

# 基于差异性的分类器集成: 有效性分析及优化集成

杨春<sup>1</sup> 殷绪成<sup>1,2</sup> 郝红卫<sup>3</sup> 闫琰<sup>1</sup> 王志彬<sup>1,4</sup>

**摘要** 差异性分类器集成具有高泛化能力的必要条件. 然而, 目前对差异性度量、有效性及分类器优化集成都没有统一的分析和处理方法. 针对上述问题, 本文一方面从差异性度量方法、差异性度量有效性分析和相应的分类器优化集成技术三个角度, 全面总结与分析了基于差异性的分类器集成. 同时, 本文还通过向量空间模型形象地论证了差异性度量的有效性. 另一方面, 本文针对多种典型的基于差异性的分类器集成技术 (Bagging, boosting GA-based, quadratic programming (QP)、semi-definite programming (SDP)、regularized selective ensemble (RSE)) 在 UCI 数据库和 USPS 数据库上进行了对比实验与性能分析, 并对如何选择差异性度量方法和具体的优化集成技术给出了可行性建议.

**关键词** 分类器集成, 差异性, 有效性分析, 优化

**引用格式** 杨春, 殷绪成, 郝红卫, 闫琰, 王志彬. 基于差异性的分类器集成: 有效性分析及优化集成. 自动化学报, 2014, 40(4): 660–674

**DOI** 10.3724/SP.J.1004.2014.00660

## Classifier Ensemble with Diversity: Effectiveness Analysis and Ensemble Optimization

YANG Chun<sup>1</sup> YIN Xu-Cheng<sup>1,2</sup> HAO Hong-Wei<sup>3</sup> YAN Yan<sup>1</sup> WANG Zhi-Bin<sup>1,4</sup>

**Abstract** Diversity is a necessary condition for high generalization capability in classifier ensemble. However, there exists no uniform analysis and operation methods for diversity measure, effectiveness analysis or ensemble optimization. To solve these issues, on the one hand, classifier ensemble with diversity is comprehensively summarized and analyzed from three aspects, i.e., diversity measurement methods, effectiveness analysis for diversity measurement methods and optimization techniques for classifier ensemble. Moreover, the effectiveness of diversity is also demonstrated by the vector space model. On the other hand, comparative experiments and analysis have been performed on UCI data sets and USPS data set with a variety of typical classifier ensemble methods (Bagging, boosting, GA-based, quadratic programming (QP), semi-definite programming (SDP), regularized selective ensemble (RSE)). Finally, we give some suggestions on how to select diversity measurement methods and optimization techniques in ensemble.

**Key words** Classifier ensemble, diversity, effectiveness analysis, optimization

**Citation** Yang Chun, Yin Xu-Cheng, Hao Hong-Wei, Yan Yan, Wang Zhi-Bin. Classifier ensemble with diversity: effectiveness analysis and ensemble optimization. *Acta Automatica Sinica*, 2014, 40(4): 660–674

分类器集成是机器学习、模式识别和数据挖掘领域的一个重要的研究方向<sup>[1–2]</sup>. 其基本思想是先

构建多个不同的基分类器, 再综合它们的结果进行决策, 从而取得比单个基分类器更好的识别性能. Lebanon 等<sup>[3]</sup>指出, 理论上可以将足够多的弱分类器 (识别率略好于随机猜测) 集成为识别精度很高的强分类器. 因而, 基分类器也常常被称为“弱分类器”. 虽然很多研究工作在弱分类器上进行, 但是集成学习也适用于强分类器的学习. 基分类器往往由已有的一些机器学习算法 (如神经网络<sup>[4–5]</sup>、朴素贝叶斯网络<sup>[6–7]</sup>、决策树<sup>[8–9]</sup> 和支持向量机<sup>[10–11]</sup> 等) 在训练集上训练得到. 大多数集成学习研究倾向于构建相同类型和参数设定的基分类器, 但是也有一些集成学习<sup>[12–13]</sup> 使用不同类型的基分类器.

在集成学习领域内已经开展了大量的研究工作, 并取得了良好的研究成果<sup>[8, 14–16]</sup>. 该领域的代表性方法, 如 Bagging<sup>[17]</sup>、Boosting<sup>[18–19]</sup>、Stacking<sup>[20]</sup> 和 Random forests<sup>[21]</sup> 等在手写数字识别<sup>[22]</sup>、人脸

收稿日期 2012-09-24 录用日期 2013-01-11  
Manuscript received September 24, 2012; Accepted January 11, 2013

国家自然科学基金 (61105018, 61175020) 资助  
Supported by National Natural Science Foundation of China (61105018, 61175020)

本文责任编辑 刘成林  
Recommended by Associate Editor LIU Cheng-Lin

1. 北京科技大学计算机与通信工程学院计算机科学与技术系 北京 100083 2. 北京科技大学材料领域知识工程北京市重点实验室 北京 100083 3. 中国科学院自动化研究所 北京 100190 4. 国家农业信息化工程技术研究中心 北京 100097

1. Department of Computer Science and Technology, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083 2. Beijing Key Laboratory of Materials Science Knowledge Engineering, University of Science and Technology Beijing, Beijing 100083 3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190 4. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097

识别<sup>[23]</sup>、年龄识别<sup>[24]</sup>、指纹识别<sup>[25]</sup>、图像处理<sup>[26]</sup>、生物信息学<sup>[27]</sup>、网络数据挖掘<sup>[28]</sup>等实际模式分类问题中都取得了较好的效果. 研究表明<sup>[29-30]</sup>集成学习的成功来自于基分类器之间具有差异性. 大多数集成学习方法<sup>[31-34]</sup>都可以看作是生成或选择更具差异性的基分类器的过程.

目前存在至少 20 种关于什么是差异性的描述, 但是仍然没有统一的定义. 较为公认的定义来自 Dietterich<sup>[35]</sup>, 他认为差异性指代“对新的样本、分类器做出不同错分的趋势”. 随着不同差异性度量方法的出现, 差异性的内涵也越加丰富.

通常, 分类器集成中的差异性学习可分为两种途径: 1) 通过不同的数据训练不同的基分类器, 来隐性地使得各分类器具有差异性, 如 Bagging<sup>[17]</sup> 和 Boosting<sup>[18]</sup>; 2) 显性地考虑差异化, 以最大化某个与差异性相关的目标函数, 来集成不同的基分类器, 如 SDP (Semi-definite programming)<sup>[33]</sup> 和 RSE (Regularized selective ensemble)<sup>[34]</sup>. 本文主要考虑后者, 即基于显性差异化分析及其相应的多分类器优化集成技术.

差异性是高泛化能力集成的必要条件, 对于提高集成学习的泛化能力具有重要意义, 有关差异性的研究是研究集成学习的基础. 将某些基分类器间的统计信息作为差异性进行度量, 在应用中也取得了一定的效果. 近几年来, 很多学者通过显性差异化分析来进行多分类器集成研究. 基于差异性的分类器集成研究主要包括三个部分:

1) 差异性的度量方法. 寻找设计更能代表集成泛化误差的差异性方法. 人们从集成学习系统<sup>[36-38]</sup>、统计学<sup>[29-30, 39-41]</sup>、信息论<sup>[42]</sup>、软件工程<sup>[43]</sup>等多个领域提出了对差异性的度量方法.

2) 差异性的有效性分析. 研究差异性与集成性能之间的关系, 为基于差异性的集成优化提供理论依据. 文献 [6, 44, 45] 通过理论分析与实验, 指出差异性度量与泛化能力相关.

3) 基于差异性的优化集成. 根据差异性与集成之间的关系, 将差异度直接 (显性的) 用于指导基分类器的设计, 以及选择较大差异性的基分类器子集. Zhou 等<sup>[46]</sup> 针对最近邻基分类器, 提出了一种基于多模型干扰来产生差异化基分类器的集成方法. Yu 等<sup>[47]</sup> 在差异化的正则优化的基础上, 序列的训练和集成线性支持向量机. Li 等<sup>[48]</sup> 试图以 PAC (Probably approximately correct) 学习理论为框架, 从理论上分析差异性与投票集成的关系, 并提出了基于差异正则化的集成裁剪方法. 荆晓远等<sup>[49]</sup> 提出了最大有效互补原则. 郝红卫等<sup>[50]</sup> 提出了基于差异性分析的动态选择与循环集成方法.

尽管人们对差异性的研究取得了一定进展, 但

仍存在以下难点:

1) 尽管存在多种差异性度量方法, 但是每种差异性与集成泛化能力之间都不是紧密关联的. 在大多数情况下, 最大化差异性并不能得到识别精度更高的集成.

2) 对差异性与集成性能之间关系的研究都没有取得较好的结果. 直到目前为止, 没有可用于指导基于差异性进行集成优化的理论基础.

3) 传统对差异性度量的研究均是在静态集成的条件下, 这些方法在动态集成中将不再适用. 在静态集成中, 由于利用不同基分类器在整个样本空间上的分类趋势计算彼此之间的差异性, 因而得到差异性与待测样本无关. 但是在动态集成中, 要根据利用不同基分类器在待测样本有效领域内的分类趋势计算彼此之间的差异性, 此时, 得到差异性与待测样本相关, 因而传统差异性度量方法的适应性将降低.

由于上述难点, 目前, 仍然没有一种差异性度量方法, 被证明对于提高集成学习泛化能力是最有效的. 没有一致的差异性度量的有效性分析及其相应的基分类器优化集成对比. 针对这些难点和问题, 本文全面的阐述与总结了上述基于差异性度量的基分类器集成的三大核心内容.

值得强调的是, 论文从向量空间关系分析来直观形象地论证了差异性度量的有效性. 同时, 针对多种典型的基于差异性的基分类器集成技术 (Bagging<sup>[17]</sup>、Boosting<sup>[18]</sup>、GA-based<sup>[51]</sup>、QP (Quadratic programming)<sup>[34]</sup>、SDP<sup>[33]</sup>、RSE<sup>[34]</sup>) 进行了大量 UCI 数据库数据集的对比试验与性能分析, 并给出了差异性度量方法和优化集成技术选择的一些可行建议.

论文余下内容安排如下: 第 1 节给出本文的基本符号定义和说明; 第 2 节简要地总结了目前常用的差异性度量方法; 第 3 节详细地论述了差异度量的有效性分析; 第 4 节介绍常用的基于差异性的集成优化方法; 第 5 节针对多种典型的基于差异性的基分类器集成技术, 全面地阐述了大量的对比实验和性能分析, 并给出了基于差异性的基分类器集成研究与应用的部分可行指南; 最后, 总结了本文的工作.

## 1 基本符号定义和说明

在具体讨论前, 给出如下定义:

1) 样本集合  $D = \{x_1, x_2, \dots, x_N\}$ ,  $|D| = N$ , 其中, 第  $i$  个样本  $x_i$  的类别为  $y_i$ .

2) 所有样本的类别  $\Omega = \{\omega_1, \omega_2, \dots, \omega_C\}$ ,  $|\Omega| = C$ .

3) 基分类器集合  $H = \{h_1, h_2, \dots, h_L\}$ ,  $|H| = L$ .

