维普资讯 http://www.cqvip.com

基于全信息相关度的动态多分类器融合*)

张健沛　程丽丽　杨　静　马　骏
(哈尔滨工程大学计算机科学与技术学院　哈尔滨 150001)

摘　要　AdaBoost采用级联方法生成各基分类器，较好地体现了分类器之间的差异性和互补性。其存在的问题是，在迭代的后期，训练分类器越来越集中在某一小区域的样本上，生成的基分类器体现不同区域的分类特征。根据基分类器的全局分类性能得到固定的投票权重，不能体现基分类器在不同区域上的局部性能差别。因此，本文基于AdaBoost融合方法，利用待测样本与各分类器的全信息相关度描述基分类器的局部分类性能，提出基于全信息相关度的动态多分类器融合方法，根据各分类器对待测样本的局部分类性能动态确定分类器组合和权重。仿真实验结果表明，该算法提高了融合分类性能。

关键词　AdaBoost，动态多分类器融合，全信息相关度

Dynamic Multiple Classifiers Combination Based on Full Information Correlation

ZHANG Jian-Pei　CHENG Li-Li　YANG Jing　MA Jun
(College of Computer Science and Technology, Harbin Engineering University of China, Harbin 150001)

Abstract　The base classifiers trained by AdaBoost combination learning algorithm are produced orderly, diversity and complementarity of base classifiers are assured. But along with the iterative process of AdaBoost, the classifier which represents different area classification performance mainly focuses on a certain small area of input space. The constant weights are obtained according to overall classification performance which can not demonstrate base classifiers' classification performance in different local areas. Based on AdaBoost, a dynamic multiple classifiers combination algorithm based on full information correlation (FIC) which describes base classifiers' local classification performance is proposed, the classifiers' selection and their weights are determined according to test samples' FIC to base classifiers. The simulated experiments show that the combination classification performance is improved greatly.

Keywords　AdaBoost, Dynamic multiple classifiers combination, Full information correlation

1　引言

实际应用的复杂性和数据的多样性使得单一的分类方法难以获得令人满意的识别性能，Suen于1990年提出了多分类器融合的概念，并已在模式识别的很多方面得到广泛应用，如字符识别、目标识别、文本分类等领域获得了较好的应用效果[1~3]。多分类器融合方法的基本假设是：利用具有不同特性和性能的多分类器，通过有效的组合可以获得更高的识别性能。

AdaBoost算法是一个比较著名的融合学习算法，该算法在训练阶段，生成分类器的训练集取决于之前的分类器的表现，被已有分类器错误判断的样本将以较大的概率出现在新的训练集中，最终生成多个具有差异性的分类器集。分类器输出一般采用多数投票规则，每个分类器产生自己的分类结果，这些分类结果被融合形成最终的融合决策，其融合规则对于所有的待测样本均采用同样的投票权重[4]。然而，各分类器对不同样本的分类效果是不同的，同一分类器在样本空间不同区域性能会有变化，因此对于不同待测样本采用固定投票权重的分类器融合不能充分体现基分类器的局域差异性特征[5]。

基于以上认识，本文提出了一种动态多分类器融合方法，给出待测样本与分类器的全信息相关度概念，根据相关度值动态调整分类器的组合和权重，而最终结果为各分类器的融合。仿真实验证明了该方法的有效性。

2　AdaBoost算法分析

首先给出经典的AdaBoost算法流程(二值情况)：$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$x_i\in X$，$y_i\in Y=\{-1,+1\}$，$i=1,\cdots,m$。

初始化：对每个$(x_i,y_i)\in S$ $D_1(x_i,y_i)=1/m$；

DO For $t=1,2,\cdots,T$

Step1　调用WeakLearn，D_t作为WeakLearn的参数；

Step2　获得基分类器$h_t:X\to Y$，确定$w_t\in R$；

Step3　更新$D_{t+1}(x_i,Y_i)=\dfrac{D_t(x_i,Y_i)\exp(-\omega_t y_i h_t(x_i))}{Z_t}$

此处Z_t是标准化常量，使得$\sum_{(x_i,y_i)}D_{t+1}(x_i,y_i)=1$

输出：最终的分类器$H(x)=sign(\sum_{t=1}^{T}w_t h_t(x))$

$D_t(x_i,y_i)$为样本(x_i,y_i)在第t次迭代时的权重，h_t为第t次迭代产生的弱分类器，$w_t=\log(\dfrac{1-\epsilon_t}{\epsilon_t})$为$h_t$的融合权重，$\epsilon_t$为弱分类器$h_t$的加权分类错误率。

从AdaBoost算法的迭代过程中可以看出：(1)$D_{t+1}(i)$通过增加不正确分类样本的$D_t(i)$计算得到，即增加分类错误的样本的分布，使得下一次训练的弱分类器更加关注那些被识别错误的样本，而正确分类的样本的权重分布下降；(2)具

