boldsymbol{w}_1(n)=1/n$ 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 $h_t(\cdot)$对原数据集 S 进行分类并且考虑权向量训练错误 ε_t:

$$\varepsilon_t=\sum_{i=1}^{n}w_t(i)I(y_i\neq h_t(x_i)) \tag{2}$$

其中当事件 E 出现时,$I(E)=1$,否则等于 0。再设

$$\beta_t=\frac{1}{2}\log=\frac{1-\varepsilon_t}{\varepsilon_t} \tag{3}$$

然后用下面规则更新权向量:

$$w_{t+1}(i)=w_t(i)\exp\{-\beta_t y_i h_t(x_i)\}/Z_t \tag{4}$$

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

$$\alpha_t=\frac{\beta_t}{\sum_{t=1}^{T}\beta_t} \tag{5}$$

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。

EMDA 基于 Adaboost 算法,采用不同的分类器算法(决策树、神经网络、K-近邻),随机的训练出不同类型的基分类器,可以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,因而利用这种差异的互补性可以提高分类的性能。

3.3　基于信息熵的成员分类器的动态选择[1]

信息融合的过程实质上也是降低系统输出不确定性、提高决策可靠性的过程。当多分类器融合中的各成员分类器相关性最小时,融合系统的输出不确定性也达到最小。因此,在存在较多可选择的融合成员的时候,应该从中选取可靠性较高、相关性较小的一部分。EMDA 正是利用信息熵来从训练数据的每个小集合中训练得到 $N(50)$个不同类型的分类器,从中选择出可靠性较高的 $n(10)$个分类器去分类对应标号的测试数据的聚类集。

对于输入的待分样本,成员分类器输出的不确定性可以用信息熵计算如下:

$$entr_i=-\sum_{k=1}^{n}(c_{ik}*\ln c_{ik})(i=1,2,\cdots,n) \tag{6}$$

其中 n 是类别数,c_{ik} 是归一化的第 i 个分类器对第 k 类的输出,可以视为类别的后验概率。信息熵越小,说明分类器的不确定性越好,即越可靠。对不同的样本,可以选择出相对可靠性最大的一部分成员进行融合。

3.4　EMDA 的思想

训练数据按类别分成 k 个小集合,k 表示类别数,其中每个小集合中的数据属于同一类。再用 k-means 方法对测试数据聚类,由于已知数据的类别数,所以可以把测试数据聚类成 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 $T_i(i=1,2,\cdots,k)$,测试数据的 k 个聚类集表示为 $t_j(j=1,2,\cdots,k)$。

对每一个 t_j 的聚类中心 $tc_j(j=1,2,\cdots,k)$

```
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))
end
```

(7)

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。

由于训练数据分成了 k 个小集合,所以每个小集合都可以得到 $N(50)$个不同类型的分类器,根据信息熵从每个小样本集训练得到的 $N(50)$个分类器选择出可靠性较高的 $n(10)$个分类器去分类对应标号的测试样本的聚类集,得到错误率。

3.5　EMDA 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 EMDA 的错误率为:

$$error_{EMDA}=\frac{\sum_{i=1}^{k}m_i}{M}(i=1,2,\cdots,k) \tag{8}$$

4 实验及结果分析

实验中采用 Adaboost 算法作为集成学习的学习方法,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 算法。利用随机数生成器在训练数据集的每个小集合上随机的生成 50 个不同类型的分类器，然后根据信息熵选择出 10 个可靠性较好的，去分类对应测试集中的聚类集。由于要对测试集进行聚类,考虑到存在小数据集的情况,EMDA 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，所以应采用多次计算求平均的方法,EMDA 选用 50 次循环,来最终求得正确率。然后使用 UCI 标准数据集，对 EMDA 以及在 Adaboost 基础上生成的三种基分类器所测得的正确率进行比较,实验结果如表 1。

