# 集成学习的泛化误差和 AUC 分解理论及其在权重优化中的应用

姜正申[1] 刘宏志[2),3)] 付 彬[2)] 吴中海[2),3)]

[1)](北京大学信息科学技术学院 北京 100871)

[2)](北京大学软件与微电子学院 北京 102600)

[3)](北京大学软件工程国家工程研究中心 北京 100871)

**摘 要** 集成学习是机器学习领域的一个重要分支,其通过整合多个学习器以获得比单个学习器更好的学习效果.多样性和间隔被认为是影响集成学习效果的两个关键因素.现有研究大多是对这两个因素的影响单独进行分析.该文的研究集中于泛化误差、AUC、多样性和间隔之间关系及其在基分类器的权重优化中的应用.该文首先在泛化误差分解理论的基础上,给出了 AUC 的分解定理.进一步地,该文讨论了泛化误差、AUC、多样性与间隔之间的关系,并指出常用的最大化间隔方法在降低经验误差的同时,也会降低基分类器之间的多样性,进而导致过拟合问题.基于这些理论结果,该文提出了两种新的基分类器的权重优化算法,通过求解一个二次优化问题,实现在准确性和多样性之间的最佳平衡.在 35 个公开数据集上的实验结果表明,该文所提出的算法在绝大多数情况下都优于现有常用的集成方法.

**关键词** 集成学习;泛化误差分解;AUC;多样性;间隔;权重优化

**中图法分类号** TP18 **DOI 号** 10.11897/SP.J.1016.2019.00001

# Decomposition Theories of Generalization Error and AUC in Ensemble Learning with Application in Weight Optimization

JIANG Zheng-Shen[1] LIU Hong-Zhi[2),3)] FU Bin[2)] WU Zhong-Hai[2),3)]

[1)](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

[2)](*School of Software and Microelectronics, Peking University, Beijing* 102600)

[3)](*National Engineering Research Center For Software Engineering, Peking University, Beijing* 100871)

**Abstract** Ensemble learning is an important branch of machine learning, which integrates multiple learners to obtain better learning performance than single learners. It has been widely accepted that the base learners in an ensemble model should be both accurate and diverse to achieve good performance. Among the factors that affect the performance of ensemble learning, diversity and margin have been considered to be two key ones. Most of the existing studies tried to analyze the impact of these two factors separately, and mainly focused on their impacts on the classification or regression error of the ensemble model. AUC is an important criterion for evaluating the classification performance of the learners. It is a pair-wise criterion that is used to evaluate the probability that the positive samples achieve higher scores than the negative ones. However, few studies have focused on the relationship between AUC and diversity or margin. In this paper, we proposed two AUC decomposition theorems based on the Ambiguity Decomposition, which is one

收稿日期:2017-08-09;在线出版日期:2018-05-15.本课题得到国家自然科学基金(61232005)、国家"八六三"高技术研究发展计划项目(2015AA016009)资助.**姜正申**,男,1990 年生,博士研究生,主要研究方向为机器学习、推荐系统. E-mail: jiangzhengshen@163.com. **刘宏志**(通信作者),男,1982 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为推荐系统、信息融合和集成学习. E-mail: liuhz@pku.edu.cn. **付 彬**,男,1990 年生,博士研究生,主要研究方向为推荐系统. **吴中海**,男,1968 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为情境感知服务、云安全与隐私保护、嵌入式智能、大数据与信息融合.

of the most important generalization error decomposition theory. Further, we discussed the relationship between generalization error, AUC, diversity and margin. According to our theoretical results, the commonly used margin maximization method not only reduces the empirical error, but also reduces the diversity among the base classifiers, which leads to the problem of over-fitting. Similar results also hold in the case of AUC. Due to the reduction in diversity, methods like margin maximization could not achieve satisfactory generalization performance with respect to classification error or AUC. Based on these theoretical results, we proposed two new weight optimization algorithms to combine the base classifiers, and the targets of these two algorithms are classification error and AUC, respectively. Existing weight optimization methods usually suffer from over-fitting problem, and these methods usually use a term that is related to diversity to avoid over-fitting. However, due to the unclear definition of diversity and the difficulty of parameter tuning, such methods usually could not achieve satisfactory performance. Inspired by our theoretical results, in both of our proposed algorithms, we got use of the margin of the ensemble model. Moreover, the objective functions are quadratic functions of the margin, thus the learning procedure can be guaranteed to be convergence. By introducing a trade-off parameter $p$, we optimized the margin to a proper level instead of maximization. Therefore, we could achieve an optimal balance between accuracy and diversity. Since the parameter $p$ is highly related to the regularization parameter, in practice, we could fix $p$ and determine the regularization parameter using grid search, thus the proposed algorithms are highly applicable. We evaluated our algorithms in 35 open datasets. The experimental results confirm that our algorithms are not sensitive to the parameter $p$. Compared with other commonly used ensemble methods, the proposed algorithms achieve significantly better results in most cases. Both our theoretical and experimental results show that there is a strong connection between diversity and the margin, and through exploiting the relationship between them, the generalization ability of ensemble models could be effectively improved.

**Keywords** ensemble learning; generalization error decomposition; area under the ROC curve; diversity; margin; weight optimization



## 1 引 言

作为机器学习领域的一个重要分支，集成学习通过将多个机器学习模型的结果综合起来，以提高其学习效果[1-2]. 这种方法类似于人类社会中的委员会. 通常一个委员会在进行决策时，会综合考虑每个成员的意见，以做出更优的决策.

组成一个集成模型的各个基本的学习器称为“基学习器”. 这些基学习器可以使用相同的学习算法以不同的参数或数据进行训练，也可以使用不同的学习算法进行训练，例如决策树、神经网络以及支持向量机等. 使用同一种学习算法生成所有的基学习器，这样的集成方法称为“同质集成”或“同态集成”. 使用不同的学习算法来训练基学习器，这样的集成方法则称为“异质集成”或“异态集成”[3].

在集成学习领域中，常常通过引入“多样性”来避免过拟合，从而提高集成模型的泛化能力. 所谓多样性，是指集成模型所使用的基学习器之间的差异. 一个被广为接受的观点是：为了得到一个较好的集成模型，基学习器需要既准确又多样[2]. 理想状态下，希望这些基学习器之间相互独立且相互补充，即单个基学习器都会犯错，但由于犯错的方面不同，集成以后效果会好于任何一个基学习器. 已有研究表明，多样性在集成学习中可以起到正则化的作用[4]，而正则化方法是机器学习方法中用来避免过拟合的常用方法.

尽管多样性被认为是集成学习中的一个关键因素，然而对于如何度量多样性，目前还没有统一的方法. 已有方法大多是基于经验启发式地定义，包括相关系数、Q 统计、Kappa 统计、信息熵[5]等.

集成学习中的另一个重要概念是“间隔”. 这一

概念是为解释 Boosting 算法[6]的良好性质时提出的[7]. 对于二分类问题，假定样本标记值 $y\in\{-1,+1\}$，那么分类器 $h$ 在样本$(\boldsymbol{x},y)$上的间隔定义为

$$\text{margin}=yh(\boldsymbol{x}) \tag{1}$$

其中 $\boldsymbol{x}$ 为样本特征向量，$h(\boldsymbol{x})$为模型的预测值. 间隔理论指出，样本间隔对集成模型的泛化误差具有重要影响. 1998 年，Schapire 等人[7]证明，在其它条件不变的情况下，集成模型在所有样本上的最小间隔越大，泛化误差越低. 最近的研究[8]给出了第 $k$ 小间隔与泛化误差界之间的关系，由此推论出间隔的分布具有重要意义.