*)本课题得到国家自然科学基金(60673131)和黑龙江省自然科学基金(F2005-02)资助。张健沛　博士，教授，主要研究方向：数据挖掘。

维普资讯 http://www.cqvip.com

有较小训练错误率(这里指全局分类错误率)的分类器,其权重大于具有较大训练错误率的分类器;(3)一旦分类器学习成功,对所有的待测样本均采用同样的投票权重。然而,具有较大全局分类错误率的基分类器,在某个局部区域可能有较好的分类性能,对属于这一区域的待测样本应给予较大的融合权重;相反,具有较小全局分类错误率的基分类器,在某个局部区域可能有较差的分类性能,对属于这一区域的待测样本应给予较小的融合权重。为了利用融合分类器的互补优势,同时兼顾待测样本的特征差异,本文提出待测样本与分类器的全信息相关度定义,以评价不同分类器对于待测样本分类输出的可信任程度,动态确定分类器的组合和权重。

3 动态多分类器融合

为确定待测样本与分类器之间的全信息相关度,需要解决:(1)找到与待测样本特征相似的样本,组成待测样本的有效邻域;(2)确定待测样本对基分类器的全信息相关度,刻画基分类器的局部分类性能。

3.1 待测样本有效邻域的判定

待测样本的邻域是指与之邻近的一组训练样本构成的区域。由于在待测样本的邻域中,通常会存在一些样本,多个分类器在这些样本上都出现了"不同类"的判断,那么这些样本与待测样本属于不同类的可能性就比较大,对分类器的性能分析造成干扰,因此应该剔除这些干扰样本,以得到待测样本的有效邻域。

设 $\omega_j(j=1,2,\cdots,M)$ 为有 M 个类别的分类问题,$H=\{h_t,t=1,2,\cdots,T\}$ 为 T 个不同的分类器,h_t 对样本 x 的分类输出为:$h_t(x)=(c_{t1},c_{t2},\cdots,c_{tM})$,$0\leqslant c_{tj}\leqslant 1$ 表示分类器 h_t 给出 x 属于类 ω_j 的概率。

定义 1 $o_t(x)=\arg\max_j(c_{tj})$ 为分类器 h_t 在样本 x 上的行为。

定义 2[6] $MCB(x)=\{o_1(x),\cdots,o_T(x)\}$ 为 T 个分类器在样本 x 上的行为(multiple classifier behavior,简称 MCB)。

定义 3 对于样本 X 和 Y 的 MCB 之间的相似度定义为 $S(X,Y)=\frac{1}{T}\sum_{t=1}^{T}Q_t(X,Y)$,其中 $Q_t(X,Y)$ 定义为

$$Q_t(X,Y)=\begin{cases}1 & \text{if } o_t(X)=o_t(Y)\\ 0 & \text{if } o_t(X)\neq o_t(Y)\end{cases}$$

根据定义 2、3 可以得出,两个样本的 MCB 之间的相似度的取值范围为[0,1]。当取值为 1 时,说明每一个分类器都将两个样本判为属于同一类;当取值为 0 时,说明每一个分类器都将两个样本判为属于不同的类。利用这种方法可以剔除与待测样本 MCB 的相似度小于阈值的样本,求得待测样本的有效邻域。对于某一待测样本,考察其在有效邻域内与基分类器的全信息相关度,实现动态分类器融合。

3.2 全信息相关度的获取

对于样本 x 的输出,我们将分类结果向量的各分量按照从大到小的顺序依次排列,第一个分量对应着第一个候选类别,依此类推。

定义 4 $h_t(x)=(c_{t1},c_{t2},\cdots,c_{tM})$ 称为分类器 h_t 对于样本 x 的全信息行为。

定义 5 $FIM(x)=[c_1(x)',c_2(x)',\cdots,c_T(x)']'$

$$=\begin{bmatrix}c_{11} & \cdots & c_{1M}\\ \cdots & c_{tj} & \cdots\\ c_{T1} & \cdots & c_{TM}\end{bmatrix}$$

称为样本 x 在多分类器 H 下的全信息矩阵[5]。

文[7]认为分类器输出的不同候选类别与待测样本之间存在某些必然的相关性,分类结果向量输出的各阶候选对相关程度的支持量是不同的。分类器输出的较低阶次的候选类别(如第一、二候选类别等)对相关度的支持作用大于较高阶次的候选类别。