4) 基分类器  $h_j$  对样本  $x_i$  的分类结果为  $h_j(x_i)$ .

5) 基分类器  $h_j$  对样本  $x_i$  的识别结果为  $O_{ij}$ . 当基分类器  $h_j$  将样本  $x_i$  分类正确 (即  $h_j(x_i) = y_i$ ) 时,  $O_{ij} = 1$ ; 否则  $O_{ij} = -1$ .

6) 对于样本  $x_i$  分类错误的基分类器比例  $l(x_i)$ .

7) 基分类器  $h_j$  的权重  $w_j$  满足:  $\sum_{j=1}^L w_j = 1$ ,  $w_j \geq 0, \forall j$ .

8) 样本  $x_i$  的边界  $m_i = \sum_{j=1}^L w_j O_{ij}$ , 具体说明参见本文第 3.3 节.

9) 基分类器  $h_j$  对样本集  $D$  的识别结果  $\mathbf{R}_j = [O_{1j}, O_{2j}, \dots, O_{Nj}]^T$ , 分类器集合对样本集  $D$  的识别结果  $\mathbf{R} = [\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_L]$ .

10) 基分类器的平均精度  $P$ .

11) 基分类器的成对差异性  $Div$ , 其平均差异性  $\overline{Div}$ .

## 2 差异性度量方法

差异性多被用于度量基分类器之间的统计关系, 如相关性、互补性、熵等. 一般来说, 差异性常常分为两类: 成对差异性和非成对差异性. 成对差异性先计算  $L$  个分类器两两之间的差异性, 再计算  $L(L-1)/2$  对分类器的平均差异性作为集成的差异性; 而非成对差异性则直接使用所有的分类器计算集成的差异性.

在度量差异性时, 往往考虑基分类器对样本的识别结果. 这是因为它不需要对数据的先验知识和基分类器进行假定. 在这种模型下得到的结果可以推广到其他集成方法中.

### 2.1 成对差异性度量方法

成对差异性先计算两两基分类器之间的统计关系, 因而依赖于两基分类器的联合分布. 基分类器对  $h_i, h_j$  的联合分布矩阵如表 1<sup>[30]</sup> 所示.

表 1 两基分类器之间的联合分布<sup>[30]</sup>

Table 1 Joint distribution for two base classifiers

	$h_j(x_k) = y_k$	$h_j(x_k) \neq y_k$
$h_i(x_k) = y_k$	$a_{ij}$	$b_{ij}$
$h_i(x_k) \neq y_k$	$c_{ij}$	$d_{ij}$

对于  $L$  个基分类器, 若基分类器对  $\{h_i, h_j\}$  的成对差异性为  $Div_{ij}$ , 则其集成的差异性如式 (1).

$$\overline{Div} = \sum_{i=1}^L \sum_{j=1, j \neq i}^L Div_{ij} \quad (1)$$

下面是 5 种常用的成对差异性.

#### 2.1.1 相关系数

相关系数<sup>[30]</sup> (Correlation coefficient,  $\rho$ ) 源自

统计学, 其计算式如式 (2).

$$\rho_{ij} = \frac{a_{ij}d_{ij} - b_{ij}c_{ij}}{\sqrt{(a_{ij} + b_{ij})(c_{ij} + d_{ij})(a_{ij} + c_{ij})(b_{ij} + d_{ij})}} \quad (2)$$

式 (2) 中,  $\rho$  取值范围  $[-1, 1]$ . 相关系数为 0 时, 两个基分类器独立.

#### 2.1.2 Q 统计量

Yule<sup>[40]</sup> 提出的  $Q$ -统计量 ( $Q$ -statistic,  $Q$ ) 可以看作是相关系数的一种简化运算. 其计算式如式 (3).

$$Q_{ij} = \frac{a_{ij}d_{ij} - b_{ij}c_{ij}}{a_{ij}d_{ij} + b_{ij}c_{ij}} \quad (3)$$

式 (3) 中,  $Q$  取值范围  $[-1, 1]$ .  $Q$ -统计量为 0 时, 两个基分类器独立.

#### 2.1.3 成对 Kappa 度量

在进行离线学习时, 常常使用矩阵运算进行数据的批量学习. 此时, 使用成对 Kappa<sup>[39]</sup> (Pairwise-kappa,  $\kappa_p$ ) 的运算量要少于  $Q$ -统计量. 成对 Kappa 越小, 基分类器的相关性越小. 其计算式如式 (4).

$$\kappa_{p_{ij}} = \frac{2(a_{ij}d_{ij} - b_{ij}c_{ij})}{(a_{ij} + b_{ij})(b_{ij} + d_{ij}) + (a_{ij} + c_{ij})(c_{ij} + d_{ij})} \quad (4)$$

#### 2.1.4 不一致度量

Skalak<sup>[36]</sup> 从“差异”这个概念出发, 提出了不一致度量 (Disagreement,  $dis$ ). 不一致度量越大, 基分类器间差异性越大, 但是平均精度也越低. 其计算式如式 (5).

$$dis_{ij} = b_{ij} + c_{ij} \quad (5)$$

#### 2.1.5 双错度量

为了避免不一致度量中, 差异性与平均精度的相关关系, Giacinto 等<sup>[37]</sup> 提出了双错度量 (Double fault,  $df$ ). 双错度量越小, 集成精度越大. 其计算式如式 (6).

$$df_{ij} = d_{ij} \quad (6)$$

### 2.2 非成对差异性度量方法

成对差异性计算集成差异性时, 忽略了基分类器整体信息. 非成对差异性考虑基分类器整体的性能, 从而获得更精确的统计信息. 非成对度量需要先计算对于样本  $x_i$  分类错误的基分类器比例  $l(x_i)$ . 下面是 6 种常见的非成对差异性度量方法.

#### 2.2.1 熵

Cunningham 等<sup>[42]</sup> 引入信息论中熵 (Entropy,  $Ent$ ) 的概念. 熵越大, 差异性越大. 其计算式如式

(7).

$$Ent = \frac{1}{N} \sum_{i=1}^N \frac{1}{L - \left\lfloor \frac{L}{2} \right\rfloor - 1} \min(l(x_i), L - l(x_i)) \quad (7)$$

### 2.2.2 KW-差异

KW-差异 (Kohavi-Wolpert variance, *KW*) 是 Kohavi 等<sup>[38]</sup> 提出的偏差-方差分解在识别问题中的应用. 定义样本  $x_i$  的方差如式 (8). 其中,  $C$  是类别数.

$$variance_x = \frac{1}{2} \left( 1 - \sum_{i=1}^C P\{y = \omega_i | x\}^2 \right) \quad (8)$$

当只考虑识别结果时, 式 (8) 等价于式 (9).

$$KW = \frac{1}{NL^2} \sum_{i=1}^N l(x_i)(L - l(x_i)) \quad (9)$$

式 (9) 中, *KW*-差异越大, 差异性越大.

### 2.2.3 Kappa 度量

Dietterich<sup>[29]</sup> 使用 Kappa 度量 (Kappa,  $\kappa$ ) 分析集成泛化能力与差异性之间的关系. Kappa 度量的计算式如式 (10).

$$\kappa = 1 - \frac{\sum_{i=1}^N l(x_i)(L - l(x_i))}{L N (L - 1) P(1 - P)} \quad (10)$$

式 (10) 中, Kappa 越小, 基分类器的相关性越小.

### 2.2.4 难度度量

Hansen 等<sup>[41]</sup> 提出的难度度量 (Difficulty,  $\theta$ ) 首先, 计算随机变量  $X$  的分布. 其中,  $X \in 0, 1/L, 2/L, \dots, 1$ , 表示随机选择样本  $x$ , 对其分类正确的基分类器个数比例所占集成的比率.

难度度量计算  $X$  的方差, 如式 (11).

$$\theta = \text{var}(X) \quad (11)$$

式 (11) 中, 难度度量越小, 差异性越大.

### 2.2.5 广义差异性

Partridge 等<sup>[43]</sup> 提出广义差异性 (Generalized diversity, *GD*). 广义差异性的计算式如式 (12).

$$GD = 1 - \frac{P(2)}{P(1)} \quad (12)$$

式 (12) 中,  $P(1)$  表示一个基分类器分类错误的概率,  $P(2)$  表示两个基分类器分类错误的概率. *GD* 越大, 差异性越大.

### 2.2.6 一致失效差异性

在广义差异性的基础上, Partridge 等<sup>[43]</sup> 提出了一致失效差异性 (Coincident failure diversity, *CFD*). 一致失效差异性的计算式如式 (13).

$$CFD = \begin{cases} 0, & p_0 = 1 \\ \frac{1}{1 - p_0} \sum_{i=1}^L \frac{L - i}{L - 1} p_i, & p_0 < 1 \end{cases} \quad (13)$$

式 (13) 中,  $p_0$  表示基分类器全部分类正确的概率. *CFD* 越大, 差异性越大.

上述 11 种差异性度量方法, 总结如表 2.

表 2 11 种差异性度量方法一览  
Table 2 A review of 11 diversity methods

度量方法	符号	成对/ 非成对	最优趋势 (↑ / ↓)	取值范围	最早使用	来源领域
相关系数	$\rho$	成对	↓	[-1, 1]	Sneath(1973)	统计学
Q-统计量	$Q$	成对	↓	[-1, 1]	Yule(1900)	统计学
成对 Kappa	$\kappa_p$	成对	↓	[-1, 1]	Dietterich(2000)	统计学
不一致度量	$dis$	成对	↑	[0, 1]	Skalak(1996)	其他
双错度量	$df$	成对	↓	[0, 1]	Giacinto(2000)	其他
熵	$Ent$	非成对	↑	[0, 1]	Cunningham(2000)	信息论
KW-差异	$KW$	非成对	↑	[0, 0.5]	Kohavi(1996)	其他
Kappa	$\kappa$	非成对	↓	[-1, 1]	Fleiss(1981)	统计学
难度度量	$\theta$	非成对	↓	[0, 0.25]	Hansen(1990)	统计学
广义差异性	$GD$	非成对	↑	[0, 1]	Partridge(1997)	软件工程
一致失效差异性	$CFD$	非成对	↑	[0, 1]	Partridge(1997)	软件工程

## 3 差异性度量的有效性分析

对于上述多种差异性度量方法, 在实际应用中, 常常使用其中的一种或几种对集成的泛化能力进行度量. 选择什么样的差异性度量更有效, 是差异性度量的重要研究内容. 本节首先综述了差异性度量方法选择的一般准则; 然后, 指出依据最大化差异性进行优化的原理及其局限; 最后, 为了克服这种局限性, 从 Margin 分析和向量空间关系两个方面分析差异性度量的有效性.

### 3.1 差异性度量的选择

为了选择更有效的差异性度量方法, 人们分析差异性度量之间, 以及差异性度量和集成方法之间的关系.