一个尚未被充分研究的问题是多样性与间隔之间的关联. 这方面的一个代表性研究是“好”与“坏”多样性理论[9]. 此理论分析了在二分类问题中，损失函数为 0-1 损失且基分类器使用多数投票进行集成的情况下，多样性与间隔的关系. 研究证明，对于集成模型分类正确的样本，间隔为正数，且间隔越小则多样性越大，同时，此时的多样性为“好”多样性，因此间隔越小越好；而对于分类错误的样本，其间隔为负数，其绝对值越大则多样性越大，但由于此时的多样性为“坏”多样性，因此间隔绝对值越小越好. 这一研究的缺点是仅适用于 0-1 损失和多数投票集成方法，适用范围比较有限.

本文主要研究集成学习中的泛化误差和 AUC 的分解定理，并基于此得出了多个关于多样性和间隔之间关联的结论. 进一步，本文还基于这些理论结果提出了两种新的基分类器权重优化算法，并在 35 个数据集上验证了本文所提算法的有效性.

本文的主要贡献在于：

(1) 基于推广的分歧分解定理，提出了集成模型的 AUC 分解定理(定理 3 和定理 4)，并给出了一些有用的推论，从而为多样性的理解和度量提供了新的角度和方法；

(2) 基于泛化误差的分歧分解定理和 AUC 分解定理，给出了泛化误差、AUC、多样性与间隔之间的关系，从而在多样性和间隔之间建立了联系，且这些结论适用于众多常用损失函数，因此具有广泛的适用性；

(3) 基于本文理论结果提出了两种新的基分类器的权重优化算法，分别用来优化分类误差和 AUC. 本文在 35 个数据集上的实验表明：本文所提算法在 0.05 的显著性水平上优于绝大部分其它集成方法，从而证明了本文算法的有效性.

## 2　相关工作

目前，集成学习泛化误差的分解理论主要包括分歧分解[10]和偏差-方差-协方差分解[11]. 本文主要涉及其中的分歧分解，因此本节将集中于分歧分解相关研究的介绍.

经典的分歧分解在1995 年由Krogh 和Vedelsby[10]提出，并在 2005 年由 Brown 等人[12]做了进一步阐释. 对单个样本，其形式为

$$(f_{\text{ens}}-y)^2=\sum_{\alpha}\omega_{\alpha}(f_{\alpha}-y)^2-\sum_{\alpha}\omega_{\alpha}(f_{\alpha}-f_{\text{ens}})^2 \tag{2}$$

或

$$e=\bar{e}-\bar{a} \tag{3}$$

第一个式子中，$f_{\alpha}$为第 $\alpha$ 个学习器的输出值，$y$ 为此样本的真值，$f_{\text{ens}}$为所有学习器输出值的加权平均，即 $f_{\text{ens}}=\sum_{\alpha}\omega_{\alpha}f_{\alpha}$，并且满足$\sum_{\alpha}\omega_{\alpha}=1$ 和 $\omega_{\alpha}\geq 0$. 第二个式子中，$e$ 代表集成模型的误差，$\bar{e}$ 代表各基学习器的平均误差，而 $\bar{a}$ 就是基学习器之间的分歧.

对上式求期望即得到整个数据集上的泛化误差分解，形式如下

$$E=\bar{E}-\bar{A} \tag{4}$$

这就是分歧分解理论的常见形式，其中，$E$ 为集成模型在数据集上的预测误差，$\bar{E}$ 为基学习器的平均误差，$\bar{A}$ 则被称为分歧，这一项通常被认为与多样性相关. 需要注意的是，经典的分歧分解是在损失函数为平方损失的假设下推导出来的，因此只适用于回归问题.

文献[13]基于泰勒定理将损失函数由平方损失推广到了任意二阶可导的函数，由此将经典的分歧分解理论由回归问题推广到分类问题中，扩展了这一理论的适用范围. 他们的理论结果表明：在分类问题中，分歧分解遵从相同的形式，即 $E=\bar{E}-\bar{A}$，区别只是在分歧项 $\bar{A}$ 中多出了一个需要估计的参数.

目前对集成模型泛化能力的分析还主要集中于对分类或回归误差的分解，而 AUC 作为一种重要的衡量分类器分类性能的指标，还没有类似的结论. 本文基于前面所述的分歧分解理论，首次给出了 AUC 的分解定理，同时给出了一些有用的推论，并讨论了间隔与 AUC 分解中各项的关系. 这些理论结果对于集成模型的算法设计和分析都有较强的指导意义.

本文还基于这些理论结果提出了两种新的基分

类器权重优化算法.权重优化是集成模型中的一个重要步骤.在训练了一组基分类器之后,需要对这些基分类器的预测值进行集成.在所有的集成方法中,最简单也最常用的方法是多数投票法,这种方法在二分类问题中等价于等权平均法.理论上讲,使用加权平均的方法进行集成可以取得比多数投票法更好的结果.然而,在实践中,要获得一组较好的权重并不容易,主要难点在于权重的优化过程十分容易发生过拟合,导致集成模型的泛化能力较差.

目前,确定基分类器的权重的主要方法包括两大类:最优化方法和贝叶斯方法.

最优化方法中,比较常见的做法是将基分类器的预测值作为特征,再训练一个新的模型来预测样本的类别,这种方法也称为 Stacking 方法[14],新训练的学习器也称为元学习器.这种方法的缺点是容易发生过拟合,并且模型对元学习器的参数比较敏感,因此在实际应用中需要较多地依赖经验进行模型和参数的调整.

最优化方法的另一个典型代表是文献[15]和文献[16]的工作.在这两个方法中,作者将权重的计算问题视为下面的最优化问题:

$$\min_{\boldsymbol{\omega}} f_{\text{loss}}(\boldsymbol{\omega}) \tag{5}$$

$$\text{s.t. } f_{\text{sparsity}}(\boldsymbol{\omega}) \geq t_1,\ f_{\text{diversity}}(\boldsymbol{\omega}) \geq t_2,\ \omega_\alpha \geq 0$$

其中,$t_1$ 和 $t_2$ 是控制模型稀疏程度和多样性的两个参数.采用不同损失函数、稀疏性约束、多样性度量方法就可以得到不同的最优化问题,进而可以使用对应的最优化方法求解.实践中,为了保证目标函数满足凸性条件,损失函数和多样性度量方法往往不能随意选取.而且,这种带约束的最优化问题往往难以直接求解,需要转化为无约束优化问题,而转化后的目标函数不仅会引入多个待定的参数,而且最终的求解结果也常常不够理想.

贝叶斯方法又称为贝叶斯模型平均法[17].设所有候选的基分类器构成集合 $H$,样本集合为 $D$,这种方法的基本原理是,假定刻画数据真实分布的函数 $h^{\rightarrow}$ 包含在所有候选的基学习器中,即 $h^{\rightarrow} \in H$,对于一个样本$(\boldsymbol{x},y)$,其后验分布为

$$p(y|\boldsymbol{x},D) = \sum_{h \in H} p(h^{\rightarrow} = h|D)\, p(y|\boldsymbol{x},h) \tag{6}$$

因此,按照最优贝叶斯决策,最终的集成模型为

$$H^{\rightarrow}(\boldsymbol{x}) = \arg\min_{y \in \boldsymbol{y}} \sum_{y' \in \boldsymbol{y}} p(y'|\boldsymbol{x},D)\, loss(y',y) \tag{7}$$

这样,$p(h^{\rightarrow}=h|D)$ 就相当于基学习器的权重,对这一分布的估计就成为模型平均法的关键问题.

在贝叶斯模型平均中,$p(h^{\rightarrow}=h|D)$ 是通过设定先验分布和统计推断得到的.相比最优化方法,贝叶斯方法不容易发生过拟合,所得模型也更加稳定.但是,这一方法也存在一些缺点.首先,先验概率的设定对结果影响较大,找到合适的先验分布比较困难.其次,有时同一个方法内会引入若干个互相矛盾的先验分布,导致模型难以解释.

本文所提出的权重优化算法隶属于最优化方法.本文所提方法不仅目标函数简单,容易求解,而且在准确性和多样性之间有较好的平衡,从而有效避免了过拟合问题,提高了集成模型的泛化能力.

## 3 符号说明

本文将集中于二分类问题的讨论,但本文的理论结果也同样适用于回归问题.

