文章编号:1671-4598(2019)04-0120-04 DOI:10.16526/j.cnki.11—4762/tp.2019.04.028 中图分类号:TP391.4 文献标识码:A

一种自适应子融合集成多分类器方法

李 敏[1],李 华[1],程茂华[2]
(1.广西师范学院 计算机与信息工程学院,南宁 530023;
2.广西科技师范学院 数学与计算机科学学院,广西 来宾 546199)

摘要:融合集成方法已经广泛应用在模式识别领域,然而一些基分类器实时性能稳定性较差,导致多分类器融合性能差,针对上述问题提出了一种新的基于多分类器的子融合集成分类器系统;该方法考虑在度量层融合层次之上通过对各类基多分类器进行动态选择,票数最多的类别作为融合系统中对特征向量识别的类别,构成一种新的自适应子融合集成分类器方法;实验表明,该方法比传统的分类器以及分类融合方法识别准确率明显更高,具有更好的鲁棒性。

关键词:分类器联合;决策置信度;决策支持度

An Adaptive Sub—fusion Integration Classification Method

Li Min[1], Li Hua[1], Cheng Maohua[2]
(1. School of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530023, China;
2. School of Mathematics and Computer Science, Guangxi Science & Technology Normal University, Laibin 546199, China)

Abstract: Fusion integration method has been widely used in the field of pattern recognition. However, some base classifiers have poor real—time performance stability, which causes poor performance of multiple classifiers. A new multi—classifier—based sub—fusion integration classification is proposed for the above problems. This method considers the dynamic selection of various classifiers at the level of measurement layer fusion, the category with the highest number of votes is the category identified by the feature vector in the fusion system to constitute a new adaptive sub—fusion integration classifier method. Experiments show that this method is significantly more accurate than conventional classifiers and classification fusion methods and has better robustness.

Keywords: classifier ensemble; decision confidence; decision support

0 引言

模式识别领域中普遍存在的一个问题是,同一个分类方法在不同的应用中分类性能不尽相同。没有哪种分类方法能够普遍适用于所有的分类情况。为了解决这样的问题,分类器融合技术成为了模式识别领域的一个重要技术。当前许多研究表明,多分类器融合技术对于模式识别的性能有较大的提高[1-3]。目前多分类器融合技术已经在很多领域上得到实践,例如图像分类、语音识别、手写技术识别等[4]。模式识别领域统一将分类器技术划分为以下两种形式:分类器动态选择[5]和分类器融合。动态分类器选择方法的核心思想是:预测当前识别任务多分类器系统中识别最准确的基分类器,选择预测的基分类器作为多分类器系统融合决策的输出。而分类器融合方法的核心思想是:全面地考虑每一个基分类器的决策输出,结合每一个基分类器的决策输出作为多分类器的最终决策输出,这种思想会得到更多的决定性决策信息。

收稿日期:2018-09-07; **修回日期**:2018-10-22。
基金项目:广西自然科学基金(2016GXNSFAA380200);2018年广西高校中青年教师基础能力提升项目(2018KY0699)。
作者简介:李 敏(1992-),男,广西玉林市人,硕士研究生,主要从事机器学习方向的研究。
通讯作者:程茂华(1990-),男,江西乐平市人,助教,主要从事机器学习和数据挖掘方向的研究。

基于这两种思想比较,更多的学者致力于研究多分类器融合方法。常规的多分类器融合技术包括多数投票法[6],人工神经网络法,加权平均值法,决策模板[7]和D—S证据理论[8],行为—知识空间方法(BKS)[9]等。存在的问题是,一些基分类器存在实时性能不稳定的情况,所以在使用多分类器融合方法时容易受到这种基分类器的影响而导致性能的不稳定。因此,更多的研究者开始把目光投向基分类器的选择,特别是集成过程中的基分类器选择[10]。这些基于基分类器选择的多分类器系统方法不再局限于基于单个或基于全部基分类器进行融合决策,而是灵活性地组合部分互补性强且对实时样本有较高识别率的基分类器来完成融合决策[11]。

一些研究发现,不同分类器对于分类具有互补性,异分类器的融合能够有效提高分类精度以及推广能力,而提高分类器间相异性的手段之一就是采用具有互补分类信息的多个不同特征集[12-13]。这些不同特征集可以是同一特征集的不同子集,也可以是异类或不同特征空间中的特征子集[13]。