Kuncheva 等<sup>[30]</sup> 指出: Kappa( $\kappa$ )、KW-差异 (*KW*) 和不一致度量 ( $\overline{Dis}$ ) 之间存在如式 (14) 所示关系.

$$\begin{aligned} \kappa &= 1 - \frac{\overline{Dis}}{2P(1 - P)} = \\ &= 1 - \frac{L}{L - 1} \frac{KW}{P(1 - P)} \\ KW &= \frac{L - 1}{2L} \overline{Dis} \end{aligned} \quad (14)$$

此外, Kuncheva 等<sup>[30]</sup> 在 BCW (Breast cancer wisconsin) 数据库上进行多组实验, 依据度量方法之间的相关性, 将常用差异性度量方法分为三组  $\{df\}, \{CFD\}, \{\rho, Q, dis, \kappa p, Ent, KW, \kappa, \theta, GD\}$ . 根据性能和运算复杂程度, Kuncheva 等推荐使用  $Q$  进行差异性度量.

Shipp 等<sup>[6]</sup> 通过统计差异性度量方法和集成方法之间的相关性, 指出使用  $df$  和  $\theta$  度量泛化能力更有效.

在排序裁剪算法<sup>[52]</sup> 中, 偏好使用 Kappa (或成对 Kappa 方法).

### 3.2 最大化差异性及其局限性

对于被选择的差异性度量方法, 实际应用中常常假定: 大的差异性意味着更好的泛化能力, 并使用最大化差异性的方法来生成或选择基分类器. 但是, 理论分析和实验结果均表明, 最大化差异性并不意味着最好的泛化能力.

Krogh 等<sup>[53]</sup> 指出: 对于回归问题, 误差函数可以进行歧义性分解, 如式 (15).

$$E = \bar{E} - \bar{A} \quad (15)$$

式 (15) 中,  $E$  为集合误差,  $\bar{E}$  为基分类器平均误差,  $\bar{A}$  为基分类器与整体的差异性.

式 (15) 表明, 当平均误差一定时, 增大差异性, 则泛化误差减小, 系统泛化能力提高. 但是, 平均误差和差异性之间不独立: 增大差异性的同时, 平均误差也会增大, 减小平均误差的同时, 差异性也会减小. 在平均误差 (或精度) 和差异性之间, 存在一个平衡 (Trade-off) 状态, 使得泛化能力较好.

因此, 如何找到在平均误差 (或精度) 和差异性之间的平衡状态, 对提高泛化能力具有重要意义. 现有的研究多是从以下两个方面考虑精度和差异性之间平衡状态的搜索问题.

一方面是在分类器集合一定的情况下, 对差异性和精度进行多目标规划.

使用遗传算法搜索多目标 (差异性、精度) 的帕累托前沿的中点. 其实验结果表明: 虽然最大化差异性的方法性能较好, 但是同时考虑差异性和精度的方法性能上更好. 使用最大化差异性和平均精度的线性或指数组合, 这些方法都针对具体实际问题, 不具有理论的支持. Trawinski 等<sup>[51]</sup> 则通过搜索平均精度  $P$  和差异性  $D$  的线性组合  $P + \lambda D$  的最大值, 实现精度与差异性的多目标搜索. Yin 等<sup>[54]</sup> 设计了基分类器差异度贡献模型, 在保持较高集成精度的前提下, 通过遗传算法来择优的选择基分类器, 使得整体集成具有更大的差异性.

另一方面是在分类器生成过程中, 在保证精度

一定的情况下, 生成更具有差异性的基分类器.

Liu 等<sup>[32]</sup> 提出了负相关学习 (Negatively correlated learning, NCL), 通过在损失函数 (精度) 中加入与差异性有关的正则项, 生成差异性更高的基分类器. Abbass<sup>[55-56]</sup> 提出 MPANN (Mesmelic pareto artificial neural networks) 方法, 根据任意神经网络之间对应网络权值差更新第三个神经网络的权值. Chandra 等<sup>[57]</sup> 通过结合 MPANN 和 NCL, 提出 DIVACE (Diverse and accurate ensemble learning algorithm) 方法. 该方法使用 NCL 生成与当前基分类器集合差异性较大的基分类器, 然后, 用 MPANN 方法更新神经网络的权值. 此外, Lee 等<sup>[4]</sup> 提出了并行使用负相关学习的方法.

虽然对于精度和差异性的平衡状态的研究有了一定进展, 可惜的是, 目前还没有文献构造一种精度和差异性的模型, 并从理论上分析该模型与集成的泛化误差界之间的关系. 此外, 人们研究最大化差异性度量有效的原因, 希望在精度一定的情况下, 通过最大化差异性实现泛化能力的提高.

### 3.3 最大化差异性与边界之间的关系

针对最大化差异性度量为什么有效的问题, Tang 等<sup>[45]</sup> 引入边界 (Margin) 概念, 并指出: 在平均精度为常量时, 最优化差异性度量等价于边界最大化.

集成边界 (Margin of ensembles) 是由 Schapire 等<sup>[31]</sup> 在解释 Boosting 算法成功时引入的概念. 对于第  $i$  个样本  $(x_i, y_i)$ , 记  $v_{i,\omega}$  为分类结果为  $\omega$  的基分类器权重,  $v_{i,y_i}$  为分类正确的基分类器权重, 则样本  $x_i$  的边界  $m_i$  定义式如式 (16).

$$m_i = v_{i,y_i} - \sum_{\omega \neq y_i} v_{i,\omega} \quad (16)$$

给定一组基分类器的集合和权重  $w$ , 基分类器  $h_j$  对样本  $x_i$  的识别结果为  $O_{ij}$  (见本文第 1 节定义), 则式 (16) 等价于式 (17).

$$m_i = \sum_{j=1}^L w_j O_{ij} \quad (17)$$

实验表明, 集成的泛化能力与其在训练集上的集成边界有关. Schapire 等<sup>[31]</sup> 指出较大的边界有助于提高集成的泛化能力. 随后, Rätsch 等<sup>[58]</sup> 从最小边界 (Minimum margin) 的角度考虑集成的泛化能力, 并提出了 Employing boosting. Vapnik<sup>[59]</sup> 指出达到最大化最小集成边界 (Largest minimum margin) 的集成, 其泛化错误边界最优.

Tang 等<sup>[45]</sup> 指出 6 种差异性度量方法  $\{dis, df, KW, \kappa, GD, \theta\}$  都可以表示成式 (18) 的形

式.

$$div = a - \left( bP + c \sum_{i=1}^N l(x_i)^2 \right) \quad (18)$$

其中,  $div$  为集成差异性,  $a, b, c$  为常系数,  $P$  为基分类器集合平均精度,  $\sum_{i=1}^N l(x_i)^2$  与集成边界相关. 式 (18) 表明, 集成边界可以看作是平均误差和差异性之间的一种平衡状态.

Tang 等<sup>[45]</sup> 通过理论分析, 给出最大化差异性和边界最大化的一致条件:

**定理 1**<sup>[45]</sup>. 当基分类器平均精度  $P$  为常量时, 若所有样本可以被相同数目的基分类器分类正确, 则差异性  $div$  达到最大. 即式 (19).

$$l(x_i) = L(1 - P), \quad \forall i \quad (19)$$

此时, 集成的泛化误差界在  $\min(m_i) = L(2P - 1)$  处达到.

**定理 2**<sup>[45]</sup>. 当基分类器平均精度  $P$  为常量且达到最大化差异性时, 在基分类器上的最大化差异性等价于在训练集上最大化最小边界.

因此, 差异性度量可以视为搜索最大化最小边界的隐性方法. 然而, Tang 等<sup>[45]</sup> 通过大量实验发现上述差异性与最小边界并不是完全正比关系, 提高差异性不能确保加大最小边界.

另一方面, 使用最小边界估计的泛化误差界, 仍然不是很紧. 在不知道真实数据分布的情况下, 使用最大化最小边界容易发生学习现象. Wang 等<sup>[60]</sup> 使用与边界分布相关的均衡边界 (E-margin) 替代最小边界, 进行泛化误差界的估计. Gao 等<sup>[61]</sup> 指出最小化边界和均衡边界均是用第  $k$  小的边界进行泛化误差界估计, 并提出一种基于边界分布的泛化误差界估计, 该估计的泛化误差下界更紧.

### 3.4 差异性与相关矢量之间的关系

Maprtínez 等<sup>[62]</sup> 从向量空间的角度, 提出相关矢量 (Reference vector), 对差异性度量进行了形象描述. 本文指出: 在基分类器精度为常量时, 基分类器矢量与相关矢量夹角余弦同集成的差异性成正比.

在具体讨论前, 给出如下定义: 基分类器  $h_j$  对样本集  $D$  的识别结果  $\mathbf{R}_j = [O_{1j}, O_{2j}, \dots, O_{Nj}]^T$ , 分类器集合对样本集  $D$  的识别结果  $\mathbf{R} = [\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_L]$ . 若基分类器等权重, 则基分类器集成对样本集  $D$  的识别结果  $\mathbf{R}_{ens} = \frac{1}{L} \sum_{j=1}^L \mathbf{R}_j$ . 集成的相关矢量  $\mathbf{R}_{ref}$  满足式 (20).

$$\mathbf{R}_{ref} = \mathbf{o} - \lambda \mathbf{R}_{ens}, \quad \mathbf{R}_{ref} \perp \mathbf{R}_{ens} \quad (20)$$

式 (20) 中,  $\mathbf{o}$  为  $N \times 1$  的全 1 矢量, 位于特征空间的第一象限;  $\lambda = \langle \mathbf{o}, \mathbf{R}_{ens} \rangle / \langle \mathbf{R}_{ens}, \mathbf{R}_{ens} \rangle$ . 当基分类器集合的平均精度  $P \geq 50\%$  时, 有  $\lambda \geq 0$ .

Martínez 等<sup>[62]</sup> 指出: 式 (20) 中,  $\mathbf{o}$  对应所有样本识别正确的集成, 是集成优化的期望目标, 为集成优化提供优化方向.  $\mathbf{R}_{ref}$  是  $\mathbf{R}_{ens}$  在第一象限内的梯度方向, 沿这个方向运动对  $\mathbf{R}_{ens}$  性能提高的效果较好.

为了更好地理解  $\mathbf{R}_{ref}$  对提高性能的作用, 考虑识别两个样本的情况 (如图 1 所示). 图 1 中, 用矢量表示单个基分类器  $H_i$  的识别性能 (这里只考虑对单个样本识别正确的分类器, 因而只有两种基分类器), 平均投票集成  $\mathbf{R}_{ens}$  的识别性能和对应的相关矢量  $\mathbf{R}_{ref}$ , 并在矢量前端标注. 其中, 对于单个基分类器, 记  $H_1$  为正确识别样本  $x_1$ , 但错误识别样本  $x_2$  的分类器集;  $H_2$  为正确识别样本  $x_2$ , 但错误识别样本  $x_1$  的分类器集. 用点划线表示全 1 矢量  $\mathbf{o}$ , 即对两个样本均分类正确, 也即最终的期望目标.