本文遵从与概率近似正确PAC(Probably Approximately Correct)学习理论[18]相同的假设.机器学习任务 $D$ 定义为 $\boldsymbol{x}\times\boldsymbol{y}$ 空间上的一个概率分布,其中 $\boldsymbol{x}$ 为样本空间,$\boldsymbol{y}$ 为标记空间.从空间 $\boldsymbol{x}\times\boldsymbol{y}$ 中抽取出来的样本表示为$(\boldsymbol{x},y)$,其中 $\boldsymbol{x}\in\mathbb{R}^m$ 为表示样本特征的向量,$y\in\boldsymbol{y}$ 表示样本的类别标记.

本文使用 $h$ 表示用来预测样本类别的分类器,$f$ 表示分类器 $h$ 在样本 $\boldsymbol{x}$ 上的预测值,即 $f=h(\boldsymbol{x})$.

### 3.1 集成方法

集成学习通常会训练一组分类器,然后通过不同方法进行集成.本文将这些分类器称为基分类器,并假定基分类器以加权平均的方式进行集成.具体地,集成模型表示为

$$H(\boldsymbol{x}) = \sum_{\alpha} \omega_\alpha h_\alpha(\boldsymbol{x}) \tag{8}$$

或

$$f_{\text{ens}} = \overline{f} = \sum_{\alpha} \omega_\alpha f_\alpha \tag{9}$$

其中,$\omega_\alpha$ 为基分类器 $h_\alpha$ 的权重.

### 3.2 损失函数和泛化误差

在机器学习任务中,通常使用损失函数来评价分类器在一个样本上的预测效果.本文使用 $l(f,y)$ 来表示损失函数.常用的损失函数包括平方损失 $l(f,y)=(f-y)^2$、对数损失 $l(f,y)=\log(1+\mathrm{e}^{-yf})$ 和指数损失 $l(f,y)=\mathrm{e}^{-yf}$ 等[19].

泛化误差通常定义为损失函数在整个样本空间 $D$ 上的期望,即

$$E_D\{l(f,y)\} = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D} l(h(\boldsymbol{x}),y) \tag{10}$$

### 3.3 AUC 指标

在二分类问题中,另外一种重要且常用的衡量分

类性能的方法是受试者工作特征(Receiver Operating Characteristic，ROC)曲线下的面积(Area Under the ROC Curve，AUC)，这是一种基于样本对的度量方法. 假定样本集合中含有 $n_+$ 个正样本和 $n_-$ 个负样本，即

$$D=\{(\boldsymbol{x}_1^+,+1),(\boldsymbol{x}_2^+,+1),\cdots,(\boldsymbol{x}_{n_+}^+,+1),$$
$$(\boldsymbol{x}_1^-,-1),(\boldsymbol{x}_2^-,-1),\cdots,(\boldsymbol{x}_{n_-}^-,-1)\}.$$

对于一个分类器 $h$，AUC 定义为[20]

$$\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\frac{\mathbb{I}[h(\boldsymbol{x}_i^+)>h(\boldsymbol{x}_j^-)]+\frac{1}{2}\mathbb{I}[h(\boldsymbol{x}_i^+)=h(\boldsymbol{x}_j^-)]}{n_+n_-} \tag{11}$$

其中，$\mathbb{I}[\cdot]$ 为指示函数，参数为真时其值为 1，否则为 0.

由于 AUC 为非凸函数，因此在实际应用时，常常使用代理函数作为近似. 具体地，可用下面函数作为 AUC 的近似：

$$-\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}l(\Delta f_{ij}) \tag{12}$$

其中，

$$\Delta f_{ij}=h(\boldsymbol{x}_i^+)-h(\boldsymbol{x}_j^-) \tag{13}$$

且 $l(\cdot)$ 为代理函数. 典型的代理函数包括二次代理函数 $l(f)=(f-1)^2$(如 OPAUC[20])、对数代理函数 $l(f)=\log(1+e^{-f})$(如 RankNet[21])和指数代理函数 $l(f)=e^{-f}$(如 RankBoost[22])等. 本文将使用损失函数来指代这些代理函数.

对于一个集成模型 $H(\boldsymbol{x})=\sum_\alpha\omega_\alpha h_\alpha(\boldsymbol{x})$，由于

$$\Delta f_{ij}^{\text{ens}}=H(\boldsymbol{x}_i^+)-H(\boldsymbol{x}_j^-) \tag{14}$$

因此有

$$\Delta f_{ij}^{\text{ens}}=\sum_\alpha\omega_\alpha\Delta f_{ij}^\alpha \tag{15}$$

相应的，集成模型的 AUC 可写为

$$-\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}l(\Delta f_{ij}^{\text{ens}}) \tag{16}$$

对于按式(12)定义的 AUC 近似函数，其值越大，表示分类器的分类效果越好. 为了叙述上的方便，并与其它泛化误差的研究保持形式上的一致，学界常常将其转化为类似误差的形式，即定义

$$\mathcal{L}^\alpha=\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}l(\Delta f_{ij}^\alpha) \tag{17}$$

$$\mathcal{L}^{\text{ens}}=\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}l(\Delta f_{ij}^{\text{ens}}) \tag{18}$$

来分别反映学习器 $h_\alpha$ 和集成模型 $H$ 的 AUC 指标. 其值越低，表示学习器的 AUC 指标越高.

在本文余下的讨论中，将采用 $\mathcal{L}^\alpha$ 和 $\mathcal{L}^{\text{ens}}$ 的形式来作为 AUC 的近似函数.

### 3.4　间隔

对于二分类问题，假定 $y\in\{-1,+1\}$，那么基分类器 $h_\alpha$ 在样本 $(\boldsymbol{x},y)$ 上的间隔定义为

$$\text{margin}_\alpha=yh_\alpha(\boldsymbol{x})=yf_\alpha \tag{19}$$

对于一个集成模型 $H(\boldsymbol{x})=\sum_\alpha\omega_\alpha h_\alpha(\boldsymbol{x})$，其预测值满足 $f_{\text{ens}}=\bar{f}=\sum_\alpha\omega_\alpha f_\alpha$，因此 $H$ 在样本 $(\boldsymbol{x},y)$ 上的间隔为

$$\text{margin}_{\text{ens}}=y\bar{f}=\sum_\alpha\omega_\alpha\text{margin}_\alpha \tag{20}$$

以 AUC 衡量分类性能时，由于 AUC 是成对的评价函数，因此间隔应定义在成对的样本上. 具体的，对一个正样本 $\boldsymbol{x}_i^+$ 和一个负样本 $\boldsymbol{x}_j^-$，分类器 $h_\alpha(\boldsymbol{x})$ 的目标为使预测两样本的类别标记之间的差值 $h(\boldsymbol{x}_i^+)-h(\boldsymbol{x}_j^-)$ 尽可能大. 考虑到间隔的含义，本文定义分类器 $h_\alpha$ 在样本对 $(\boldsymbol{x}_i^+,\boldsymbol{x}_j^-)$ 上的间隔为

$$\text{margin}_\alpha^{\text{auc}}=\Delta f_{ij}^\alpha=h_\alpha(\boldsymbol{x}_i^+)-h_\alpha(\boldsymbol{x}_j^-) \tag{21}$$

类似地，集成模型 $H$ 在样本对 $(\boldsymbol{x}_i^+,\boldsymbol{x}_j^-)$ 上的间隔为

$$\text{margin}_{\text{ens}}^{\text{auc}}=\Delta f_{ij}^{\text{ens}}=\sum_\alpha\omega_\alpha\Delta f_{ij}^\alpha \tag{22}$$

## 4　集成学习的泛化误差和 AUC 分解理论

本节将首先给出文献[13]中提出的两种推广的分歧分解理论，然后基于其理论结果，提出两种关于 AUC 的分解定理，并给出一些有用的结论. 需要指出的是，尽管下面的讨论都假定 $\sum_{\alpha=1}^T\omega_\alpha=1$，但下面的结论可以轻易地推广到 $\sum_{\alpha=1}^T\omega_\alpha\neq1$ 的情形. 此外，下面的推导并不要求 $\omega_\alpha\geq0$ 这一条件.