定义 6 分类器 $h_l(1\leqslant l\leqslant T)$ 对于样本 x_i 和 x_j 的全信息输出的第 $p(1\leqslant p\leqslant M)$ 个分量之间的相关值定义为

$$r_p=1-\left|\frac{c_{lp}^i}{\sum_{s=1}^{M}c_{ls}^i}-\frac{c_{lp}^j}{\sum_{s=1}^{M}c_{ls}^j}\right|\Big/\left|\frac{c_{lp}^i}{\sum_{s=1}^{M}c_{ls}^i}+\frac{c_{lp}^j}{\sum_{s=1}^{M}c_{ls}^j}\right|$$

其中 c_{lp}^i 表示分类器 h_l 对样本 x_i 分类输出的第 p 阶分量。如果两个分类结果向量的第 p 阶分量对应着不同的候选类别,那么它们之间的相关性比较没有任何意义,即该分量的相关程度为 0;如果对应着相同的候选类别,且大小相等(当分量值做了归一化之后),则相关程度为 1,所以相关值落在[0,1]之间。分量之间的差除以分量之间的和,表示了它们的相对的差异程度,而不是绝对的差异程度。其好处是将度量尺度归一化,使衡量标准统一。

定义 7 分类器 h_l 对于样本 x_i 和 x_j 的全信息相关度定义为

$$R_l(x_i,x_j)=\sum_{p=1}^{M}\eta_p r_p \tag{1}$$

η_p 表示第 p 阶候选的支持因子。η_p 可有多种形式,如 $\eta_p=e^{-\alpha_1(p-\alpha_2)}$ 或 $\eta_p=1.0-\alpha_3\times p$,其中 α_1,α_2 和 α_3 为非负常数。

定义 8 样本 x 在有效邻域上的全信息相关度矩阵(Full Information Correlation Matrix, FICM)定义为

$$FICM(x)=\begin{bmatrix}R_1(x,x_1),\cdots,R_t(x,x_1),\cdots,R_T(x,x_1)\\ \cdots\\ R_1(x,x_j),\cdots,R_t(x,x_j),\cdots,R_T(x,x_j)\\ \cdots\\ R_1(x,x_k),\cdots,R_t(x,x_k),\cdots,R_T(x,x_k)\end{bmatrix}$$

其中,$x_j\in N(x)$,$K=|N(x)|$,$N(x)$ 为 x 的有效邻域。当分类器 h_t 对训练样本 x_j 分类正确时,$R_t(x,x_j)$ 的值按照公式(1)计算求得;如果分类器 h_t 对训练样本 x_j 分类错误时,$R_t(x,x_j)$ 的取值为 0。

定义 9 样本 x 与基分类器 h_t 的相关度值取对有效邻域内样本的相关度值的平均值:

$$\delta_t=\frac{1}{K}\sum_{j=1}^{k}R_t(x,x_j) \tag{2}$$

全信息相关度充分利用了分类器对于训练样本的分类输出,描述了样本和基分类器之间的相关度,体现了分类结果向量输出的各阶候选对相关程度的支持作用,避免了只利用分类器分类结果中概率最大的那一类标号作为相关度度量所带来的信息丢失问题。举例说明,假设两个分类器的分类结果向量为[0.5,0.4,0.1]和[0.7,0.2,0.1],虽然两个分类器输出相同的类别,但是后者的可信度明显高于前者。如果我们只是取类别概率最大的类标作为分类结果参与相关度的计算,就会丢失第二类别概率、第三类别概率所提供给我们的信息以及两者之间可信度的差异,影响最终的融合分类。

3.3 动态分类器组合及权重

设分类器集合 $H=\{h_t,t=1,2,\cdots,T\}$,待测样本 x 和基分类器 h_t 的相关度为 δ_t,则分类器 h_t 对于待测样本的加权参数为

维普资讯 http://www.cqvip.com

$$w'_t=\begin{cases}\log(\frac{\delta_t}{1-\delta_t}) & \text{if } \delta_t>\text{阈值}\\ 0 & \text{if } \delta_t<\text{阈值}\end{cases} \quad (3)$$

动态融合决策为

$$H(x)=\text{sign}(\sum_{t=1}^{T}w'_t h_t(x)) \quad (4)$$

动态多分类器融合算法可描述如下。

步骤 1：根据 AdaBoost，训练形成各分类器；

步骤 2：输入待测样本 x，确定 x 的 MCB 值；

步骤 3：对于 x，如果 MCB 中的各项值都相等，则说明各个分类器对 x 的类别判断一致，则直接输出该类别作为 x 的类别，算法结束；否则，求出待测样本的有效邻域；

步骤 4：根据公式(2)计算 x 对于每个分类器的相关度，根据公式(3)确定融合权重；

步骤 5：根据公式(4)得到分类器的融合输出。

各分类器的分类效果随着待测样本的特征和区域的不同而不同，因此需要针对不同样本进行分类器的组合及权重调整。本文利用待测样本与分类器的相关度来刻画基分类器的局部分类性能，对于相关度小于某一阈值的基分类器，不参与融合决策；否则，按照相关度值的大小动态确定融合权重。