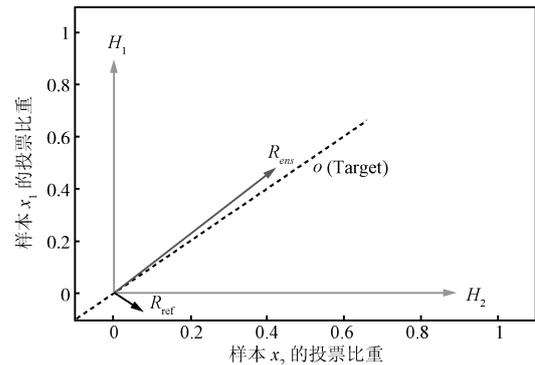


图 1 相关矢量举例

Fig. 1 An example of reference vector

图 1 中,  $\mathbf{R}_{ens}$  在  $\mathbf{o}$  上方, 说明集成对样本  $x_1$  的识别正确, 对样本  $x_2$  识别错误. 此时加大对样本  $x_2$  识别正确的基分类器  $H_2$  在集成中所占的权重, 则  $\mathbf{R}_{ens}$  与  $\mathbf{o}$  的夹角减小, 从而使整体的识别精度提高. Martínez 等<sup>[62]</sup> 指出: 添加沿着  $\mathbf{R}_{ref}$  正方向运动的基分类器, 可以改善系统的集成性能.

本文进一步考虑沿着  $\mathbf{R}_{ref}$  正方向运动的基分类器这一概念. 基分类器  $h_j$  与相关矢量  $\mathbf{R}_{ref}$  的夹角余弦  $\cos(\theta)$  满足式 (21).

$$\cos(\theta_j) = \frac{\langle \mathbf{R}_j, \mathbf{R}_{ref} \rangle}{|\mathbf{R}_j| |\mathbf{R}_{ref}|} = \frac{1}{|\mathbf{R}_j| |\mathbf{R}_{ref}|} (\langle \mathbf{R}_j, \mathbf{o} \rangle - \lambda \langle \mathbf{R}_j, \mathbf{R}_{ens} \rangle) \quad (21)$$

其中, 基分类器  $h_j$  对应矢量  $\mathbf{R}_j$  与  $\mathbf{R}_{ref}$  的夹角余弦  $\cos(\theta_j)$  越大,  $\mathbf{R}_j$  在  $\mathbf{R}_{ref}$  上的投影  $\langle \mathbf{R}_j, \mathbf{R}_{ref} \rangle$  越大, 基分类器  $h_j$  沿  $\mathbf{R}_{ref}$  正方向运动的越多.

在进一步分析  $\langle \mathbf{R}_j, \mathbf{R}_{ens} \rangle$  前, 首先, 考虑两个基分类器  $\{h_j, h_k\}$  对应矢量点积  $\langle \mathbf{R}_j, \mathbf{R}_k \rangle$ , 其满足式

(22).

$$\begin{aligned} \langle \mathbf{R}_j, \mathbf{R}_k \rangle &= \sum_{q=1}^N O_{qj} O_{qk} = \\ &N(a_{jk} + d_{jk} - b_{jk} - c_{jk}) = \\ &N(1 - 2dis_{jk}) \end{aligned} \quad (22)$$

式 (22) 表明, 两个基分类器之间的矢量点积与不一致度量  $dis$  成反比. 将式 (22) 中  $\mathbf{R}_k$  换成  $\mathbf{R}_{ens}$ , 结论依然成立, 即  $\langle \mathbf{R}_j, \mathbf{R}_{ens} \rangle$  同基分类器  $h_j$  与集成之间的一致差异  $dis_j$  成反比. 由此, 基分类器  $h_j$  与相关矢量对应点积  $\langle \mathbf{R}_j, \mathbf{R}_{ref} \rangle$  满足式 (23).

$$\begin{aligned} \langle \mathbf{R}_j, \mathbf{R}_{ref} \rangle &= \langle \mathbf{R}_j, \mathbf{o} \rangle - \lambda \langle \mathbf{R}_j, \mathbf{R}_{ens} \rangle = \\ &\langle \mathbf{R}_j, \mathbf{o} \rangle - \lambda N(1 - 2dis_j) \end{aligned} \quad (23)$$

将式 (23) 带入式 (21), 得到式 (24).

$$\cos(\theta_j) = \frac{1}{|\mathbf{R}_j| |\mathbf{R}_{ref}|} (\langle \mathbf{R}_j, \mathbf{o} \rangle - \lambda N + 2\lambda N dis_j) \quad (24)$$

式 (24) 中,  $1/(|\mathbf{R}_j| |\mathbf{R}_{ref}|)$  和  $N$  均为正的常系数;  $\langle \mathbf{R}_j, \mathbf{o} \rangle$  项与基分类器精度成正比;  $\lambda$  为非负的常系数. 因而, 在基分类器精度为常量时, 基分类器矢量与相关矢量夹角余弦  $\cos(\theta_j)$  同集成的差异性  $dis_j$  成正比.

综上所述, 添加沿着  $\mathbf{R}_{ref}$  方向运动的基分类器, 即与  $\mathbf{R}_{ref}$  夹角余弦  $\cos(\theta_j)$  较大的基分类器  $h_j$ , 可以较好地提高泛化能力. 同时,  $\cos(\theta_j)$  与集成的差异性  $dis_j$  成正比. 因而, 大的差异性意味着好的泛化能力.

## 4 基于差异性度量的集成优化技术

基于差异性与集成泛化能力之间的关系, 实际中多通过最大化差异性来实现优化集成的效果. 目前使用差异性度量优化集成基分类器主要有两种思路: 1) 在基分类器生成的过程中考虑差异性, 即生成方法; 2) 先生成一组基分类器, 再根据差异性从中选择基分类器子集, 即集成修剪 (Ensemble pruning) 方法. 本节对这两种思路加以介绍.

### 4.1 生成方法

生成方法主要有两种思路, 一种是在性能函数中显性地添加代表差异性的正则项; 另一种是通过改变样本权重从而隐性调节整体差异性.

Liu 等<sup>[32]</sup> 提出了负相关学习 (Negatively correlated learning, NCL), 通过在损失函数中加入与差异性有关的正则项, 生成差异性更高的基分类器. 这种显性添加差异性的方法只适用于神经网络等考虑性能函数的方法较为有效, 但是对于决策树等方法则完全无效.

Freund<sup>[18]</sup> 提出 Boosting 方法, 通过更新样本权重的方法来获得更有差异性的基分类器. 这种方法在实验中取得了较好的结果, 是目前主流方法之一.

生成方法的优点在于每次生成的基分类器都使整体差异性趋于更大, 即基分类器集合内的冗余较少. 但是生成方法对于生成基分类器的个数没有限制. 当样本噪声较大时, 生成方法容易产生过拟合现象. 为了解决这些问题, 集成修剪方法.

### 4.2 修剪方法

修剪方法的重要代表为 Zhou 等<sup>[63]</sup> 和张春霞等<sup>[64]</sup> 提出的选择性集成: 通过选择部分基分类器进行集成, 比集成全部基分类器的性能更好. Zhou 等<sup>[63]</sup> 指出: 选择的标准是个体的精度高、差异性大.

目前, 集成修剪的方法主要分为直接搜索方法、排序方法和数学优化方法.

#### 4.2.1 直接搜索方法

直接搜索方法是一种最优化方法. 搜索的目标是差异性大的基分类器子集. 为了提高搜索效率, 常常使用遗传算法进行搜索.

Zhou 等<sup>[63]</sup> 提出 GASEN (Genetic algorithm based selective ensemble) 方法. 该方法使用遗传算法搜索验证集上集成精度最高的分类器子集.

Tuve 等<sup>[65]</sup> 在此基础上, 使用遗传算法对集成精度 (Ensemble accuracy,  $EA$ )、平均精度 (Basic accuracy,  $BA$ )、双错 (Double fault,  $DF$ ) 和难度 ( $\theta$ ) 的任意组合达到最优的子集进行搜索. 其中, 只用  $EA$  进行优化的方法就是 GASEN; 对于多目标优化问题 (如组合 ( $EA, BA$ )), 首先, 用多目标优化遗传算法计算帕累托前沿 (Pareto front), 然后, 将帕累托前沿中最靠近中心位置 (Med) 的点, 所对应的分类器子集作为搜索结果.

#### 4.2.2 排序方法

直接搜索方法由于搜索空间很大, 导致选择的执行时间很长. Margineantu 等<sup>[52]</sup> 提出了排序修剪 (Ordering pruning), 通过限制搜索空间的大小, 减少选择的执行时间.

基于差异性的排序修剪使用上文提及的 5 种成对差异性度量方法, 对整个分类器集合进行排序. 其具体流程如算法 1<sup>[52]</sup>.

前向排序时, 首先, 选择差异性最大的基分类器对; 随后依次选择与已选基分类器集的平均差异性最大的基分类器. 后向排序时, 每次选择与待选基分类器集平均差异性最小的基分类器. 不论前向还是后向方法都存在只能人为设定选择的基分类器个数的缺点. Martinez 等<sup>[66]</sup> 根据二八定律, 推荐选取前 20% ~ 40% 的基分类器. Li 等<sup>[48]</sup> 给出了排序修

剪在 PAC 框架下的解释, 并提出 DREP (Diversity regularized ensemble pruning) 方法.

### 算法 1. 排序裁减方法<sup>[52]</sup>

**Input:**

$D_V$ : 验证集.

$H = \{h_1, h_2, \dots, h_L\}$ : 待选分类器集合,  
 $|H| = L$ .

$D_M$ : 成对差异性度量方法.

**Output:**

$S = \{h_1^*, h_2^*, \dots, h_T^*\}$ : 已选择分类器集合.

**Parameter:**

$T$ : 要选择的分类器子集大小.

**Procedure:**

$h \leftarrow D_V$  上错误率最小的分类器,  $S \leftarrow \{h\}$ ,

$H \leftarrow H \setminus S$ .

**For**  $i = 1, 2, \dots, T$ ;

    计算分类器  $h_j \in H$  与  $S$  中所有分类器的平均  
    差异性:

$$\text{div}_{h_j, S} \leftarrow \frac{1}{|S|} \sum_{h_k \in S} D_M(h_j, h_k)$$

    按照  $\text{div}_{h_i, S}$  对  $H$  中分类器进行降序排序, 得  
    到序列  $R_i$ .

$h \leftarrow$  排在序列  $R_i$  首位的分类器,  $S \leftarrow S \cup \{h\}$ ,

$H \leftarrow H \setminus \{h\}$ .

**End**

### 4.2.3 数学优化方法

直接搜索方法和排序方法目标都在于搜索最优解, 但是在搜索过程中可能会陷入局部最优解而非全局最优解. 为了减少搜索时间, 提高搜索性能, 使用数学优化方法.

目前, 基于差异性的数学优化方法主要分为两种: 二次规划 (Quadratic programming, QP)<sup>[34]</sup> 和半定规划 (Semi-definite programming, SDP)<sup>[33]</sup>.

基于差异性的二次规划方法, 其目标函数如式 (25).

$$\begin{aligned} \mathbf{w}_{\text{opt}} &= \min_{\mathbf{w}} \mathbf{w}^T D \mathbf{w} + \mathbf{A} \mathbf{w} \\ \text{s.t. } \mathbf{1}^T \mathbf{w} &= 1, \quad \mathbf{w} \geq 0 \end{aligned} \quad (25)$$

其中,  $w_j$  为基分类器  $h_j$  的权重.  $D$  与基分类器间差异性成反比,  $\mathbf{A}$  与单个基分类器的精度成反比.