### 4.1　推广的泛化误差分歧分解理论

文献[13]提出了两种推广的分歧分解理论，将传统的分歧分解由回归问题推广到了分类问题中. 具体地，文章提出了下面两个定理.

**定理 1**(第一种推广的分歧分解)．对于一个二分类问题或回归问题，假定已经训练好了一组学习器 $\{h_1,h_2,\cdots,h_T\}$，并且这些学习器使用加权平均的方式进行集成，即 $f_{\text{ens}}=\sum_{\alpha=1}^T\omega_\alpha f_\alpha$，其中 $f_\alpha=h_\alpha(\boldsymbol{x})$，且 $\sum_{\alpha=1}^T\omega_\alpha=1$. 那么，对于任意二阶可导的损失函数，

集成模型的误差可以分解为

$$E=\bar{E}-\bar{A} \tag{23}$$

其中

$$E=E_D\{l(f_{\text{ens}},y)\} \tag{24}$$

$$\bar{E}=\sum_{\alpha=1}^{T}\omega_\alpha E_D\{l(f_\alpha,y)\} \tag{25}$$

$$\bar{A}=\frac{1}{2}\sum_{\alpha=1}^{T}\omega_\alpha E_D\{l''(f_\sigma^*,y)f_\sigma^2-l''(f_{\text{ens}}^*,y)f_{\text{ens}}^2\} \tag{26}$$

上式中，$E_D\{\cdot\}$表示在样本空间中求期望，$f_\alpha^*$ 为介于 0 和 $f_\alpha$之间的实数，$f_{\text{ens}}^*$为介于 0 和 $f_{\text{ens}}$之间的实数. 类似于拉格朗日中值定理，$f_\alpha^*$ 和 $f_{\text{ens}}^*$的具体数值依赖于损失函数 $l(f,y)$的具体形式以及 $f_\alpha$和 $f_{\text{ens}}$的具体取值.

**定理 2**(第二种推广的分歧分解). 在与定理 1 相同的条件和假设下，对于任意二阶可导的损失函数，集成模型的损失还可以做另外一种分解，其中 $E$ 和 $\bar{E}$ 与定理 1 具有相同的形式，而 $\bar{A}$ 则为

$$\bar{A}=\frac{1}{2}\sum_{\alpha=1}^{T}\omega_\alpha E_D\{l''(f_\alpha^*,y)(f_\alpha-f_{\text{ens}})^2\} \tag{27}$$

其中，$f_\alpha^*$为介于 $f_\alpha$和 $f_{\text{ens}}$之间的实数.

在上述两个定理中，$\bar{A}$ 即为“分歧项”，它常被认为与集成学习中至关重要的“多样性”存在关联，因此常被用来作为多样性的评价指标.

### 4.2 AUC 分解理论

本文在推广的分歧分解理论的基础上，进一步的提出了如下的两种关于 AUC 的分解定理.

**定理 3**(第一种形式的 AUC 分解定理). 对于一个二分类问题，假定已经训练好了一组学习器 $\{h_1,h_2,\cdots,h_T\}$，并且这些学习器使用加权平均的方式进行集成，即 $f_{\text{ens}}=\sum_{\alpha=1}^{T}\omega_\alpha f_\alpha$，其中 $f_\alpha=h_\alpha(\boldsymbol{x})$，且 $\sum_{\alpha=1}^{T}\omega_\alpha=1$. 并且，由 $\Delta f_{ij}^\alpha$和 $\Delta f_{ij}^{\text{ens}}$的定义，有 $\Delta f_{ij}^{\text{ens}}=\sum_\alpha\omega_\alpha\Delta f_{ij}^\alpha$. 那么，对于任意二阶可导的损失函数$l(\cdot)$，按式(12)所定义的集成模型的 AUC 指标可以分解为

$$\mathcal{L}=\bar{\mathcal{L}}-\bar{\mathcal{A}} \tag{28}$$

其中

$$\mathcal{L}=\mathcal{L}^{\text{ens}} \tag{29}$$

$$\bar{\mathcal{L}}=\sum_\alpha\omega_\alpha\mathcal{L}^\alpha \tag{30}$$

$$\bar{\mathcal{A}}=\frac{1}{2n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\sum_{\alpha=1}^{T}\omega_\alpha l''(f^*)(\Delta f_{ij}^\alpha-\Delta f_{ij}^{\text{ens}})^2 \tag{31}$$

且 $f^*$ 为介于 $\Delta f_{ij}^\alpha$和 $\Delta f_{ij}^{\text{ens}}$之间的实数.

证明. 根据泰勒定理，函数 $l(\Delta f_{ij}^\alpha)$可以在 $\Delta f_{ij}^{\text{ens}}$附近展开为

$$l(\Delta f_{ij}^\alpha)=l(\Delta f_{ij}^{\text{ens}})+l'(\Delta f_{ij}^{\text{ens}})(\Delta f_{ij}^\alpha-\Delta f_{ij}^{\text{ens}})+\frac{1}{2}l''(f^*)(\Delta f_{ij}^\alpha-\Delta f_{ij}^{\text{ens}})^2 \tag{32}$$

其中 $f^*$ 为介于 $\Delta f_{ij}^\alpha$和 $\Delta f_{ij}^{\text{ens}}$之间的实数.

对上式进行加权平均，可得

$$\sum_\alpha\omega_\alpha l(\Delta f_{ij}^\alpha)=l(\Delta f_{ij}^{\text{ens}})+\frac{1}{2}\sum_\alpha\omega_\alpha l''(f^*)(\Delta f_{ij}^\alpha-\Delta f_{ij}^{\text{ens}})^2 \tag{33}$$

对所有正负样本对求和，得

$$\sum_\alpha\omega_\alpha\mathcal{L}^\alpha-\mathcal{L}^{\text{ens}}=\frac{1}{2n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\sum_{\alpha=1}^{T}\omega_\alpha l''(f^*)(\Delta f_{ij}^\alpha-\Delta f_{ij}^{\text{ens}})^2 \tag{34}$$

即为定理所述形式. 证毕.

**定理 4**(第二种形式的 AUC 分解定理). 在与定理 3 相同的条件和假设下，集成模型的 AUC 的代理函数$\mathcal{L}$可以做另外一种形式的分解，其中$\mathcal{L}$和$\bar{\mathcal{L}}$与定理 3 具有相同的形式，而$\bar{\mathcal{A}}$则为

$$\bar{\mathcal{A}}=\frac{1}{2n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\sum_{\alpha=1}^{T}\omega_\alpha[l''(f_\alpha^*)(\Delta f_{ij}^\alpha)^2-l''(f_{\text{ens}}^*)(\Delta f_{ij}^{\text{ens}})^2] \tag{35}$$

其中，$f_\alpha^*$ 为介于 0 和 $\Delta f_{ij}^\alpha$之间的实数，$f_{\text{ens}}^*$为介于 0 和 $\Delta f_{ij}^{\text{ens}}$之间的实数.

证明. 根据泰勒定理，函数$l(\Delta f_{ij}^\alpha)$可以在 0 点附近展开为

$$l(\Delta f_{ij}^\alpha)=l(0)+l'(0)(\Delta f_{ij}^\alpha)+\frac{1}{2}l''(f_\alpha^*)(\Delta f_{ij}^\alpha)^2 \tag{36}$$

其中，$f_\alpha^*$为介于 0 和 $\Delta f_{ij}^\alpha$之间的实数. 对式(36)加权平均，可得

$$\sum_\alpha\omega_\alpha l(\Delta f_{ij}^\alpha)=l(0)+l'(0)(\Delta f_{ij}^{\text{ens}})+\frac{1}{2}\sum_\alpha\omega_\alpha l''(f_\alpha^*)(\Delta f_{ij}^\alpha)^2 \tag{37}$$

同理，函数 $l(\Delta f_{ij}^{\text{ens}})$也可以在 0 点附近展开

$$l(\Delta f_{ij}^{\text{ens}})=l(0)+l'(0)\Delta f_{ij}^{\text{ens}}+\frac{1}{2}l''(\Delta f_{\text{ens}}^*)(\Delta f_{ij}^{\text{ens}})^2 \tag{38}$$

其中 $f_{\text{ens}}^*$为介于 0 和 $\Delta f_{ij}^{\text{ens}}$之间的实数.