表 1 EMDA 及在 Adaboost 基础上生成的 3 种分类器所测正确率比较

序号	数据集(dataset)	正确率/%			
		EMDA	Adaboost (J48)	Adaboost (NB)	Adaboost (IBk)
1	breastcancer-w	98.614 3	96.137 3	95.851 2	95.994 3
2	heart-statlog	97.886 0	80.000 0	81.481 5	75.555 6
3	machine	95.709 4	87.081 3	79.904 3	87.559 8
4	labor	92.142 9	91.228 1	89.473 7	85.964 9
5	hepatitis	91.589 2	83.870 0	78.709 7	81.935 5
6	credit-g	91.400 0	75.400 0	75.400 0	72.400 0
7	letter1	89.681 0	91.360 0	63.220 0	89.980 0
8	sonar	86.538 5	83.173 1	85.096 2	86.057 7
9	audiology	85.975 9	84.513 3	79.203 5	78.318 6
10	artificial	78.185 7	60.246 6	30.182 0	56.821 3
11	glass	75.166 6	73.364 5	46.729 0	67.289 7
12	automobile	72.124 8	83.414 6	60.000 0	74.146 3
13	vehicle	71.750 1	78.487 0	45.981 1	69.267 1
14	clean1	65.336 1	92.016 8	83.193 3	85.504 2
15	hayes-roth	55.787 9	74.242 4	75.757 6	61.363 6

注:数据集 letter1 是原 letter 中前 5 000 个实例,由于考虑到运行速度的问题。

根据表 1 中的实验数据对 EMDA 做如下分析:

(1)在表 1 中的 15 个 UCI 数据集,EMDA 算法正确率高于其他 3 种方法的有 10 个。因此从数据集的特点来看,EMDA 算法适合于处理样本数较大,属性个数较多的数据集,当然更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,对于这样的数据集,它的分类效果有明显的优势。

(2)EMDA 算法采用多种不同类型的基分类器比其他方法中只使用一种类型的基分类器的分类正确率要好,充分说明不同分类器之间的差异互补性。从实验数据中也可以看出,基分类器的分类效果对 EMDA 影响很大，如正确率差距很小的数据集:breastcancer-w、creat-g 等,EMDA 的正确率较高;而数据集 glass、letter1、vehicle 等的分类效果差距很大,EMDA 的正确率就相对较低。因此,考虑使用何种类型的基分类器来提高分类的正确率是非常重要的。

(3)由于 EMDA 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,因此在时间的耗费上较其他方法可能要多一些，但是考虑到正确率,时间复杂度问题应该可以忽略。

5 结论

EMDA 算法的优越性:EMDA 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来用于测试，体现出了分类器多样性的特点。同时 EMDA 采用了有导师学习与无导师学习相结合的思想,使用了有导师的决策树、神经网络、K-近邻分类方法与无导师的聚类方法相结合,使得分类正确率明显提高。由于采用了聚类这种无导师的分类方法,对测试样本进行了聚类,然后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，再使用这些分类器组去分类测试数据,所以在实际中可以利用该方法处理没有类别标签的样本。

因此,在以后的实际应用中,应充分考虑分类器差异互补的特点以及有导师与无导师的分类方法相结合的思想,有关如何结合这些思想还需进一步的研究。

参考文献:

[1] 谢华,夏顺仁,高光金.基于分类器融合的骨髓细胞识别研究[J].计算机工程与应用,2005,41(27):184-186.

[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] 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.

[4] Mitchell T M.机器学习[M].北京:机械工业出版社,2006:166-167.

[5] 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.

[6] Witten I H,Eibe F.数据挖掘实用机器学习技术[M].2 版.北京:机械工业出版社,2006:212-214.

[7] Dymitr Ruta,Bogdan Gabrys.Classifier selection for majority voting[J].Information Fusion,2005,6(1):63-81.

[8] Wang Xiao,Wang Han.Classification by evolutionary ensembles[J].Pattern Recognition,2006,39(4):595-607.

(上接 83 页)

[2] Coupean M.Electrostatic spraying of liquids:main function models[J].Journal of Electrostatic,1990,25(1):165-184.

[3] 周浩生,冼福生.荷电射流雾化研究[J].江苏大学学报:自然科学版,1995,16(4):7-12.

[4] 高全杰,陈馨.基于粒子系统的静电喷涂雾化模拟研究[J].冶金设备,2004,12(6):47-49.

[5] van Wijk J J.Flow visualization with surface particles[J].IEEE Computer Graphics & Applications,1993(7).

[6] 黄晶晶.基于 OpenGL 的发动机试车仿真动画设计[J].计算机仿真,2005,22(4):214-217.

[7] Wright R S,Sweet Jr M.OpenGL[M].北京:人民邮电出版社,2001.

[8] 江早.OpenGL VC/VB 图形编程[M].北京:科学出版社,2001.