在此基础上, Li 等添加与基分类器权值  $\mathbf{w}$  有关的正则项, 提出了 RSE (Regularized selective ensemble)<sup>[34]</sup> 方法. 其目标函数如式 (26).

$$R = \lambda V(\mathbf{w}) + \Omega(\mathbf{w}) \quad (26)$$

其中,  $V(\mathbf{w})$  为集成在  $D_V$  上的目标函数, 正则项  $\Omega(\mathbf{w})$  用于平滑和简化最终结果.

当使用式 (25) 作为目标函数, Graph Lapla-

caian 作为正则项时, 问题等价于式 (27).

$$\begin{aligned} \mathbf{w}_{\text{opt}} &= \min_{\mathbf{w}} \mathbf{w}^T O^T L O \mathbf{w} + \\ &\quad \lambda \{ \mathbf{w}^T D \mathbf{w} + \mathbf{A} \mathbf{w} \} \\ \text{s.t. } \mathbf{1}^T \mathbf{w} &= 1, \quad \mathbf{w} \geq 0 \end{aligned} \quad (27)$$

其中,  $O \in \{-1, 1\}^{N \times L}$ , 对应分类器集合对样本集 的识别结果. 若记数据集  $D$  的邻接图为  $G = (V; E)$ , 其邻接矩阵为  $T$ ; 又记对角矩阵  $S$ , 满足  $S_{ii} = \sum_{j=1}^L T_{ij}$ . 则式 (27) 中,  $L = S^{-1/2}(T - S)S^{-1/2}$ .

QP 方法的优点在于其约束条件是  $l_1$  范数.  $l_1$  是一种稀疏约束, 因而 QP 的结果更倾向于选择更小的基分类器子集. 此外, 由于 RSE 中使用 Graph Laplacian 正则项, 因而也可用于半监督学习中.

QP 方法适用于加权集成. 当选用投票集成时, 根据给定阈值和 QP 生成的基分类器权重大小来选择基分类器子集, 性能较差. Zhang 等<sup>[33]</sup> 将问题视为 0-1 规划, 并用 SDP 方法加以解决.

Zhang 给出平均差异性最大的目标函数如式 (25).

$$\begin{aligned} \mathbf{z}_{\text{opt}} &= \min_{\mathbf{z}} \mathbf{z}^T G \mathbf{z} \\ \text{s.t. } \mathbf{1}^T \mathbf{z} &= T, \quad \mathbf{z} \in \{0, 1\}^{L \times 1} \end{aligned} \quad (28)$$

其中,  $G$  为基分类器间的成对差异性,  $G_{ij} = \text{div}(h_i, h_j)$ ;  $L$  为待选基分类器个数,  $T$  为被选中的基分类器个数. 对于式 (28) 的 0-1 规划问题, 定义  $L \times 1$  矢量  $\mathbf{v}$ , 其元素  $v_i = 2z_i - 1 \in \{-1, 1\}$ . 记单位矩阵  $I$ , 并定义  $V, H, D$  如式 (29).

$$\begin{aligned} V &= \mathbf{v} \mathbf{v}^T \\ H &= \begin{pmatrix} \mathbf{1}^T G \mathbf{1} & \mathbf{1}^T G \\ G \mathbf{1} & G \end{pmatrix} \\ D &= \begin{pmatrix} L & \mathbf{1}^T \\ \mathbf{1} & I \end{pmatrix} \end{aligned} \quad (29)$$

得到式 (30).

$$\begin{aligned} V_{\text{opt}} &= \min_V H \otimes V \\ \text{s.t. } D \otimes V &= 4T \\ \text{diag}\{V\} &= \mathbf{1}, \quad V \geq 0 \end{aligned} \quad (30)$$

其中,  $H \otimes V = \sum_{i,j} H_{ij} V_{ij}$ .

与 QP 方法相比, SDP 方法在解决投票问题的基分类器选择时, 性能更好.

## 5 分类器优化集成实验与对比分析

上述的多种集成优化技术各有利弊, 对于选择哪一种方法仍没有统一的标准. 本节对集成优化技

术进行对比实验,分析在不同基分类器集大小情况下,各种技术的优劣.其中,第5.1节重点对集成学习中各种优化方法进行实验和对比分析;第5.2节对传统多分类器系统(分类器个数较少)中各种优化方法进行实验和对比分析.

## 5.1 集成优化实验

### 5.1.1 实验方法与实验数据

目前基于差异性的集成优化技术主要分为生成方法和修剪方法两类.与生成方法相比,修剪方法不关心基分类器的生成过程和集成方式,因而易于推广,在理论研究和实际应用中使用较多.同时,当基分类器存在大量冗余时,使用修剪方法得到的基分类器子集具有较好的泛化能力.本节实验中,重点比较多种修剪方法.

实验时,从UCI数据库<sup>[67]</sup>选取10个数据集(数据库信息见表3),以CART(Classification and regression tree)树作为基分类器,使用Bagging分别生成25, 51, 75, 101, 151个基分类器.使用10-cross-cv,对比考虑双错度量( $df$ )和成对Kappa( $\kappa p$ )两种差异性时, Bagging、Gasen、AdaBoost、GA、前向排序(Forward ordering, FO)、后向排序(Backward ordering, BO)、QP、RSE、SDP等9种优化方法的性能.具体测试方法如表4所示.

表3 数据库信息  
Table 3 Information for datasets

数据集	样本个数	测试方法	特征维数	类别
German	1 000	10-fold-cv	24	2
Imageseg	2 310	10-fold-cv	19	7
Wave40	5 000	10-fold-cv	40	3
Wave21	5 000	10-fold-cv	21	3
Chess	3 196	10-fold-cv	36	2
Sick	3 773	10-fold-cv	29	2
Sick-euthyroid	3 263	10-fold-cv	24	2
Allbp	3 773	10-fold-cv	29	3
Led7	3 200	10-fold-cv	7	10
Led24	3 200	10-fold-cv	24	10

表4中,1) Bagging、Gasen和Adaboost方法作为参考基准;QP、RSE和Adaboost使用加权投票集成,其他方法均使用投票集成;2) GA方法优化目标为平均精度和平均成对差异性的线性组合;3) 前向排序,后向排序取前21%个分类器;4) SDP方法取 $T$ 为基分类器集合大小的21%.

表4 测试集成方法  
Table 4 Testing methods

方法名称	出处	备注
Bagging	Breiman(1996)	
Gasen	Zhou(2002)	适应度函数为集成精度倒数
Adaboost	Freund(1995)	
GA- $\kappa p$	Trawinski(2009)	适应度函数为平均精度与成对Kappa度量的线性组合
GA- $df$	Trawinski(2009)	适应度函数为平均精度与DF度量的线性组合
FO- $\kappa p$	Margineantu(1997)	按照成对Kappa进行前向排序
FO- $df$	Margineantu(1997)	按照DF进行前向排序
BO- $\kappa p$	Margineantu(1997)	按照成对Kappa进行后向排序
BO- $df$	Margineantu(1997)	按照DF进行后向排序
QP- $\kappa p$	Li(2012)	按照成对Kappa进行二次规划
QP- $df$	Li(2012)	按照DF进行二次规划
RSE- $\kappa p$	Li(2012)	按照成对Kappa进行数学优化
RSE- $df$	Li(2012)	按照DF进行数学优化
SDP- $\kappa p$	Zhang(2006)	按照成对Kappa进行半定规划
SDP- $df$	Zhang(2006)	按照DF进行半定规划

### 5.1.2 实验结果与对比分析

集成101个基分类器时,各种方法集成识别率如表5,修剪后分类器子集大小如表6.其中每个数据集上识别精度与最大值之差不大于0.5%的被标粗.表5中,最后一行列出了各种方法平均排序(Ranks).对于某个测试数据集,性能最好的排序为1,性能次好的排序为2,以此类推.表6中,最后一行列出了各种方法所用基分类器的平均个数(Average).

通过分析表5和表6,有:

1) 与成对Kappa( $\kappa p$ )方法相比,使用双错度量( $df$ )进行差异性度量时,得到的集成优化效果较好.表5中,对除GA-based和QP方法外的大多数方法,使用 $df$ 作为差异性指标在性能上要优于 $\kappa p$ .在整体性能较好的数学优化方法中, $df$ 作为优化目标在性能上要优于 $\kappa p$ 方法.

2) 多种优化方法中,数学优化的方法要明显好于直接搜索和排序的方法.表6中,使用数学优化的方法(QP和RSE)获得的分类器个数要多于直接搜索和排序的方法.多种数学优化方法相比,SDP方法执行时间较长,且需要预先设定分类器子集的分类器个数.

3) 使用 $df$ 进行差异性度量时,RSE(Ranks = 4.1)和SDP(Ranks = 3.1)识别精度与Adaboost(Ranks = 3.7)方法相近,且明显优于其他方法.

为了进一步分析不同集成优化方法的性能.在Imageseg、Wave40、Wave21、Chess、Sick-euthyroid和Allbp等5个UCI数据库上考虑使用 $df$ 进行差异性度量时,统计分类器集合大小对集成精度的影响.实验结果如图2所示.

表 5 集成 101 个基分类器时各种方法集成识别率 (均值  $\pm$  标准差) (%)  
 Table 5 Recognition rate (Average  $\pm$  Standard deviation) for ensemble with 101 base classifiers (%)