以上两式相减，得

$$\sum_\alpha\omega_\alpha l(\Delta f_{ij}^\alpha)-l(\Delta f_{ij}^{\text{ens}})=\frac{1}{2}\sum_\alpha\omega_\alpha l''(f_\alpha^*)(\Delta f_{ij}^\alpha)^2-l''(f_{\text{ens}}^*)(\Delta f_{ij}^{\text{ens}})^2 \tag{39}$$

对所有正负样本对求和，得

$$\sum_{\alpha}\omega_{\alpha}\mathcal{L}^{\alpha}-\mathcal{L}^{\mathrm{ens}}=\frac{1}{2n_{+}n_{-}}\sum_{i=1}^{n_{+}}\sum_{j=1}^{n_{-}}\sum_{\alpha=1}^{T}\omega_{\alpha}\left[l''(f_{\alpha}^{*})(\Delta f_{ij}^{\alpha})^{2}-l''(f_{\mathrm{ens}}^{*})(\Delta f_{ij}^{\mathrm{ens}})^{2}\right] \tag{40}$$

即为定理所述形式. 证毕.

类似于泛化误差的分歧分解，定理 3 和定理 4 中的 $\bar{A}$ 就相当于“分歧”，可以用来衡量基分类器之间的多样性.

### 4.3　一些有用的推论

根据上面四个定理，可以得出下面推论.

**推论 1**(集成学习的有效性). 若损失函数 $l(\cdot)$ 为凸函数，则有

(1) 集成模型的泛化误差一定低于各基分类器的平均误差；

(2) 以代理函数衡量的集成模型的 AUC 一定好于各基分类器的平均 AUC.

这是因为在定理 2 和定理 3 中，当损失函数 $l(\cdot)$ 为凸函数时，$l''\geq 0$ 恒成立，因此有 $A\geq 0$ 与 $\bar{A}\geq 0$. 因此可知 $E\leq\bar{E}$ 与 $\mathcal{L}\leq\bar{\mathcal{L}}$ 恒成立，由此可得上述推论.

上述推论证明了集成学习的有效性. 在实际应用中，损失函数通常都是凸的，例如平方损失、对数损失和指数损失等. 因此，上述推论具有十分广泛的适用性.

**推论 2**(AUC 分解的二阶近似). 在与定理 3 相同的条件和假设下，AUC 分解中的 $\bar{A}$ 项具有如下的二阶近似形式

$$\bar{A}=\frac{l''(0)}{2n_{+}n_{-}}\sum_{i=1}^{n_{+}}\sum_{j=1}^{n_{-}}\left[\sum_{\alpha=1}^{T}\omega_{\alpha}(\Delta f_{ij}^{\alpha})^{2}-(\Delta f_{ij}^{\mathrm{ens}})^{2}\right] \tag{41}$$

且其误差不超过 $o((\Delta f)^{2})$，其中 $\Delta f=\max_{\alpha,i,j}\{\Delta f_{ij}^{\alpha}\}$.

这可以通过对损失函数进行带有皮亚诺(Peano)型余项的泰勒定理推导出来. 此处省略了其详细的推导过程.

由推论 2，在 $|\Delta f_{ij}^{\alpha}|\leq 1$ 时，可以使用推论 2 中的 $\bar{A}$ 作为分歧项的一个较好的近似. 而 $|\Delta f_{ij}^{\alpha}|\leq 1$ 通常比较容易满足，因为如果基分类器的输出 $f=h(\boldsymbol{x})$ 为概率值，那么就有 $0\leq h(\boldsymbol{x})\leq 1$，这样根据 $\Delta f_{ij}^{\alpha}$ 的定义 $\Delta f_{ij}^{\alpha}=h_{\alpha}(\boldsymbol{x}_{i}^{+})-h_{\alpha}(\boldsymbol{x}_{j}^{-})$，可知 $|\Delta f_{ij}^{\alpha}|\leq 1$ 会自然满足. 对于输出值不是概率值的模型，也可以使用概率校准方法[23]来将输出值转化为概率值. 因此，上述近似形式具有广泛的应用场景.

### 4.4　讨论

泛化误差的分歧分解理论(定理 1 和定理 2)与 AUC 的分解理论(定理 3 和定理 4)存在密切的关联. 二者均基于泰勒定理推导得到，结论的形式也是类似的. 它们都可以用来对集成模型的泛化能力进行分析，主要区别在于定理 1 和定理 2 是针对分类误差进行的分解，而定理 3 和定理 4 是针对 AUC 进行的分解. 分类误差是针对单个样本而言的，整个数据集的分类误差是所有样本分类误差的平均值. 而 AUC 是一种成对的衡量指标，计算时是针对每一对样本进行的，对于整个数据集，它衡量的是正样本比负样本取得更高的预测值的概率. 正因如此，AUC 分解定理比分类误差的分解具有更为复杂的形式.

分类误差和 AUC 从两个不同的方面刻画了模型的泛化能力. 相比于分类误差，AUC 指标更加注重样本之间的顺序关系. 一般的分类问题通常使用分类误差来反映模型的分类能力，而对于排序问题(例如信息检索和推荐系统)，则更加关心 AUC 指标. 因此，定理 1、2 和定理 3、4 都具有很广泛的应用场景.

泛化误差和 AUC 的分解定理均涉及到了“间隔”的概念. 如前所述，间隔理论指出[8]，样本间隔的分布对集成模型的泛化误差具有重要影响. 本文中的泛化误差和 AUC 分解定理均提供了关于间隔的进一步的理论结果.

下一节中，本文将基于泛化误差和 AUC 的分解定理对间隔做进一步的分析.

## 5　泛化误差、AUC 与间隔的关系

如前所述，机器学习的目标在于最小化泛化误差. 然而由于训练数据有限，通常只能通过学习器在训练集上的经验误差来估计泛化误差. 但是，直接最小化经验误差常常会导致过拟合问题，即学习器在训练数据上的效果很好，而在新样本数据上的效果较差.

本文的理论结果为泛化误差和 AUC 的分析和优化提供了新的视角. 从前面的泛化误差和 AUC 的分解理论上看，最小化误差 $E$(或 $\mathcal{L}$)可以分解为两个任务：最小化平均误差 $\bar{E}$(或 $\bar{\mathcal{L}}$)，同时最大化分歧 $\bar{A}$(或 $\bar{A}$).

为了进一步分析影响集成模型泛化能力的因素，本节将对泛化误差、AUC 与间隔的关系做进一步的分析和讨论. 具体地，本节将证明：

集成模型的间隔越大，将使得：

(1) 集成模型的误差 $E$(或 $\mathcal{L}$)越低；

(2)基分类器的平均误差 $\bar{E}$(或 $\bar{\mathcal{L}}$)越低;

(3)基分类器之间的分歧 $\bar{A}$(或 $\bar{\mathcal{A}}$)越低.

如前所述,集成模型需要同时最小化平均误差 $\bar{E}$(或 $\bar{\mathcal{L}}$)和最大化分歧 $\bar{A}$(或 $\bar{\mathcal{A}}$).但本节的理论结果表明,过小的间隔会导致经验误差 $E$(或 $\mathcal{L}$)较高,从而使得集成模型"欠拟合";而过大的间隔会导致基分类器之间多样性降低,从而降低集成模型的泛化能力.

因此,在集成学习中,过大或过小的间隔都会影响集成模型的最终效果.基于此,本文将在第 6 节中提出两种新的基分类器权重优化算法,使集成模型在准确性和多样性之间取得平衡.