4 仿真实验

4.1 所用分类器及数据集

实验中采用前向 BP 神经网络(BPNN)、支持向量机(SVM)、Bayes 分类器分别作为基分类器的学习算法。为了验证上述所提算法(这里简称 FIC)，我们选用 UCI 机器学习测试数据集中的 8 个数据集进行实验。对于每一个数据集，取出 2/3 的样本作为训练集，1/3 的样本作为测试集。数据集的特征如表 1 所示。

4.2 实验结果分析

表 2 给出了单个分类器在所有数据集上的性能比较。对于同一个学习任务，没有一种分类算法能对所有应用都取得很好的结果，因此多分类器融合系统势在必行。对于多分类器融合，在确定待测样本的有效邻域时，与待测样本 MCB 的相似度阈值取 0.5，有效邻域中的样本根据待测样本的不同而动态确定。

表 3 给出了融合分类实验结果对比情况，前 3 列为权值固定的多分类器组合方法，简称为 FIX*；后 3 列为本文所提出的动态分类器融合方法，简称为 FIC*。从中可以看出，动态分类器融合方法相对于固定权重的融合方法从整体上提高了分类性能，对不同分类器和样本都有较高的分类准确率。

表 1 UCI 数据集合

Name	Cases	classes	features	
			Numeric	Nominal
Anneal	898	6	9	29
Autos	205	6	15	10
Credit-a	690	2	6	9
Glass	214	6	9	—
Heart-c	303	2	8	5
Hepatitis	155	2	6	13
Colic	368	2	10	12
Iris	150	3	4	—

表 2 单分类器的准确率比较

Data set	BPNN	SVM	Bayes
Anneal	0.840	0.851	0.845
Autos	0.832	0.856	0.857
Credit-a	0.838	0.852	0.849
Glass	0.836	0.854	0.835
Heart-c	0.837	0.860	0.831
Hepatitis	0.831	0.861	0.837
Colic	0.833	0.855	0.839
Iris	0.829	0.857	0.843

表 3 融合分类方法准确率比较

Data set	FIXBPNN	FIXSVM	FIXBayes	FICBPNN	FICSVM	FICBayes
Anneal	0.844	0.858	0.852	0.892	0.883	0.885
Autos	0.829	0.856	0.839	0.884	0.892	0.878
Credit-a	0.861	0.875	0.853	0.882	0.894	0.876
Glass	0.838	0.864	0.861	0.883	0.883	0.887
Heart-c	0.827	0.868	0.845	0.881	0.887	0.882
Hepatitis	0.835	0.858	0.825	0.885	0.881	0.883
Colic	0.859	0.862	0.852	0.887	0.894	0.884
Iris	0.843	0.878	0.848	0.885	0.886	0.891

结束语 在模式识别领域，为了提高分类系统的性能，多分类器融合的识别方法是一种发展趋势。多分类器融合的关键是寻找一种合适的融合规则，使分类器融合的分类效果最好。

本文利用分类器输出信息的各阶分量对实际类别的支持作用得到待测样本与分类器的相关度，评价分类器的局部分类性能。发挥各分类器对于不同样本和不同区域的分类优势，指导分类结果的融合。通过对标准 UCI 数据集的测试结果可以看出，基于全信息相关度的动态多分类器融合算法相对于传统的固定权值融合算法的分类性能有很大的提高。

参 考 文 献

1 Florian R, Ittycheriah A, Jing H, et al. Named entity recognition through classifier combination[C]. In: CoNLL-2003. San Francisco: Morgan Kaufmann Publishers, 2003. 168～171

2 Larkey L S, Croft W B. Combining classifiers in text categorization[A]. In: Proc. SIGIR'96, New York: ACM Press, 1996. 289～297

3 Schapire R E, Singer Y. Boostexter: A boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2-3): 135～168

4 Puuronen S, Terziyan V, Tsymbal A. A dynamic integration algorithm for an ensemble of classifiers[A]. In: Ras Z W, Skowron A, eds. Foundations of Intelligent System: ISMIS; 99, Lecture Notes in AI, Spring-verlag, Warsaw, 1999. 592～600

5 唐春生，金以慧. 基于全信息矩阵的多分类器集成方法[J]. 软件学报，2003, 14(6): 1103～1109

6 Giacinto G, Roli F. Dynamic classifier selection based on multiple classifier behavior[J]. Pattern Recognition, 2001, 34(9): 1879～1881

7 荆晓远，杨静宇. 基于相关性和有效互补性分析的多分类器组合方法[J]. 自动化学报，2000, 26 (6): 741～747