DataSet	Bag	Gasen	AB	GA		FO		BO		QP		RSE		SDP	
				$\kappa p$	$df$										
German	76.58	77.81	75.14	78.02	<b>78.20</b>	77.01	76.30	75.42	77.43	77.85	76.46	77.42	<b>78.60</b>	77.00	<b>78.30</b>
	$\pm 4.07$	$\pm 5.97$	$\pm 4.75$	$\pm 6.24$	$\pm 5.15$	$\pm 4.85$	$\pm 7.08$	$\pm 0.90$	$\pm 3.06$	$\pm 6.00$	$\pm 4.75$	$\pm 5.09$	$\pm 4.51$	$\pm 0.92$	$\pm 3.13$
Imageseg	97.50	97.60	<b>99.00</b>	97.38	97.33	97.72	97.77	97.73	97.77	97.51	97.72	97.51	97.70	97.51	97.70
	$\pm 0.82$	$\pm 1.45$	$\pm 0.40$	$\pm 0.80$	$\pm 0.72$	$\pm 0.58$	$\pm 1.01$	$\pm 0.40$	$\pm 3.63$	$\pm 1.07$	$\pm 0.65$	$\pm 0.63$	$\pm 0.60$	$\pm 0.71$	$\pm 3.98$
Wave40	83.87	83.09	<b>84.60</b>	83.84	83.89	82.90	83.01	82.81	82.68	84.03	83.57	84.00	83.90	83.80	<b>84.10</b>
	$\pm 1.05$	$\pm 1.67$	$\pm 0.09$	$\pm 1.05$	$\pm 1.06$	$\pm 1.07$	$\pm 1.46$	$\pm 0.5$	$\pm 3.03$	$\pm 1.02$	$\pm 0.94$	$\pm 1.26$	$\pm 1.11$	$\pm 0.79$	$\pm 4.22$
Wave21	83.98	84.00	<b>84.80</b>	83.86	83.57	83.20	83.14	82.86	83.19	84.24	83.91	<b>84.50</b>	<b>84.50</b>	<b>84.60</b>	<b>84.40</b>
	$\pm 1.23$	$\pm 2.37$	$\pm 2.20$	$\pm 0.96$	$\pm 1.33$	$\pm 1.12$	$\pm 0.99$	$\pm 1.10$	$\pm 1.00$	$\pm 1.14$	$\pm 0.86$	$\pm 1.22$	$\pm 1.36$	$\pm 1.41$	$\pm 1.34$
Chess	<b>99.60</b>	<b>99.59</b>	<b>99.65</b>	<b>99.54</b>	<b>99.57</b>	<b>99.54</b>	<b>99.54</b>	<b>99.47</b>	<b>99.47</b>	<b>99.53</b>	99.51	<b>99.60</b>	<b>99.60</b>	<b>99.70</b>	<b>99.7</b>
	$\pm 0.33$	$\pm 0.35$	$\pm 0.43$	$\pm 0.37$	$\pm 0.34$	$\pm 0.36$	$\pm 0.31$	$\pm 0.50$	$\pm 2.13$	$\pm 0.36$	$\pm 0.34$	$\pm 0.36$	$\pm 0.43$	$\pm 0.67$	$\pm 4.06$
Sick	<b>98.83</b>	<b>99.01</b>	<b>99.07</b>	<b>98.89</b>	<b>98.81</b>	<b>99.02</b>	<b>99.05</b>	<b>99.02</b>	<b>99.02</b>	<b>98.75</b>	<b>99.02</b>	<b>99.02</b>	<b>99.02</b>	<b>99.02</b>	<b>99.12</b>
	$\pm 0.54$	$\pm 0.74$	$\pm 0.51$	$\pm 0.54$	$\pm 0.67$	$\pm 0.72$	$\pm 0.57$	$\pm 0.60$	$\pm 3.64$	$\pm 0.7$	$\pm 0.65$	$\pm 0.72$	$\pm 0.57$	$\pm 0.63$	$\pm 0.40$
Sick-euthyroid	<b>97.81</b>	<b>97.85</b>	<b>97.92</b>	<b>97.75</b>	<b>97.78</b>	<b>97.62</b>	<b>97.59</b>	<b>97.59</b>	<b>97.59</b>	<b>97.64</b>	<b>97.59</b>	<b>97.85</b>	<b>97.81</b>	<b>97.90</b>	<b>97.80</b>
	$\pm 1.10$	$\pm 1.15$	$\pm 1.18$	$\pm 1.12$	$\pm 1.07$	$\pm 1.11$	$\pm 1.19$	$\pm 1.23$	$\pm 2.47$	$\pm 1.15$	$\pm 1.03$	$\pm 1.11$	$\pm 1.01$	$\pm 1.38$	$\pm 4.04$
Allbp	97.59	97.64	<b>98.30</b>	97.70	97.57	97.65	97.65	97.51	97.67	97.62	97.67	97.70	97.70	<b>98.00</b>	<b>98.04</b>
	$\pm 0.93$	$\pm 1.12$	$\pm 0.30$	$\pm 0.85$	$\pm 0.89$	$\pm 1.01$	$\pm 1.09$	$\pm 1.02$	$\pm 2.07$	$\pm 0.90$	$\pm 0.98$	$\pm 0.85$	$\pm 0.62$	$\pm 1.96$	$\pm 2.65$
Led7	72.45	72.90	73.00	72.95	72.52	72.86	72.82	73.42	73.42	<b>74.44</b>	72.87	73.42	<b>74.00</b>	73.20	<b>74.00</b>
	$\pm 2.72$	$\pm 2.48$	$\pm 1$	$\pm 3.01$	$\pm 3$	$\pm 2.53$	$\pm 2.86$	$\pm 3.39$	$\pm 2.16$	$\pm 3.03$	$\pm 2.76$	$\pm 2.61$	$\pm 3.15$	$\pm 5.91$	$\pm 2.82$
Led24	<b>71.79</b>	71.50	<b>72.00</b>	<b>71.85</b>	<b>72.17</b>	71.36	71.64	71.02	71.23	<b>71.91</b>	<b>71.61</b>	<b>71.88</b>	<b>71.90</b>	<b>72.02</b>	<b>72.04</b>
	$\pm 1.82$	$\pm 4.89$	$\pm 0.10$	$\pm 2.03$	$\pm 1.60$	$\pm 1.81$	$\pm 1.68$	$\pm 0.70$	$\pm 2.83$	$\pm 1.78$	$\pm 1.90$	$\pm 1.51$	$\pm 1.73$	$\pm 0.96$	$\pm 3.73$
Ranks	9.8	8.8	3.7	8.7	9.0	9.8	9.8	11.1	9.2	7.9	9.5	5.1	4.1	5.1	3.1

表 6 集成 101 个基分类器时各种方法分类器集合大小 (均值  $\pm$  标准差)  
 Table 6 Pruned set size (Average  $\pm$  standard deviation) for ensemble with 101 base classifiers

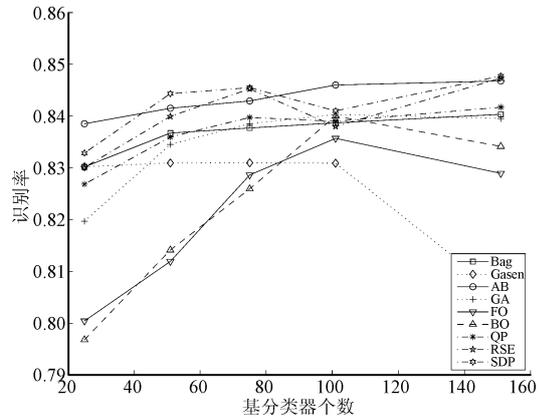
DataSet	Bag	Gasen	AB	GA		FO		BO		QP		RSE		SDP	
				$\kappa p$	$df$	$\kappa p$	$df$	$\kappa p$	$df$	$\kappa p$	$df$	$\kappa p$	$df$	$\kappa p$	$df$
German	101.00	38.90	101.00	79.40	77.60	21.00	21.00	21.00	21.00	37.30	46.20	45.97	53.36	21.00	21.00
		$\pm 13.78$		$\pm 2.27$	$\pm 4.22$					$\pm 2.31$	$\pm 3.74$	$\pm 6.89$	$\pm 6.64$		
Imageseg	101.00	21.50	101.00	78.10	78.80	21.00	21.00	21.00	21.00	22.30	29.20	19.84	21.88	21.00	21.00
		$\pm 21.14$		$\pm 3.41$	$\pm 3.29$					$\pm 2.54$	$\pm 3.43$	$\pm 5.56$	$\pm 5.20$		
Wave40	101.00	31.10	101.00	77.00	78.10	21.00	21.00	21.00	21.00	56.40	76.40	45.54	61.51	21.00	21.00
		$\pm 17.92$		$\pm 2.72$	$\pm 4.01$					$\pm 2.37$	$\pm 3.20$	$\pm 6.70$	$\pm 5.25$		
Wave21	101.00	14.40	101.00	79.00	75.70	21.00	21.00	21.00	21.00	60.30	76.30	48.69	63.49	21.00	21.00
		$\pm 11.99$		$\pm 3.89$	$\pm 3.06$					$\pm 3.20$	$\pm 3.47$	$\pm 7.30$	$\pm 6.96$		
Chess	101.00	32.70	101.00	76.70	76.30	21.00	21.00	21.00	21.00	11.90	15.40	11.84	16.02	21.00	21.00
		$\pm 21.45$		$\pm 3.62$	$\pm 3.59$					$\pm 1.85$	$\pm 2.5$	$\pm 3.92$	$\pm 4.59$		
Sick	101.00	28.00	101.00	77.8	76.9	21.00	21.00	21.00	21.00	16.3	22.00	16.34	19.82	21.00	21.00
		$\pm 15.84$		$\pm 3.68$	$\pm 3.18$					$\pm 2.58$	$\pm 2.26$	$\pm 6.87$	$\pm 4.07$		
Sick-euthyroid	101.00	18.50	101.00	78.50	76.50	21.00	21.00	21.00	21.00	21.20	27.30	21.19	29.74	21.00	21.00
		$\pm 12.97$		$\pm 3.63$	$\pm 2.68$					$\pm 3.55$	$\pm 3.34$	$\pm 6.63$	$\pm 6.12$		
Allbp	101.00	25.90	101.00	78.5	79.1	21.00	21.00	21.00	21.00	23.60	30.20	21.05	28.15	21.00	21.00
		$\pm 21.98$		$\pm 2.27$	$\pm 4.28$					$\pm 2.32$	$\pm 2.35$	$\pm 5.55$	$\pm 3.74$		
Led7	101.00	39.70	101.00	78.10	67.50	21.00	21.00	21.00	21.00	17.00	25.20	21.87	29.07	21.00	21.00
		$\pm 21.30$		$\pm 2.92$	$\pm 22.74$					$\pm 2.16$	$\pm 2.74$	$\pm 7.77$	$\pm 4.82$		
Led24	101.00	39.10	101.00	78.50	77.80	21.00	21.00	21.00	21.00	47.60	58.20	53.98	66.38	21.00	21.00
		$\pm 17.48$		$\pm 3.89$	$\pm 4.02$					$\pm 3.53$	$\pm 3.52$	$\pm 7.34$	$\pm 6.10$		
Average	101.00	28.98	101.00	78.22	76.43	21.00	21.00	21.00	21.00	31.39	40.64	30.53	38.94	21.00	21.00

图 2 中, 当基分类器个数在 80~120 之间时, RSE 和 SDP 两种方法的性能较为稳定. 因而, 前文根据 101 个基分类器集成指出 RSE 和 SDP 方法具有较好的性能是有意义的.

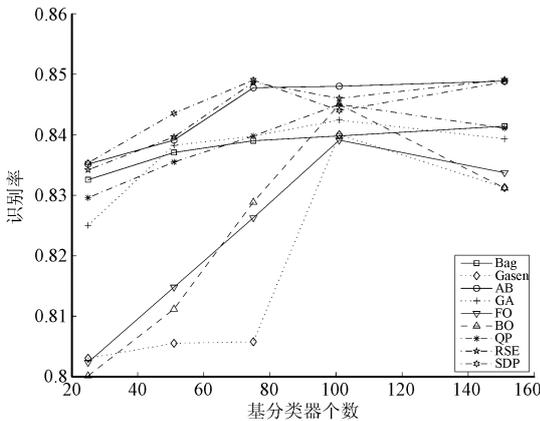
### 5.2 多分类器集成优化试验

上节主要讨论了在集成学习 (Bagging, Boosting) 中多分类器的选择问题.

