基于 GA 的多分类器融合算法

段敬红[1],王 黎[2]
DUAN Jing-hong[1],WANG Li[2]

1.西安理工大学 计算机科学与工程学院,西安 710048
2.西安理工大学 信息科学系,西安 710048
1.School of Computer Science and Engineering,Xi'an University of Technology,Xi'an 710048,China
2.Department of Information Science,Xi'an University of Technology,Xi'an 710048,China

DUAN Jing-hong,WANG Li.Multiple classifiers fusion algorithm based on GA.Computer Engineering and Applications, 2010,46(3):163-165.

Abstract: In order to improve the performance of single classifier,multi-classifier fusion methods have been widely used.This paper gives a new multi-classifiers fusion algorithm based on GA.To begin with,the genetic algorithm is used to partition the feature set into subsets of features for generating member classifiers,and then a new multi-classifier combining method based on weighted coefficient matrix is proposed according to the concept of class distinguishing ability presented by us.Experiments with UCI datasets show that the performance of the proposed algorithm is improved with high correct recognition rate.
Key words: multi-classifiers fusion;genetic algorithm;weighted coefficient matrix

摘 要:为了提高单一分类器的识别性能,在模式识别领域经常采用多分类器集成的方法。提出了一种基于 GA 的多分类器融合算法,首先通过 GA 算法对特征集的分割进行优化选择,形成了较优的成员分类器;然后通过对成员分类器分辨能力的度量,提出了一种加权系数矩阵的多分类器组合方法。在 UCI 数据库上进行了实验,结果表明所提出的算法具有较高的识别率。
关键词:多分类器融合;遗传算法;加权系数矩阵
DOI:10.3778/j.issn.1002-8331.2010.03.049 **文章编号:**1002-8331(2010)03-0163-03 **文献标识码:**A **中图分类号:**TP391

1 引言

模式识别分类技术被广泛应用于各个领域, 对于识别分类,传统的做法是通过实验寻求性能最优的单一分类器,但实际的效果往往并不理想。研究表明,尽管各分类算法的性能不同,但它们的误识集合却并不一定交叉,即存在互补信息,于是人们更倾向于将一些简单的分类器集成在一起,以达到提高识别性能的目的,这就是多分类器融合。Suen[1]于 1990 年提出了多分类器融合的概念,Kittler(1998)[2]对多分类器融合框架进行了分析,并给出了理论框架。在多分类器融合系统中,多样性是其成功的关键之一,Krogh[3]等就证明过要获得理想的效果,各分类器的性能都要很高,且它们之间的差异要尽可能大。Huang 等[4]提出了将单个分类器识别率作为权重的加权融合规则。从以上可以看出,多分类器融合中的关键是成员分类器的设计和集成方法的研究。Breiman 提出的 Bagging[5]算法采用从大小为 n 的原始数据集中,通过对训练样本的重抽样,以一个基算法生成相异的成员,解决了训练集样本有限的问题,但是它的取样选择是随机的,从而随机产生分类器组合,并没有理论依据证明成员选择的正确性。Bryll 提出的 AB[6](Attribute Bagging)作为 Bagging 算法的一种改进,利用对样本特征的抽取代替样本实例的抽取,可以较好地解决高维数据及样本不足现象。但是 AB 的特征提取具有随机性,不能保证系统的性能最优。为此,提出采用遗传算法(GA),对特征集的分割进行优化,从而形成相应成员分类器,这样可以保证各成员分类器从整体上是最优的。而且通过对特征集的分割,不同分类器采用不同的特征子集,使分类器具有一定的差异性,各分类器特征维数的减少,可以解决"维数灾难"的问题,降低对样本数的需求。在此基础上,提出了分类器对各类别分辨能力的概念,以此形成了一种加权系数矩阵的多分类器组合方法。

2 算法描述

算法的模型如图 1 所示。对于待测 N 维样本空间 U,测试样本 $x \in U, x=(a_1,a_2 \cdots a_N)$,(其中 $a_1,a_2 \cdots a_N$ 为样本 x 的特征向量)。空间包含 M 个模式类别,分别表示为 $(C_1,C_2 \cdots C_M)$,并有 L 个分类器 $e_k(k=1,2,\cdots L)$。为了构造多分类器系统,必须通过某种形式得到多个分量分类器。这里采用了特征分割的方法来获得成员分类器。算法通过对 N 维样本特征空间进行分割,形成 L 个特征子集,训练出 L 个分类器。然后使用这 L 个分类器对训练集进行测试,根据各分类器的分辨能力,提出了一种加权系数矩阵的组合方法,通过加权融合形成决策结果。