针对上述动态选择基分类器与分类器融合方法存在实时性能不稳定的问题,本文提出一种自适应子融合集成分类器方法,首先通过有放回地随机选择样本完成样本集采样,产生多个不同的训练集,然后通过线性判决思想(Fisher线性判决思想是:一个好的特征应该使类内离散度

尽可能小，而类间离散度尽可能大。）在不同训练子集中进行特征提取，并利用简单的分类器对输入的特征变量单独进行分类，最后基于本文提出的一种基分类器选择模型完成实时的子融合系统构建，并在该子融合系统上按分类的结果进行投票，选择得票最多的作为分类结果输出。

1　问题定义

多分类器系统作为一种集成分类算法（Ensemble learning），通过基分类器集合和组合规则或组合算法模型构成。根据基分类器决策输出信息的不同，多分类器系统一般被划分为三个不同的层次[14]：决策层融合（Abstract level），排序层融合（Rank level）和度量层融合（Measurement level）。在决策层融合层次上，各个基分类器的输出为某个确定的类别号；在排序层融合层次上，各个基分类器的输出为测试样本属于各类可能性的一个排序列表；在度量层融合层次上，各个基分类器的输出为测试样本属于各类的后验概率。

在实际应用中，大部分用于集成的基分类器可以获取类似于后验概率的中间度量值，如 k－NN 分类器可以利用测试样本到各类中心的最近邻距离来构建函数求取测试样本属于各类的可能性。这种可能性在同质基分类器构成的多分类器系统中可以作为基分类器选择的考虑因素。因此，本文主要研究度量层融合层次之上的多分类器联合方法。

1.1　数学定义

度量层融合层次的多分类器系统问题可以定义如下：

输入：

$[e_1(x)\quad e_2(x)\quad \cdots\quad e_K(x)]$：各基分类器对样本 x 的识别输出，其中，$e_k(x)=[\omega(C_1)\quad \omega(C_2)\quad \cdots\quad \omega(C_M)](k\in\{1,2,\cdots,K\})$，$\omega(C_i)\in[0,1]$，$\omega(C_i)(i\in\{1,2,\cdots,M\})$ 为后验概率、隶属度或某种模糊测度，说明样本 x 归属于各类的程度。

输出：

$E(x)=C_i$：多分类器系统识别样本所归属的类别，其中 $i\in\{1,2,\cdots,M\}$。

输出结果的获取可以通过多种不同形式实现，常见的有提取最大值、计算平均值和加权平均等。

1.2　相关定义

定义 1：对于样本 x，多分类器系统中各基分类器将其归属为类 C_i 的程度为 $[\omega_1(C_i)\quad \omega_2(C_i)\quad \cdots\quad \omega_K(C_i)]$，则对于某个分量 $\omega_k(C_i)$ 的类内决策支持度记为 $wDS(\omega_{ki})=1-\dfrac{\sum_{j=1}^{K}V_j\times|\omega_k(C_i)-\omega_j(C_i)|}{\sum_{j=1}^{K}V_j}$，其中，$i\in\{1,2,\cdots,M\}$，$V_j$ 为基分类器在分类器系统中的权重。

上述定义中，分量 $\omega_k(C_i)$ 与 $\omega_j(C_i)$ 的距离越小，说明它们之间的决策支持度越大。反之，则说明决策支持度越小。

定义 2：对于样本 x，多分类器系统中各基分类器 $e_k(k\in\{1,2,\cdots,K\})$ 的实时决策支持度记为 $DS(e_k)=\sum_{i=1}^{M}wDS(\omega_{ki})$。

定义 3：对于样本 x，多分类器系统中各基分类器 $e_k(k\in\{1,2,\cdots,K\})$ 的实时决策置信度记为 $DC(e_k)=\sum_{i=1}^{M}[(\omega_k(C_i)-0.5)\times 2]^2$。

上述定义中，第 k 个基分类器识别样本 x 归属于 C_i 类的程度 $\omega_k(C_i)$ 越靠近 [0，1] 区间中值 0.5，其决策置信度越小．反之，则说明决策置信度越大。

2　自适应子融合系统

自适应子融合系统可以针对不同的输入样本，动态挑选出不同数目的基分类器组成子融合系统进行样本识别。根据上述实时决策支持度和实时决策置信度的定义，设计基分类器动态挑选的策略，其过程为：首先提取实时决策支持度最高的基分类器，然后在多分类器系统中将其它基分类器的实时决策置信度一一与该基分类器的实时决策置信度进行比较，动态选择出比该基分类器实时决策置信度高的基分类器，并一起构成子融合系统，最后通过简单多数投票决定输入样本所归属的类别号。