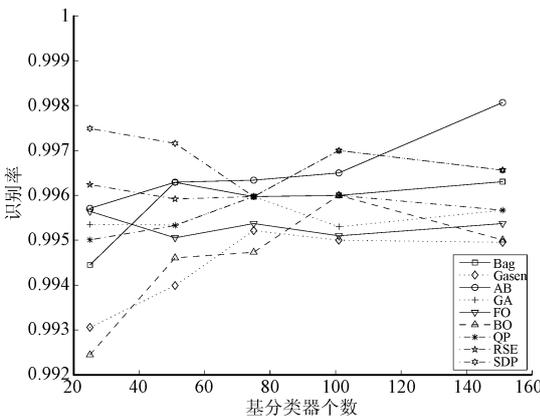
在传统多分类器学习系统中, 往往使用各种方法生成少量基分类器, 再使用集成的方法对结果进



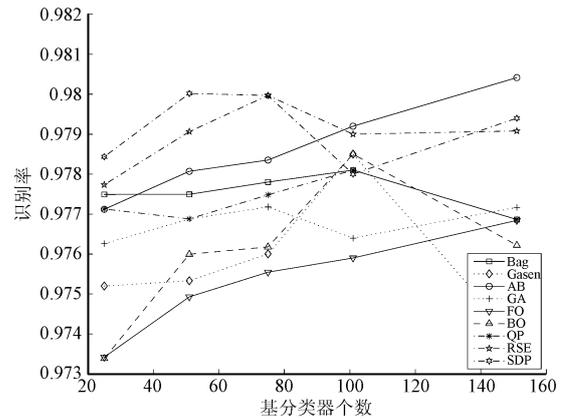
(a) Wave40



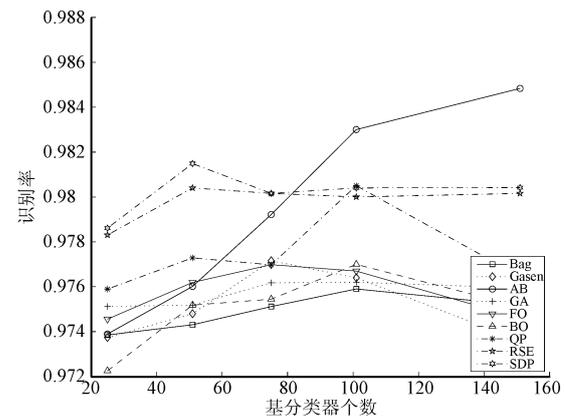
(b) Wave21



(c) Chess



(d) Sick-euthyroid



(e) Allbp

图 2 5 组 UCI 数据库上集成不同分类器个数对性能的影响  
Fig. 2 Influences of ensembles different sizes of classifiers on 5 UCI databases

行综合. 为了进一步讨论此时表 4 中修剪方法的性能, 使用手写数字样本库 USPS<sup>[68]</sup> 进行实验. USPS 库中每个样本用 16 × 16 的灰度图像的值表示, 灰度值已经被归一化. 库中共有 9 298 个手写数字图像, 其中, 7 291 个用于训练, 2 007 个用于测试. 实验中, 使用不同特征选择和机器学习算法分别对 USPS 数据库进行学习, 对结果使用投票集成, 并使用各种修剪方法对集成进行优化.

由于实验旨在分析集成少量基分类器时, 所以实验时选用的特征提取方法和学习算法都较为简单. 实验中, 共采用三种特征提取方法: 原始特征、PCA (维数为 64 和 128 维); 并使用 4 种机器学习算法: 支持向量机 (Support vector machine, SVM)、决策树 (Decision tree, DT)、K-近邻 (K-nearest neighbor, KNN) 和神经网络 (Neural network, NN) 进行学习. 各种特征提取-学习方法组合在测试集上的错误率如表 7 所示.

集成上述 12 个基分类器时, 各种方法集成识别率如表 8 所示. 表 8 中, 排序方法和 SDP 方法均取 6 个分类器.

表 7 各种特征提取 (FE)-学习方法 (LM) 组合的识别率 (%)

Table 7 Recognition rate (%) for different feature extraction (FE)-learning method (LM) combination (%)

FE \ LM	SVM	DT	KNN	NN
Original	92.33	82.02	93.37	92.18
PCA-64	90.94	81.22	90.12	90.62
PCA-128	92.03	81.37	92.2	91.12

表 8 各种集成识别率 (%)

Table 8 Recognition rate (%) for ensemble

方法名称	识别率	方法名称	识别率
Voting	94.18	Gasen	<b>95.47</b>
GA- $\kappa p$	<b>95.45</b>	GA- $df$	<b>95.49</b>
FO- $\kappa p$	<b>95.50</b>	FO- $df$	<b>95.60</b>
BO- $\kappa p$	93.87	BO- $df$	94.30
QP- $\kappa p$	93.50	QP- $df$	93.86
RSE- $\kappa p$	94.12	RSE- $df$	94.20
SDP- $\kappa p$	<b>95.49</b>	SDP- $df$	<b>95.51</b>

表 8 可以看出, 当集成少数基分类器时, 使用遗传算法搜索 (Gasen、GA- $\kappa p$ 、GA- $df$ ) 和排序方法进行优化时, 可以使集成后识别率提高 1% 左右; 使用二次规划 (QP- $\kappa p$ 、QP- $df$ 、RSE- $\kappa p$ 、RSE- $df$ ) 方法进行优化容易产生过学习现象, 导致识别率下降; 使用半定规划 (SDP- $\kappa p$ 、SDP- $df$ ) 进行优化性能较为稳定。

### 5.3 研究与应用指导

通过前面的方法综述及对比实验, 在基于差异性的分类器集成研究与应用中, 相关的差异性度量方法选择和具体的优化集成技术设计的主要指导与建议包括:

1) 在多种差异性度量方法中, 双错度量 ( $df$ ) 能够较好地代表泛化能力. 对于直接搜索方法 (GA-based) 和排序方法 (Forward ordering, FO; Backward ordering, BO), 使用  $df$  和  $\kappa p$  进行差异性度量, 对性能影响不大; 而使用数学优化方法时, 使用  $df$  度量差异性在性能上要明显好于  $\kappa p$  方法。

2) 数学优化方法在识别精度上要优于直接搜索和排序方法. 使用  $df$  进行差异性度量时, RSE- $df$  和 SDP- $df$  在性能上要优于 GA- $df$ 、FO- $df$  和 BO- $df$ . SDP- $df$  在性能上与 Adaboost 和 RSE- $df$  方法相近, 但是 SDP- $df$  需要预先设定分类器子集的分类器个数。

3) 进行数学优化时, 适宜使用 80~120 个待选基分类器. 图 2 中, 当分类器个数较少时, 使用数学优化方法容易产生过学习现象, 此时性能较不稳定;

当分类器个数较多时, 更多的分类器加大了选择方法的执行时间, 同时对提高性能的作用较少。

4) 当集成数目较少的强分类器时, 推荐使用遗传算法搜索、排序方法和半定规划可以使识别性能提高。

综上, 当使用差异性对数目较多的分类器学习系统进行集成优化时, 推荐使用  $df$  进行差异性度量, 并使用 RSE 或 SDP 方法进行分类器选择. 使用 RSE 和 SDP 方法时, 适宜选用 80~120 个基分类器进行选择. 使用 SDP 方法时,  $T$  推荐为 21; 当对数目较少的基分类器进行集成时, 推荐使用遗传算法搜索、排序方法和半定规划方法。

## 6 结论

差异性获得高性能集成的重要指标. 本文从差异性的度量方法、差异性的有效性分析和基于差异性的优化集成等三个重要方面, 总结了前人的研究成果, 对基于差异性的集成进行了阐述. 本文首先根据相关向量关系, 直观形象地论证了差异性度量的有效性. 同时, 通过对比实验与性能分析, 本文推荐使用双错度量进行差异性度量, 并在 80~120 的待选基分类器集上使用 RSE 和 SDP 方法进行集成优化。

目前, 对差异性的度量和操作仍然没有统一的标准. 为了更好地利用差异性来优化集成, 建议从以下两个方面进行基于差异性的分类器集成的研究:

1) 设计更能体现集成泛化能力的差异性. 虽然, 目前存在一些差异性 (如 Kappa、双错度量) 度量方法, 它们在集成优化时取得了较好的结果. 但是, 不论哪种差异性方法, 其在一定程度上都忽略了平均精度. 因而, 需要研究一种新的差异性度量方法, 其与平均精度具有一定的关联性, 并与集成泛化能力具有更高的相关性。

2) 构造同时考虑差异性和平均精度的集成优化模型. 本文第 3.2 节提及, 最大化差异性并不一定意味着泛化能力最好. 在平均误差 (或精度) 和差异性之间, 存在一个平衡 (Trade-off) 状态, 使得泛化能力较好. 然而, 现有的基于差异性的集成优化模型, 往往只考虑了差异性. 此外, 考虑差异性和平均精度的几种尝试虽然有一定效果, 但是缺乏理论依据. 因而, 需要研究一种集成优化模型, 其能够同时考虑差异性和平均精度, 从而更有效地进行分类器集成。

## References

- Polikar R. Ensemble learning. *Ensemble Machine Learning: Methods and Applications*. New York: Springer, 2012. 1-34
- Zhou Z H. *Ensemble Methods: Foundations and Algorithms*. New York: CRC Press, 2012