本节余下部分将以 AUC 分解定理为例,在一般情形下对上述三条结论进行定性的讨论.对于这三条结论的详细证明可参见附录 1.

第一条结论比较容易证明.因为代理函数 $l(\Delta f_{ij}^{\text{ens}})$通常为单调下降的,因此间隔越大,误差越低.

对于第二条结论,类似前面的证明,基分类器的平均误差可以展开为

$$\bar{\mathcal{L}} \propto \sum_{\alpha}\omega_{\alpha} l(\Delta f_{ij}^{\alpha}) = l(0)+l'(0)\Delta f_{ij}^{\text{ens}}+R \tag{42}$$

其中 $R=\sum_{\alpha}\omega_{\alpha}\sum_{n=2}^{+\infty}\frac{l^{(n)}(0)}{n!}(\Delta f_{ij}^{\alpha})^n$ 为泰勒展开的余项.再考虑到损失函数的单调递减性质,可得 $l'(0)<0$,因此在其它条件不变时,间隔 $\Delta f_{ij}^{\text{ens}}$ 越大,$\bar{\mathcal{L}}$ 越低.

对于第三条结论,考虑到集成模型误差可以展开为

$$l(\Delta f_{ij}^{\text{ens}}) = l(0)+l'(0)\Delta f_{ij}^{\text{ens}}+R' \tag{43}$$

其中 $R'=\sum_{n=2}^{+\infty}\frac{l^{(n)}(0)}{n!}(\Delta f_{ij}^{\text{ens}})^n$ 为泰勒展开的余项.进而可得

$$\bar{\mathcal{A}} \propto \sum_{\alpha}\omega_{\alpha} l(\Delta f_{ij}^{\alpha}) - l(\Delta f_{ij}^{\text{ens}}) = R-R' \tag{44}$$

注意到在 $\Delta f_{ij}^{\text{ens}}\to 0$ 时,$R'$ 是与 $\frac{l''(0)}{2}(\Delta f_{ij}^{\text{ens}})^2$ 同阶的,而由于损失函数通常是凸的,即 $l''\geqslant 0$,可知 $\Delta f_{ij}^{\text{ens}}$ 越大,$R'$ 越大,进而在其它条件不变时,将导致 $\bar{\mathcal{A}}$ 越低.

综上所述,过小的间隔会导致集成模型误差较高,而过大的间隔会导致集成模型多样性不足从而泛化能力较差.因此,在优化基分类器的权重时,不能简单地最大化间隔或最小化间隔,而应该在准确性和多样性之间取得平衡.

## 6 两种新的集成学习权重优化算法

基于前文的理论结果,本节将提出两种新的集成学习权重优化算法,两种算法分别以误差和 AUC 作为集成模型分类性能的评价标准和优化目标.

### 6.1 以分类误差为评价指标

上文已经证明,集成模型的间隔 $y\bar{f}$ 越大,集成模型的误差 $E$ 越低,同时基分类器之间的分歧 $\bar{A}$ 越低.

然而,直接最小化集成模型的误差 $E$ 会导致过拟合问题,因此不能简单地最大化间隔 $y\bar{f}$.避免过拟合问题的一种常用方法是引入"多样性",即同时最小化误差和最大化多样性,这就要求间隔 $y\bar{f}$ 不能过大或过小,而是处于一个合理的范围之内.

本文采用如下目标函数来平衡误差和多样性,其中 $p$ 为权衡参数,$\lambda$ 为正则化系数.需要注意的是,基于第 4 节中的理论推导,本文所提出的最优化目标不要求 $\sum\omega_{\alpha}=1$ 这一条件.

$$\begin{aligned}&\arg\min_{\boldsymbol{\omega}} E_D\{(y\bar{f}-p)^2\}+\lambda\|\boldsymbol{\omega}\|_2 \\ &\text{s.t. } \omega_{\alpha}\geqslant 0\end{aligned} \tag{45}$$

这一最优化问题还可以等价地写成下面的形式

$$\begin{aligned}&\arg\min_{\boldsymbol{\omega}} E_D\{(y\bar{f})^2-2p\cdot y\bar{f}\}+\lambda\|\boldsymbol{\omega}\|_2 \\ &\text{s.t. } \omega_{\alpha}\geqslant 0\end{aligned} \tag{46}$$

其中,二次项 $(y\bar{f})^2$ 代表了集成模型的多样性,因为其值越小,$\bar{f}$ 越接近于 0,基分类器之间的分歧也就越大.而一次项 $2p\cdot y\bar{f}$ 为间隔的倍数,由第 5 节的讨论可知,间隔越大,准确性越高,因此这一项代表了准确性.因此,从整体上看,上述最优化问题使用了参数 $p$ 来平衡准确性和多样性,$p$ 越大,则越强调准确性,$p$ 越小,则越强调多样性.

尽管上面的最优化问题中包含两个参数:$p$ 和 $\lambda$,但二者存在较强的关联.$\lambda$ 为正则化系数,对 $\omega_{\alpha}$ 的取值范围进行约束,$\lambda$ 值越小,对 $\omega_{\alpha}$ 的取值范围的约束就越松,则 $\omega_{\alpha}$ 的取值范围就越大.考虑到集成模型的输出值 $\bar{f}$ 定义为 $\bar{f}=\sum_{\alpha=1}^{T}\omega_{\alpha}f_{\alpha}$,可知随着 $\omega_{\alpha}$ 取值范围的扩大,$\bar{f}$ 的取值范围也会相应扩大,进一步可知集成模型的间隔 $y\bar{f}$ 的取值范围也会扩大,这样,为了取得准确性和多样性的平衡,$p$ 值也需要相应的提高.因此,在实际的应用中,可以将 $p$ 值固定为一个常数,然后使用网格搜索的方法确定 $\lambda$ 值.

对于参数 $p$ 的选取,本文建议将其取为基分类器数目的一半.如果所有 $T$ 个基分类器取均等权重,例如均为 1,那么 $y\bar{f}$ 的取值范围为 $[-T,T]$,$y\bar{f}=0$ 时,多样性最大,$y\bar{f}=T$ 时,准确性最高.因此,$p$ 取 $T/2$ 时,可以在多样性和准确性之间取得平衡.

为了求解上面的最优化问题，本文采用随机梯度下降的方法，每轮迭代的更新公式为

$$\omega_\alpha \leftarrow \omega_\alpha - \eta \cdot [2(y\bar{f} - p)yf_\alpha + 2\lambda\omega_\alpha)] \quad (47)$$

其中 $\eta$ 为学习速率. 为了保证 $\omega_\alpha \geq 0$ 这一条件，在迭代时如果出现 $\omega_\alpha < 0$，则将学习速率减半.

### 6.2　以 AUC 为评价指标

以 AUC 作为评价指标时，本文通过求解下面的最优化问题来确定集成权重

$$\arg\min_{\boldsymbol{\omega}} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} (\Delta f_{ij}^{\text{ens}} - p)^2 + \lambda \| \boldsymbol{\omega} \|_2 \quad (48)$$

$$\text{s.t. } \omega_\alpha \geq 0$$

对应的随机梯度下降的迭代更新公式为

$$\omega_\alpha \leftarrow \omega_\alpha - \eta \cdot [2(\Delta f_{ij}^{\text{ens}} - p)\Delta f_{ij}^{\alpha} + 2\lambda\omega_\alpha] \quad (49)$$

其中，$\Delta f_{ij}^{\text{ens}} = H(\boldsymbol{x}_i^+) - H(\boldsymbol{x}_j^-)$，$\Delta f_{ij}^{\alpha} = h_\alpha(\boldsymbol{x}_i^+) - h_\alpha(\boldsymbol{x}_j^-)$，$\eta$ 为学习速率.

以上两种权重优化算法的伪码描述如算法 1 所示.