在这一方法中,特征子集的分割组合数量非常多。例如原

基金项目:陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2006F26)。
作者简介:段敬红(1966-),女,工程师,主要研究方向:嵌入式系统与图像处理。
收稿日期:2008-07-30 **修回日期:**2008-10-23

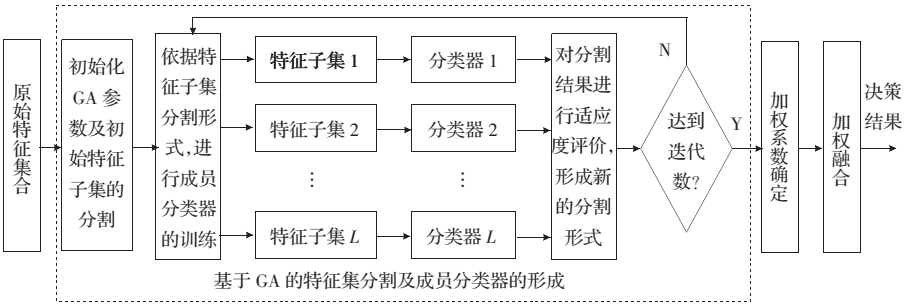

图1 基于GA的多分类器融合模型图

始特征维数为9,选取特征分割形式为[3,3,3],则组合的数量就有 $C_9^3C_6^3C_3^3$=1 680 种，这还不包括其他的分割形式及特征分量元素可重叠的情况。由此可见,要在如此多的分割组合中选择出最优的分割方式是非常困难的,必须采用一定的优化方法。图1中的虚线部分给出了采用GA对原始特征集进行分割的方法,其中特征子集的分割和成员分类器的形成是同时进行的。

2.1 基于GA的特征子集分割优化方法

在使用遗传算法进行优化设计时,主要涉及到染色体的编码、适应度函数的选择及遗传算子的设计等问题。

对于特征子集的分割有两种形式:一是无重叠的分割;二是有重叠的划分。实验表明,有重叠的划分效果较好[7],这里也采用这一种划分的方法。对于染色体的编码,其实就是如何记录特征子集的划分方式。参照文献[7]中的编码方式,对于分类器 $e_k(k=1,2,\cdots,L)$,用 N 位二进制数 $b_{k1}b_{k2}\cdots b_{kN}$ 作为一个掩模,对特征向量$(a_1,a_2,\cdots a_i,\cdots a_N)$进行筛选,$b_{kj}(k=1,2,\cdots,L;j=1,2,\cdots,N)$表示分类器 e_k 是否包含特征分量 a_j,当包含时 $b_{kj}=1$,否则 $b_{ij}=0$。将各分类器的掩模顺序连接起来,就构成了一条染色体,表示了一种特征子集的分割形式。例如9个特征3个分类器的一条染色体的编码为:000111010000010101111000000,则相应的特征子集分割为:L_1:{1,3,4,5},L_2:{2,4},L_3:{7,8,9}。按照这样的编码方式，一个具有 N 维特征 L 个分类器的特征子集的染色体编码的长度就为 NL 位。

对于每一种特征子集的划分,用训练样本对这 L 个分类器进行训练,可以得到各个分类器的训练正确率,设为 $f_1,f_2\cdots f_L$,则可以用总的正确率:$f=f_1+f_2+\cdots+f_L$ 作为适应度评价函数。

对于遗传算子的设计,复制操作采用轮盘赌算法,交叉算子采用单点交叉,变异算子采用对相应位取反操作。交叉概率取 p_c=0.6~0.8，变异概率 p_m=0.01~0.02，其他参数根据实验选择,终止条件采用迭代次数的方式。

2.2 基于分类器分辨能力加权系数矩阵组合方法

由上节得到的各分类器,其分类能力是不相同的。传统的做法是对不同的分类器分配不同的权重系数[9],但没有考虑到同一个分类器对各类别的分辨能力是不相同的。基于这一情况,提出分类器分辨能力的概念,用其构造加权系数矩阵。

设用于分类器训练的样本总数为 n,属于 C_j 类的样本数为 m,其他样本视为它的反例,个数为 $n-m$。用分类器 e_i 对训练样本进行测试,设将 C_j 类样本分类为 C_j 类的样本个数为 a,将其他类分类为 C_j 类的样本个数为 b，将 C_j 类样本分类为其他类的样本个数为 c。则定义分类器 e_i 对 C_j 类识别信息[8]:

正确识别率为:

$$Rec=\frac{a}{m}=\frac{a}{a+c} \tag{1}$$

准确识别率为:

$$Pre=\frac{a}{a+b} \tag{2}$$

对 C_j 类的分辨能力为:

$$d_{ij}=\frac{2Rec\cdot Pre}{Rec+Pre}=\frac{2a}{a+b+m} \tag{3}$$

将所有分类器对每一类别的分辨能力结合在一起,就形成了式(4)所示分类器的分辨性能矩阵 $\boldsymbol{DC}$:

$$\boldsymbol{DC}=\begin{bmatrix} d_{11} & \cdots & d_{1j} & \cdots & d_{1M} \\ \vdots & & \vdots & & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & d_{iM} \\ \vdots & & \vdots & & \vdots \\ d_{K1} & \cdots & d_{K2} & \cdots & d_{KM} \end{bmatrix} \tag{4}$$

令 $\alpha_{ij}=d_{ij}/\sum_{j=1}^{M}d_{ij}$,则依据分辨性能矩阵 $\boldsymbol{DC}$ 可以定义加权系数矩阵为:

$$\boldsymbol{\alpha}=\begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1j} & \cdots & \alpha_{1M} \\ \vdots & & \vdots & & \vdots \\ \alpha_{i1} & \cdots & \alpha_{ij} & \cdots & \alpha_{iM} \\ \vdots & & \vdots & & \vdots \\ \alpha_{K1} & \cdots & \alpha_{Kj} & \cdots & \alpha_{KM} \end{bmatrix} \tag{5}$$

设输入的待分类的样本为 x,将其输入多分类器系统后得到的输出为:

$$\boldsymbol{y}=\begin{bmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1M} \\ \vdots & & \vdots & & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{iM} \\ \vdots & & \vdots & & \vdots \\ y_{K1} & \cdots & y_{Kj} & \cdots & y_{KM} \end{bmatrix} \tag{6}$$

则样本 x 属于 C_j 类,若满足:

$$\sum_{i=1}^{K}=\alpha_{ij}y_{ij}=\max_{l=1}^{M}\sum_{i=1}^{K}\alpha_{il}y_{il} \tag{7}$$

3 实验分析

为了验证算法的有效性，选用了UCI标准数据库(http://mlearn.ics.uci.edu/MLRepository.html)提供的4个数据集进行验证比较。分别是Glass Database、Pima Indians Diabetes Database、Wine Database和Wisconsin Breast Cancer Database。表1中给出了这些数据集的信息,包括数据集的样本数、类别数、特征向量维数及训练集、测试集样本数。随机抽取数据集的3/4作为训练集,1/4作为测试集。

实验成员分类器选用最小欧式距离分类器,通过GA算法优化出3组特征子集,通过成员分类器在训练集上对各类识别

表 1　实验数据集分析

数据名称	类别数	样本个数	特征维数	训练集	测试集
Glass	7	214	9	160	54
Pima Indian Diabetes	2	768	8	576	192
Wine	3	178	13	133	45
Wisconsin Breast Cancer	2	699	9	524	175

能力,形成了相应的加权系数矩阵,通过加权组合进行最终综合决策。

为进行比较,对于 4 个数据集,按照以下特征分割方式来获得成员分类器:Glass Database(3,3,3),Pima Indians Diabetes Database(3,3,2),Wine Database(5,4,4),Wisconsin Breast Cancer Database(3,3,3),形成 3 个成员分类器 NN1,NN2,NN3,并分别应用了投票法、和规则、积规则、最大规则、最小规则和中值规则等组合规则,与该文算法比较结果如表 2 所示。

表 2　实验结果比较

组合算法	正确率			
	Glass	Pima Indian Diabetes	Wine	Wisconsin Breast Cancer
NN1	0.425 9	0.630 2	0.777 8	0.817 1
NN2	0.444 4	0.609 4	0.844 4	0.817 1
NN3	0.574 1	0.666 7	0.666 7	0.737 1
投票法	0.537 0	0.671 9	0.800 0	0.862 9
和规则	0.500 0	0.666 7	0.755 6	0.862 9
积规则	0.370 4	0.666 7	0.755 6	0.862 9
最大规则	0.537 0	0.666 7	0.688 9	0.817 1
最小规则	0.388 9	0.666 7	0.755 6	0.817 1
中值规则	0.518 5	0.671 9	0.800 0	0.862 9
该文算法	0.574 1	0.682 3	0.844 4	0.885 7