- 3 Lebanon G, Lafferty J. Boosting and maximum likelihood for exponential models. *Advances in Neural Information Processing Systems 14*. Cambridge: MIT Press, 2002. 447–454
- 4 Lee H, Kim E, Pedrycz W. A new selective neural network ensemble with negative correlation. *Applied Intelligence*, 2012, **37**(4): 488–498
- 5 Liu C L. Classifier combination based on confidence transformation. *Pattern Recognition*, 2005, **38**(1): 11–28
- 6 Shipp C A, Kuncheva L K. Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 2002, **3**(2): 135–148
- 7 Jiang L X, Cai Z H, Zhang H, Wang D H. Naive Bayes text classifiers: a locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence*, 2013, **25**(2): 273–286
- 8 Yuksel S E, Wilson J N, Gader P D. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(8): 1177–1193
- 9 Shi L, Wang Q, Ma X M, Weng M, Qiao H B. Spam email classification using decision tree ensemble. *Journal of Computational Information Systems*, 2012, **8**(3): 949–956
- 10 Malisiewicz T, Gupta A, Efros A A. Ensemble of exemplar-SVMs for object detection and beyond. In: Proceedings of the 13th International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 89–96
- 11 Zhou Jin-Zhu, Huang Jin. Multiple kernel linear programming support vector regression incorporating prior knowledge. *Acta Automatica Sinica*, 2011, **37**(3): 360–370  
(周金柱, 黄进. 集成先验知识的多核线性规划支持向量回归. 自动化学报, 2011, **37**(3): 360–370)
- 12 Nguyen H L, Woon Y K, Ng W K, Wan L. Heterogeneous ensemble for feature drifts in data streams. In: Proceedings of the 16th Pacific-Asia Conference of Advances in Knowledge Discovery and Data Mining. Kuala Lumpur, Malaysia: Springer, 2012. 1–12
- 13 Tahir M A, Kittlera J, Bouridaneb A. Multilabel classification using heterogeneous ensemble of multi-label classifiers. *Pattern Recognition Letters*, 2012, **33**(5): 513–523
- 14 Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 2007, **22**(4): 477–505
- 15 Mease D, Wyner A. Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research*, 2008, **9**: 131–156
- 16 Shen C H, Li H X. On the dual formulation of boosting algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(12): 2216–2231
- 17 Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123–140
- 18 Freund Y. Boosting a weak learning algorithm by majority. *Information and Computation*, 1995, **121**(2): 256–285
- 19 Leistner C, Saffari A, Roth P M, Bischof H. On robustness of on-line boosting—a competitive study. In: Proceedings of the 12th International Conference on Computer Vision Workshops. Kyoto, Japan: IEEE, 2009. 1362–1369
- 20 Wolpert D H. Stacked generalization. *Neural Networks*, 1992, **5**(2): 241–260
- 21 Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5–32
- 22 Jain A K, Duin R P W, Mao J C. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(1): 4–37
- 23 Yu Ling, Wu Tie-Jun. LS-Ensem: a ensemble method for regression. *Chinese Journal of Computers*, 2006, **29**(5): 719–726  
(于玲, 吴铁军. LS-Ensem: 一种用于回归的集成算法. 计算机学报, 2006, **29**(5): 719–726)
- 24 Zhang Yu, Zhou Zhi-Hua. A new age estimation method based on ensemble learning. *Acta Automatica Sinica*, 2008, **34**(8): 997–1000  
(张宇, 周志华. 基于集成的年龄估计方法. 自动化学报, 2008, **34**(8): 997–1000)
- 25 Liu Ming, Yuan Bao-Zong, Miao Zhen-Jiang. A double-objective rank level classifier fusion method. *Acta Automatica Sinica*, 2007, **33**(12): 1276–1282  
(刘明, 袁保宗, 苗振江. 一种双目标排序层分类器融合方法. 自动化学报, 2007, **33**(12): 1276–1282)
- 26 Jiang Li-Xing, Hou Jin. Image annotation using the ensemble learning. *Acta Automatica Sinica*, 2012, **38**(8): 1257–1262  
(蒋黎星, 侯进. 基于集成分类算法的自动图像标注. 自动化学报, 2012, **38**(8): 1257–1262)
- 27 Zhang Liang, Huang Shu-Guang, Hu Rong-Gui. Ensemble system of double granularity RNN by linear combination. *Acta Automatica Sinica*, 2011, **37**(11): 1402–1406  
(张亮, 黄曙光, 胡荣贵. 线性合成的双粒度 RNN 集成系统. 自动化学报, 2011, **37**(11): 1402–1406)
- 28 Yang Bo, Liu Jie, Liu Da-You. A random network ensemble model based generalized network community mining algorithm. *Acta Automatica Sinica*, 2012, **38**(5): 812–822  
(杨博, 刘杰, 刘大有. 基于随机网络集成模型的广义网络社区挖掘算法. 自动化学报, 2012, **38**(5): 812–822)
- 29 Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 2000, **40**(2): 139–158
- 30 Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 2003, **51**(2): 181–207
- 31 Schapire R E, Freund Y, Bartlett P L, Lee W S. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998, **26**(5): 1651–1686

- 32 Liu Y, Yao X. Ensemble learning via negative correlation. *Neural Networks*, 1999, **12**(10): 1399–1404
- 33 Zhang Y, Burer S, Street W N. Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, 2006, **7**: 1315–1338
- 34 Li N, Zhou Z H. Selective ensemble under regularization framework. In: Proceedings of the 8th International Workshop on Multiple Classifier Systems. Reykjavik, Iceland: Springer, 2009. 293–303
- 35 Dietterich T G. Machine learning research: four current directions. *AI Magazine*, 1997, **18**(4): 97–136
- 36 Skalak D B. The sources of increased accuracy for two proposed boosting algorithms. In: Proceedings of the 13th American Association for Artificial Intelligence, Integrating Multiple Learned Models Workshop. Portland, Oregon: AAAI Press, 1996. 120–125
- 37 Giacinto G, Roli F. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 2000, **19**: 699–707
- 38 Kohavi R, Wolpert D H. Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the 13th International Conference on Machine Learning. Bari, Italy: Springer, 1996. 275–283
- 39 Sim J, Wright C C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 2005, **85**(3): 257–268
- 40 Yule G U. On the association of attributes in statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 1900, **194**: 257–319
- 41 Hansen L K, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, **12**(10): 993–1001
- 42 Cunningham P, Carney J. Diversity versus quality in classification ensembles based on feature selection. Technical Report TCD-CS-2000-02, Department of Computer Science, Trinity College Dublin, Ireland, 2000
- 43 Partridge D, Krzanowski W J. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 1997, **39**(10): 707–717
- 44 Tumber K, Ghosh J. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 1996, **29**(2): 341–348
- 45 Tang E K, Suganthan P N, Yao X. An analysis of diversity measures. *Machine Learning*, 2006, **65**(1): 247–271
- 46 Zhou Z H, Yu Y. Ensembling local learners through multimodal perturbation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(4): 725–735
- 47 Yu Y, Li Y F, Zhou Z H. Diversity regularized machine. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Catalonia, Spain: Morgan Kaufmann, 2011. 1603–1608
- 48 Li N, Yu Y, Zhou Z H. Diversity regularized ensemble pruning. In: Proceedings of the 23rd European Conference on Machine Learning. Bristol, UK: Springer, 2012. 330–345
- 49 Jing Xiao-Yuan, Yang Jing-Yu. Combining classifiers based on analysis of correlation and effective supplement. *Acta Automatica Sinica*, 2000, **26**(6): 741–747  
(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. *自动化学报*, 2000, **26**(6): 741–747)
- 50 Hao Hong-Wei, Wang Zhi-Bin, Yin Xu-Cheng, Chen Zhi-Qiang. Dynamic selection and circulating combination for multiple classifier systems. *Acta Automatica Sinica*, 2011, **37**(11): 1290–1295  
(郝红卫, 王志彬, 殷绪成, 陈志强. 分类器的动态选择与循环集成方法. *自动化学报*, 2011, **37**(11): 1290–1295)
- 51 Trawinski K, Quirin A, Cordon O. On the combination of accuracy and diversity measures for genetic selection of bagging fuzzy rule-based multiclassification systems. In: Proceedings of the 9th International Conference on Intelligent Systems Design and Applications. Pisa, Italy: IEEE, 2009. 121–127
- 52 Margineantu D D, Dietterich T G. Pruning adaptive boosting. In: Proceedings of the 14th International Conference on Machine Learning. Nashville, Tennessee, USA: Morgan Kaufmann, 1997. 211–218
- 53 Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. *Neural Information Processing Systems*, 1995, **7**: 231–238
- 54 Yin X C, Huang K Z, Hao H W, Iqbal K, Wang Z B. Classifier ensemble using a heuristic learning with sparsity and diversity. In: Proceedings of the 19th International Conference on Neural Information Processing. Doha, Qatar: Springer, 2012. 100–107
- 55 Abbass H A. Pareto neuro-evolution: constructing ensemble of neural networks using multi-objective optimization. In: Proceedings of the 2003 IEEE Conference on Evolutionary Computation. Canberra, Australia: IEEE, 2003. 2074–2080
- 56 Abbass H A. Pareto neuro-ensembles. In: Proceedings of the 16th Australian Joint Conference on Artificial Intelligence. Perth, Australia: Springer, 2003. 554–566
- 57 Chandra A, Yao X. DIVACE: diverse and accurate ensemble learning algorithm. *Computer Science*, 2004, **3177**: 619–625
- 58 Rätsch G, Onoda T, Müller K R. Soft margins for AdaBoost. *Machine Learning*, 2001, **42**(3): 287–320
- 59 Vapnik V N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995
- 60 Wang L W, Sugiyama M, Jing Z X, Yang C, Zhou Z H, Feng J F. A refined margin analysis for boosting algorithms via equilibrium margin. *Journal of Machine Learning Research*, 2011, **12**: 1835–1863

- 61 Gao W, Zhou Z H. On the Doubt About Margin Explanation of Boosting [Online], available: <http://arxiv.org/abs/1009.3613>, September 19, 2010
- 62 Martínez-Muñoz G, Suárez A. Aggregation ordering in bagging. In: Proceedings of the 2004 IASTED International Conference on Artificial Intelligence and Applications. Innsbruck, Austria: Acta Press, 2004. 258–263
- 63 Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 2002, **137**(1–2): 239–263
- 64 Zhang Chun-Xia, Zhang Jiang-She. A survey of selective ensemble learning algorithms. *Chinese Journal of Computers*, 2011, **34**(8): 1399–1410  
(张春霞, 张讲社. 选择性集成学习算法综述. 计算机学报, 2011, **34**(8): 1399–1410)
- 65 Tuve L, Johansson U, Bostrom H. On the use of accuracy and diversity measures for evaluating and selecting ensembles of classifiers. In: Proceedings of the 7th International Conference on Machine Learning and Applications. San Diego, California, USA: IEEE, 2008. 127–132
- 66 Martinez-Munoz G, Hernandez-Lobato D, Suarez A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2009, **31**(2): 245–259
- 67 Frank A, Asuncion A. UCI Machine Learning Repository [Online], available: <http://www.ics.uci.edu/~mllearn/>, October 25, 2010
- 68 Image Processing Research Laboratory in Hefei University of Technology [Online], available: <http://wwwil.hfut.edu.cn/organ/images/imagelab/download/usps.htm>, November 19, 2007



**杨 春** 北京科技大学博士研究生. 主要研究方向为图像处理与模式识别.  
E-mail: ych.learning@gmail.com  
(**YANG Chun** Ph.D. candidate at University of Science and Technology Beijing. His research interest covers image processing and pattern recognition.)



**殷绪成** 北京科技大学副教授. 主要研究方向为模式识别, 机器学习, 信息检索. 本文通信作者.

E-mail: xuchengyin@ustb.edu.cn

(**YIN Xu-Cheng** Associate professor at University of Science and Technology Beijing. His research interest covers pattern recognition, machine

learning, and information retrieval. Corresponding author of this paper.)



**郝红卫** 中国科学院自动化研究所教授. 主要研究方向为大规模语义计算理论和 技术, 大规模机器学习理论, 海量信息智能处理.

E-mail: hongwei.hao@ia.ac.cn

(**HAO Hong-Wei** Professor at Institute of Automation, Chinese Academy of Sciences. His research inter-

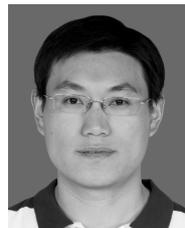
est covers large-scale semantic computing theory and technology, large-scale machine learning theory, and intelligent massive information processing.)



**闫 琰** 北京科技大学博士研究生. 主要研究方向为图像处理与模式识别.

E-mail: happyyyan@163.com

(**YAN Yan** Ph.D. candidate at University of Science and Technology Beijing. Her research interest covers image processing and pattern recognition.)



**王志彬** 国家农业信息化工程技术研究中心助理研究员, 博士. 主要研究方向为图像处理与模式识别.

E-mail: wangzb@necita.org.cn

(**WANG Zhi-Bin** Ph.D, assistant professor at National Engineering research Center for Information Technology in Agriculture. His research inter-

est covers image processing and pattern recognition.)