**算法 1.**　权重优化的 MarginWeight 算法.

输入：基分类器集合 $\{h_1, h_2, \cdots, h_T\}$，学习速率 $\eta$，参数 $p$ 和 $\lambda$

输出：基分类器的权重 $\omega_1, \omega_2, \cdots, \omega_T$

```
初始化权重 ω_α ← 1.0
REPEAT
    随机抽取一个训练样本(以分类误差为目标)或随机抽取一个正样本 x_i^+ 和一个负样本 x_j^- (以AUC为目标)
    ω'_α ← 根据式(47)(或式(49))计算权重
    WHILE 存在 ω'_α < 0
        η ← η/2
        ω'_α ← 根据式(47)(或式(49))计算权重
    END
    ω_α ← ω'_α
UNTIL 迭代收敛
RETURN ω_1, ω_2, ..., ω_T
```

## 7　实验验证

### 7.1　实验设定

本文使用了 UCI 机器学习仓库[24]中的 35 个数据集对本文所提出的算法进行了验证. 限于篇幅，本文省略了这 35 个数据集的信息. 本文的实验集中于二分类问题，对于包含多个类别的数据集，本文采用了与文献[4]类似的方法对其进行了预处理，以将其转换为二分类问题.

实验中，每个数据集被均等地划分为三部分：训练集、验证集和测试集. 训练集用来训练基分类器，验证集用来确定各基分类器的权重，测试集用来评估集成模型的分类效果. 为了便于比较，与已有方法[25]类似，本文使用了 Bagging 的方法训练了 101 棵 CART 决策树[26]作为基分类器，然后使用下面的集成方法确定这 101 棵决策树的权重，并进行加权集成.

本文对比了如下集成方法：

(1) ArgMin：此方法选择经验误差最低(或 AUC 最高)的基分类器作为集成模型，即将误差最低(或 AUC 最高)的分类器权重设为 1，其它均设为 0；

(2) Bag[27]：各分类器之间以均等权重进行集成；

(3) AdaBoost[6]：以串行的方式训练一组基分类器，并根据基分类器的误差确定每个分类器的权重；

(4) MetaSVM：将各基分类器的预测值作为特征，训练一个 SVM 模型来预测最终的样本类别；

(5) MetaRidge：类似于 MetaSVM，区别是在第二层使用岭回归作为分类模型；

(6) SoftMin：一种贝叶斯模型平均方法，具体公式可参见文献[28]；

(7) ConvexDS[15]：在目标函数中综合考虑多样性和稀疏性，并通过凸优化方法进行权重优化；

(8) Agnostic：文献[28]提出的一种基于不可知贝叶斯理论的模型集成方法；

(9) MarginWeight：本文所提算法(如算法 1 所示).

本文实验中将学习速率 $\eta$ 设为 0.5. 如前所述，由于参数 $p$ 和 $\lambda$ 存在关联，因此本文将参数 $p$ 取为基分类器数目的一半，即 $p=50$，并对 $\lambda$ 进行网格搜索. 关于参数 $p$ 的不同取值对结果的影响，将在第 7.2.3 节进行讨论.

为了确定 $\lambda$ 的搜索范围，可以将式(49)改写为下面公式

$$\omega_\alpha \leftarrow (1-2\lambda\eta)\omega_\alpha - \eta \cdot [2(\Delta f_{ij}^{\text{ens}} - p)\Delta f_{ij}^{\alpha}] \quad (50)$$

可以看出，在模型训练过程中，$\lambda$ 充当了 $\omega_\alpha$ 的衰减因子，$\lambda$ 取值越大，$\omega_\alpha$ 衰减越快. 考虑到随机梯度下降算法通常需要较多的迭代次数才会收敛，因此 $\lambda$ 取值不宜过大. 因此，本文对 $\lambda$ 在集合 $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ 内进行了网格搜索，这也是已有研究中建议的正则化系数的范围[29].

上述所有算法中未提及的参数均使用网格搜索的方法来确定.

本文分别使用了分类误差和 AUC 作为各集成方法的评价指标. 在使用分类误差作为评价指标时，本文所提出的 MarginWeight 算法使用式(47)更新

权重，使用 AUC 作为评价指标时，则使用式(49)来更新权重.

本文对每个算法都运行了 30 次，然后将这 30 次所得结果进行平均，就得到每个算法的最终结果.

### 7.2 实验结果和讨论

#### 7.2.1 以分类误差为评价指标

各算法在 35 个数据集上的分类误差结果见表 1 和图 1.

**表 1 各算法在 35 个数据集上的分类误差(均值±标准差)汇总表**

| 数据集 | Bag | AdaBoost | MetaSVM | MetaRidge | SoftMin | ConvexDS | Agnostic | MarginWeight |
|---|---|---|---|---|---|---|---|---|
| abalone | .221±.013 | .233±.014● | .227±.011● | .225±.012● | .221±.013 | **.220±.013** | .240±.024● | .221±.013 |
| ad | .036±.008● | .045±.011● | .036±.008● | .030±.007● | .032±.008● | .033±.007● | .037±.009● | **.028±.006** |
| australian | .134±.021 | .158±.026● | .202±.030● | .168±.025● | .136±.022 | **.132±.022** | .158±.028● | .134±.020 |
| breast | .040±.018 | .052±.017● | .058±.017● | .082±.027● | .037±.015 | .038±.015 | .047±.015● | **.037±.012** |
| cervical_biopsy | .096±.023● | .103±.021● | .129±.024● | .126±.028● | .093±.021 | .097±.024● | .097±.021● | **.091±.019** |
| cervical_schiller | .070±.018● | .082±.017● | .096±.015● | .133±.149● | .067±.014 | .070±.018● | .068±.014 | **.067±.014** |
| connect4 | .214±.000● | .189±.000○ | **.184±.000**○ | .202±.000● | .203±.000● | .214±.000● | .290±.000● | .192±.000 |
| diabetes | .255±.032 | .259±.035 | .313±.031● | .296±.027● | .257±.028 | **.253±.031**○ | .261±.027 | .261±.030 |
| ecoli | .078±.024● | .093±.028● | .077±.025● | .131±.031● | .068±.021● | .074±.025● | .069±.026● | **.062±.022** |
| german | .255±.016 | .269±.021● | .296±.025● | .283±.026● | .258±.022 | **.253±.016** | .276±.033● | .254±.020 |
| glass | .233±.075● | .257±.093● | .173±.079● | .274±.092● | .189±.074● | .210±.074● | .198±.076● | **.150±.070** |
| haberman | .299±.064● | .304±.055● | .337±.051● | .387±.051● | .284±.061 | .301±.063● | .287±.056 | **.284±.061** |
| heart | .214±.054● | .244±.047● | .266±.046● | .305±.062● | .203±.044● | .202±.045 | .209±.052● | **.191±.043** |
| hepatitis | **.169±.052** | .202±.058● | .197±.058 | .214±.066● | .170±.055 | .173±.057 | .188±.062 | .179±.054 |
| ionosphere | .077±.029 | .129±.039● | .096±.032● | .138±.038● | .076±.027 | .074±.025 | .096±.029● | **.074±.025** |
| kr-vs-kp | .031±.008● | .049±.009● | .021±.005 | **.020±.006**○ | .030±.009● | .029±.008● | .046±.010● | .023±.007 |
| krkopt | .177±.017● | .214±.021● | .176±.019● | .169±.018 | .173±.017● | .172±.019● | .197±.027● | **.166±.020** |
| letter | .002±.002● | .004±.003● | **.000±.001** | .001±.002 | .001±.002● | .002±.002● | .006±.005● | .001±.001 |
| liver | .349±.043● | .352±.047● | .387±.053● | .395±.062● | .334±.040 | .335±.045 | .348±.047● | **.326±.054** |
| magic | .167±.018● | .190±.017● | .169±.018● | .162±.017● | .167±.016● | .164±.016● | .190±.023● | **.156±.016** |
| mnist | .014±.006● | .016±.005● | .013±.005 | .012±.005 | .015±.006● | .014±.006● | .025±.010● | **.012±.005** |
| optdigits | .032±.004● | .132±.009● | .034±.004● | .029±.004 | .031±.004● | .030±.004● | .111±.027● | **.029±.004** |
| pendigits | .002±.002 | .004±.003● | **.001±.001**○ | .004±.003● | .001±.001 | .001±.002 | .002±.002 | .001±.002 |
| poker | .152±.000● | .159±.000● | .169±.000● | .159±.000● | .152±.000● | .154±.000● | .167±.000● | **.144±.000** |
| satimage | .019±.003● | .020±.003● | .022±.003● | .018±.003 | .019±.002● | .019±.002● | .030±.008● | **.018±.003** |
| segment | .006±.006● | .106±.201● | .003±.005 | .006±.006● | .006±.007● | .006±.006● | .009±.007● | **.003±.005** |
| sonar | .240±.056● | .259±.063● | .242±.046● | .377±.094● | .234±.064 | .240±.051● | .262±.074● | **.220±.048** |
| spambase | .063±.007● | .070±.008● | .064±.007● | .062±.007● | .064±.007● | .063±.007● | .094±.020● | **.060±.007** |
| tic-tac-toe | .108±.032● | .131±.031● | .064±.023 | **.058±.014**○ | .092±.030● | .099±.031● | .165±.044● | .069±.025 |
| vehicle | .066±.022● | .074±.031● | .041±.013 | .064±.023● | .056±.020● | .054±.018● | .075±.030● | **.041±.022** |
| vote | .051±.026 | .061±.026● | .057±.020● | .098±.037● | **.044±.020** | .046±.019 | .049±.020 | .046±.019 |
| vowel | .149±.068● | .202±.062● | **.067±.041**○ | .238±.088● | .123±.055● | .126±.055● | .151±.047● | .083±.041 |
| waveform | .090±.008 | .097±.009● | .100±.009● | .094±.010● | .093±.009● | .090±.008 | .131±.022● | **.090±.009** |
| wine-red | .255±.016● | .280±.017● | .264±.016● | .256±.013● | .251±.014● | .253±.016● | .273±.023● | **.247±.016** |
| wine-white | .221±.010● | .250±.009● | .220±.009● | .219±.010● | .219±.011● | .219±.010● | .249±.023● | **.216±.008** |
| Average | 0.131 | 0.151 | 0.137 | 0.155 | 0.126 | 0.127 | 0.146 | **0.119** |