为了提高多分类器系统的泛化能力，自适应子融合系统通过有放回随机选择多个不同的训练集，并在这些训练集上通过线性判决思想随机动态地提取特征构成各基分类器训练的特征子集。自适应子融合系统的方法模型框架如图 1 所示。训练样本和训练特征集的差异保证了多分类器系统中基分类器的互补性。

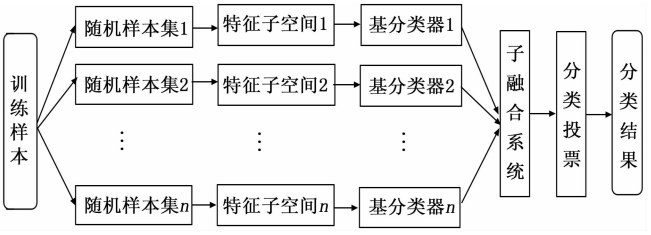

图 1　自适应子融合集成分类器方法模型

2.1　特征子集生成

在每个随机训练样本集基础上随机提取有较优线性可分性的特征子集，首先在特征集上随机地限定特征提取范围，该提取范围为随意的部分特征组合，以提高基分类器的差异性。然后，在随机挑选出第一个特征的基础上利用线性判决思想在这些随机提取的特征组合中通过迭代重组出线性可分性较强的特征子集。具体特征子集生成算法如算法 1 所示。

算法 1：特征选择：

Input：特征集 F.

Output：特征子集 S.

1)获取特征集 F 的特征个数 m；

2)初始化：Lsd=0，max_Lsd=0，first_i=0，S= φ ，i=0；

3)随机生成长度为 m 的二进制字符串 a；

4)在 a 中随机选择值为 1 的某个位置 first_i;
5)S=S ⋃ {F[first_i]};
6)max_Lsd=calculate_Lsd(S);
7)while i<m
8) if (a[i]==1 && i! =first_i) then
9) Lsd=calculate_Lsd(S ⋃ { F[i]});
10) if Lsd>max_Lsd then
11) S=S ⋃ {F[i]}; max_Lsd=Lsd;
12) end if
13) end if
14) i++;
15)end while
16)return(S,a).

其中，步骤 6) 中 calculate _ Lsd 函数为特征集输入参数 S 在当前随机样本集中的线性可分度，线性可分度 Lsd 的计算公式如式 (1) 所示。其中，c 为特征集 S 存在的类别数，X_i 为当前随机样本集中属于第 i 类的样本集合。

$$Lsd(S)=\frac{\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}(\overline{X}_i-\overline{X}_j)(\overline{X}_i-\overline{X}_j)}{\sum_{i=1}^{c}\sum_{x^j\in X_i}(x^j-\overline{X}_i)(x^j-\overline{X}_i)} \tag{1}$$

特征子集生成算法在自适应子融合系统中是基于多个不同样本集分别实现的，其实现过程可以并行处理。因此，有可能存在相同的特征子集被不同基分类器提取。本文通过两种不同的策略来优化提取的特征子集，提高基分类器的差异性。这两种策略分别是变异策略和交叉策略，具体方法如下所示：

变异策略：在特征选择向量 $a=(a_1, a_2, \cdots, a_n)^T$ 中随机选择一块区域，如 $(a_i, a_{i+1}, \cdots, a_j)^T$，然后进行取反运算，即：$(a_i,a_{i+1},\cdots,a_j)^T \leftarrow (\bar{a}_i,\bar{a}_{i+1},\cdots,\bar{a}_j)^T$。

交叉策略：随机选择一个不同的特征选择向量 a_2，在 a_2 中随机选择一个交叉区域，将 a 的相应交叉区域由 a_2 交叉区域代替。

例如，存在相同特征子集的特征选择向量为 $a=10011100$，选择的 a_2 为 $a_2=00100110$，交叉区域为 0011，则进行交叉操作后有：$a=10000110$。

通过双重循环将所有生成的特征子集进行比较，存在相同的特征子集进行 1 次或多次变异和交叉操作，直至得到一个与现有所有特征子集不重复的新特征子集。

2.2 基分类器动态选择

在随机样本和特征子空间生成后，分别训练基分类器，因为自适应子融合系统基于 1.2 节中定义的实时决策支持度和实时决策置信度动态选择集成，所以动态选择基分类器操作在测试阶段进行。