从表 2 可以看出,大多数组合算法都表现出好于或等于最优单分类器的性能。可以看出,通过分类器集成方案来提高系统性能,不失为一种好的选择。在组合算法中,该文的算法取得了较好的结果,其原因一是通过 GA 算法对特征子集的分割进行了优化选择,从而形成了较优的成员分类器,二是在分类器组合方面,考虑了各分类器对不同类别的分辨能力是不一样的,引入权值矩阵方法,使分类的准确性更好。

4　结论

对于模式识别问题,其最终的目标是尽可能好地识别样本。多分类器集成不仅能避免选取性能较差的学习算法,还能充分利用各成员分类器的互补信息,提高对目标的分类识别效果。从成员分类器的选择及信息互补性方面出发,提出了一种基于 GA 的特征分割方法及基于分类器分辨能力的加权系数矩阵组合法,取得了较好的实验效果,算法具有一定的通用性。

参考文献:

[1] Suen C Y,Nadal C,Mai T A,et al.Recognition of totally unconstrained handwriting numerals based on the concept of multiple experts[C]//Suen C Y.Frontiers in Handwriting Recognition:International Workshop on Frontiers in Handwriting Recognition,1990:131-143.

[2] Kittler J,Hatef M,Duin R P W,et al.On combining classifiers[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(3):226-239.

[3] Krogh A,Vedelsby J.Neural network ensembles,cross Validation active learning[C]//Tesauro G,Touretzky D,Leen T.Advances in Neural Information Processing Systems.Cambridge,MA:MIT Press,1995,7:231.

[4] Huang J,Yuen P C,Lai J H.Face recognition using local and global features[J].EURASIP Journal on Applied Signal Processing,2004(4):530-541.

[5] Breiman L.Bagging predictors[J].Machine Learning,1996,24(2):123-140.

[6] Bryll R,Gutierrez O R,Quek F.Attribute bagging:Improving accuracy of classifier ensembles by using random features subsets[J].Pattren Recognition Letters,2003,36(6):1291-1302.

[7] Ludmila I K,Lakhmi C J.Designing classifier fusion systems by Genetic Algorithms[J].IEEE Trans On Evolutionary Compution,2000,4(4):327-336.

[8] 宋枫溪,高林.文本分类器性能评估指标[J].计算机工程,2004,30(13):107-109.

(上接 162 页)

[2] Rezk S C,Engel K,Bauer M,et al.Interactive volume rendering on standard PC graphics hardware using multi-textures and multi-stage rasterization[C]//Proceedings of the ACM SIGGRAPH/Eurographics Workshop on Graphics Hardware,Interlaken,Switzerland,2000:109-119.

[3] Stegmaier S,Strengert M ,Klein T.A simple and flexible volume rendering framework for graphics hardware-based raycasting[C]//Proceedings of Volume Graphics,New York,2005:187-195.

[4] Röβler F,Botchen R P,Ertl T.Dynamic shader generation for flexible multi-volume visualization[C]//Proceedings of Pacific,2008.

[5] Kraus M,Strengert M,Klein T,et al.Adaptive sampling in three dimensions for volume rendering on GPUs[C]//Proceedings Asia Pacific Symposium on Visualization,2007:113-120.

[6] Weiskopf D,Schafhitzel T,Ertl T.Texture-based visualization of 3D unsteady flow by real-time advection and volumetric illumination[C]//IEEE Transactions on Visualization and Computer Graphics,2007:119-125.

[7] Cullip T J,Neumann U.Technical Report UNC-1993-027,Accelerating volume reconst-ruction with 3D texture mapping hardware[R].Unversity of North Carilian,1993.

[8] 童欣,唐泽圣.基于空间跳跃的三维纹理硬件体绘制算法[J].计算机学报,1998,21(9):807-812.

[9] Boada I,Navazo I,Scopigno R.Multiresolution volume visualization with a texture-based octree[J].The Visual Computer,2001,17(3):185-197.

[10] Li W,Mueller K,Kanfman A.Empty space skipping and occlusion clipping for texture-based volume rendering[C]//Proceedings of the 14th IEEE Visualization 2003,Seattle,WA,USA,2003:317-324.