首先通过多分类器系统中的各个基分类器对输入测试样本进行分类识别，然后分别计算各基分类器的实时决策支持度 DS，并从中挑选出获得当前实时决策支持度最高的基分类器，将其作为自适应子融合系统的基分类器，并用该基分类器的实时决策置信度与其它基分类器的实时决策置信度进行比较，进一步挑选出实时决策置信度比其高的基分类器作为自适应子融合系统的成员，完成用来融合决策的子系统构建，算法流程如下：

算法 2：基分类器动态选择.
Input：分类器集合 E.
Output：分类器子集合 S.
1)初始化：$S=\varphi$;
2)从 E 中选择当前样本识别中 DS 最高的基分类器 e_c;
3)$S=\{e_c\}$;
4)$E=E-\{e_c\}$;
5)$\theta=DC(e_c)$;
6)while E! =NULL
7) if DC(E[0])>θ then
8) $S=S\bigcup\{e_i\}$;
9) end if
10) $E=E-\{e_i\}$;
11)end while
12)return(S).

该方法对于输出结果带有类似后验概率的分类器进行直接软迭代集成，对于其他输出形式的基分类器需要先将其输出值转化到 [0，1] 上的可信度，然后再利用算法。本文定义其输出值转化方法为：

$$e_k(x)=[P_k(C_1\mid x),P_k(C_2\mid x),\cdots,P_k(C_M\mid x)]$$

其中：各决策分量满足：$P_k(C_i\mid x)\in[0,1]$，并且 $\sum_{i=1}^{M}P_k(C_i\mid x)=1(\forall i\in\{1,2,\cdots,K\})$。

基于上述方法可以得到多分类器系统的决策矩阵如下：

$$\begin{bmatrix} P_{s(1)}(C_1\mid x) & P_{s(1)}(C_2\mid x) & \cdots & P_{s(1)}(C_M\mid x) \\ P_{s(2)}(C_1\mid x) & P_{s(2)}(C_2\mid x) & \cdots & P_{s(2)}(C_M\mid x) \\ \cdots & \cdots & \ddots & \cdots \\ P_{s(s)}(C_1\mid x) & P_{s(s)}(C_2\mid x) & \cdots & P_{s(s)}(C_M\mid x) \end{bmatrix}$$

2.3 融合决策过程

自适应子融合集成分类方法融合了一系列基分类器的分类结果，直接采用多数投票法来决定识别结果，让当前被自适应子融合系统选中的基分类器都对输入的特征向量进行投票，汇总各类得票数，找出其中拥有票数最多的类别作为融合系统对该特征向量识别的类别。

3 实验结果与分析

本实验使用的是 UCI 机器学习数据库中的四类数据集进行相关测试。数据集样本如表 1 所示。实验数据属于多分类样本数据集，需限定使用方法为多分类方法，以保证分类的效果，实验基分类器如表 2 所示。有效划分训练集与测试集比重往往可以提高分类的效率，参照先验知识且经过多次试验测试集与训练集比例，最终发现 30%作为训练集、70%作为测试集的实验效果最好，因此我们将各类数据集分别按照 0.3 的比例划分。

本文将分类准确率作为衡量融合集成分类器方法识别效果的衡量标准，具体方法是测试集中分类正确数量占总测试集的百分比，公式如式 (2)：

$$Accuracy = \frac{N_k}{N_c} \times 100\% \quad (2)$$

其中：N_k表示测试集中分类正确的数量，N_c 表示测试集的总数。

表 2 实验结果数据表明，本文提出的自适应子融合集成分类方法与其他基分类器比较，本文方法的识别效果更优，在所用数据集都得到了有效提升。同时，表 2 也表明了在 Vehicle 数据集、Glass 数据集上一些基分类器识别性能较差的现象。验证了本文前面提到的基分类器实时稳定性差从而导致一些融合方法的性能不稳定的问题。本文提出的自适应子融合集成多分类器方法从表 3 中明显证明识别性能优于其他两种多分类器融合方法，并且在 Wine 数据集和 Vehicle 数据集效果提升稍好于其他两类数据集。通过表 2、表 3，我们可以得出以下结论：多分类问题，数据类别越多，分类的准确率越高，即分类效果越好。

表 1　实验的四类数据集

数据集	类数	特征数	样本数
Wine	3	13	178
Vehicle	4	18	846
Glass	6	10	214
Statlog	6	36	4435

表 2　本文方法与基本分类器识别准确度比较　　%

基分类器	Wine	Vehicle	Glass	Statlog
Parzen 分类器	84.90	81.67	93.24	86.35
k－NN 分类器	81.90	84.23	94.20	87.76
OQDF－分类器	80.20	67.12	82.50	83.21
本文方法	89.10	89.36	98.81	90.12

表 3　本文方法与其他多分类器联合方法识别准确度比较%

多分类器联合方法	Wine	Vehicle	Glass	Statlog
多数投票法	86.72	86.69	97.65	89.15
平均值规则法	87.65	87.03	98.10	89.36
本文方法	89.10	89.36	98.81	92.12

从图 2 中，我们可以直观看到各基分类器与多分类器融合方法的分类性能，并且在分类性能上多分类器融合方法普遍优于基分类器方法，本文方法在识别准确率上同样高于所比较的其他分类融合方法。

4　结论

本文基于 Fisher 线性判决思想来完成随机特征子集内的特征选择有效提高基分类器的差异性，结合决策支持度 DS 与决策置信度 DC 完成基分类器的动态选择，并让每一个被选中的基分类器对输入的特征向量进行投票，计算所有投票数，获取子融合系统中投票数最多的类别作为当前输入样本的分类结果，有效提高了分类器识别性能。实验结果表明，本文研究的度量层融合层次之上的多分类器联合方法能获得较好的识别性能，较单个分类器的识别准确

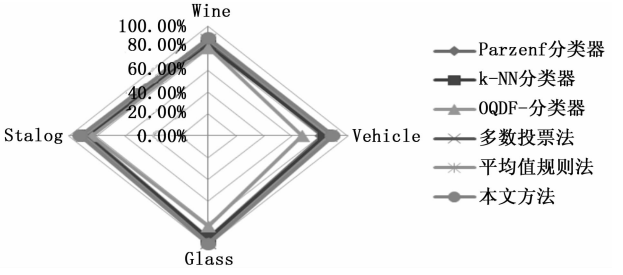

图 2　各基分类器与分类器融合方法性能比较

度都有所提高。

我们的工作存在如下不足：在未来的研究中需要解决的问题，如基分类器选择当前实时决策支持度最高者，是否可以通过先验概率或判别函数确定基分类器会有更好的分类效果。

参考文献：

[1] 李艳秋，任福继，胡　敏. 动态模糊密度的多分类器融合算法 [J]. 电子学报，2018，46 (5)：1246－1252.

[2] 周　星，刁兴春，曹建军，等. 基于重采样和集成选择的适用于实体识别的多分类器系统 [J]. 数字采集与处理，2017 (5)：931－938.

[3] 贾澎涛，李　阳. 基于选择性集成分类器的面部表情识别研究 [J]. 计算机应用研究，2017，34 (12)：3825－3827.

[4] 楚浩宇，高　萌，刘永生. 基于并行组合分类器的脱机手写体数字识别 [J]. 计算机技术与发展，2018 (3)：105－108.

[5] 郝红卫，王志彬，殷绪成，等. 分类器的动态选择与循环集成方法 [J]. 自动化学报，2011，37 (11)：1290－1295.

[6] Zia M S，Hussain M，Jaffar M A. A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier [J]. Multimedia Tools and Applications，2018，77 (19)：25537－25567.

[7] 唐　彪，金　炜，符冉迪，等. 多稀疏表示分类器决策融合的人脸识别 [J]. 电信科学，2018 (4)：31－40.

[8] 张　扬，杨建华，侯　宏. 一种基于证据理论的数据分聚类融合算法 [J]. 科学技术与工程，2018 (1)：54－58.

[9] 王　江，孙美凤，张　炜. 基于行为知识空间的多分类器网络流量分类方法 [J]. 扬州大学学报：自然科学版，2016 (4)：54－57.

[10] 袁　立，穆志纯. 基于子分类器融合的部分遮挡人耳识别 [J]. 仪器仪表学报，2011，32 (1)：186－193.

[11] 朱　波，陈　科，徐　君，等. 平均分布集成策略：一种新的分类器融合方法 [J]. 小型微型计算机系统，2016，(7)：1546－1550.

[12] Borja S P，et al . Ensemble feature selection for rankings of features [J]. Springer International Publishing，2015，9095：29－42.

[13] 陶晓玲，亢蕊楠，刘丽燕. 基于选择性集成的并行多分类器融合方法 [J]. 计算机工程与科学，2018 (5)：787－792.

[14] 刘　明，袁保宗，苗振江，等. 从局部分类精度到分类置信度的变换 [J]. 计算机研究与发展，2008，45 (9)：1612－1619.