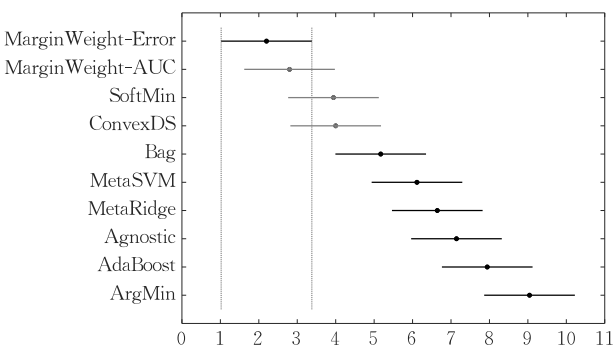



图 1 各算法分类误差的 Friedman 检验结果

表 1 给出了各算法在各数据集上的分类误差，表中，MarginWeight 算法使用式(47)进行权重的更新，即以分类误差作为优化目标. 由于空间所限，表中省略了最差的 ArgMin 算法的结果. 其中，每一个数据集上的最好结果被加粗显示. 表 1 的最后一行给出了各算法在所有数据集上的分类误差的平均值.

此外，我们还对实验结果进行了 T 检验，如果某个算法在 0.05 的显著性水平上差于本文提出的 MarginWeight 算法，则在其结果后面标注了实心圆圈(●)，反之，如果某算法显著的优于本文所提算

法，则其后面用空心圆圈(○)标注. 后面没有标注的结果则表示对应算法与本文所提算法在 0.05 的显著性水平上没有显著差异.

从表 1 可以看出，在绝大多数情况下，本文所提算法都以 0.05 的显著性水平优于其它算法. 并且，在全部 35 个数据集中，本文所提算法在其中的 23 个数据集上取得了最优的结果. 从最后一行的平均分类误差上看，本文所提的算法也取得了最好的分类效果.

为了进一步显示各集成方法的分类效果，本文还进行了 0.05 显著性水平上的 Friedman 检验. 检验结果如图 1 所示. 图中，MarginWeight-Error 表示按式(47)更新权重，而 MarginWeight-AUC 则表示用式(49)更新权重.

图 1 中，横线上的圆点表示各算法的平均排名(统计学上也称为平均秩)，例如本文提出的 MarginWeight-Error 算法，在所有参与比较的 10 个算法中，在大多数数据集上的效果都排在前 3 名，对所有 35 个数据集上的效果的排名求平均，即得到 MarginWeight-Error 算法的平均排名为 2.2. 横线表示 0.05 的显著性水平上的 Bonferroni 检验的置信区间. 如果两个算法的横线没有重合，则表示二者的分类效果存在显著的差异. 然而反过来则不一定成立，即使两个横线存在重合的部分，二者也可能存在显著的差异.

从图 1 可以看出，所有集成方法都优于 ArgMin 算法. 需要注意的是，ArgMin 算法会将效果最好的基分类器的权重设为 1，其它基分类器的权重设为 0，因此，严格地说，这并不是一种集成方法，而是单个的机器学习模型. 因此，这一结果说明了集成方法的有效性.

从图 1 中还可以看出，本文所提出的 MarginWeight 算法取得了最好的效果，且显著地优于参与比较的其它大部分算法. 同时，由于此实验中以分类误差作为评价标准，而 MarginWeight-Error 算法直接以分类误差作为目标，因此取得了比 MarginWeight-AUC 更好的结果. 此外，还需要指出的是，尽管 MarginWeight-AUC 没有以分类误差作为优化目标，但仍取得了相当好的结果，这反映了分类误差和 AUC 在反映模型泛化能力上的一致性[30]，同时也说明，本文所提出的两种方法都可以有效地降低集成模型的分类误差.

7.2.2　以 AUC 为评价指标

各算法在 35 个数据集上的 AUC 结果见表 2 和图 2. 表 2 中，MarginWeight 算法使用式(49)进行权重的更新，即以 AUC 作为优化目标. 由于空间所限，表中省略了最差的 ArgMin 算法的结果. 注意，表中只保留了小数部分，其中 .000 表示 AUC 指标值为 1.0. 表中的标注信息与表 1 类似，此处不再重复.

**表 2　各算法在 35 个数据集上的 AUC 结果(均值±标准差)汇总表**

| 数据集 | Bag | AdaBoost | MetaSVM | MetaRidge | SoftMin | ConvexDS | Agnostic | MarginWeight |
|---|---|---|---|---|---|---|---|---|
| abalone | .859±.010 | .769±.014● | .773±.011● | .850±.009● | .859±.009● | .860±.010 | .841±.012● | **.860±.010** |
| ad | .975±.011● | .877±.028● | .932±.014● | .972±.013● | .976±.011● | .975±.011● | .970±.011● | **.977±.010** |
| australian | .926±.017● | .835±.027● | .799±.035● | .895±.024● | .926±.019● | .928±.017● | .919±.020● | **.930±.017** |
| breast | .990±.007 | .951±.017● | .941±.017● | .932±.036● | **.991±.006** | .990±.007 | .990±.007 | .990±.007 |
| cervical_biopsy | .618±.068● | .527±.030● | .528±.039● | .574±.083● | .623±.071 | .618±.066● | .616±.071 | **.627±.070** |
| cervical_schiller | .627±.081 | .516±.026● | .525±.047● | .571±.096● | .629±.078 | .626±.083 | .616±.078● | **.631±.082** |
| connect4 | .819±.000● | .683±.000● | .690±.000● | .818±.000● | **.833±.000**○ | .827±.000● | .787±.000● | .829±.000 |
| diabetes | .804±.029● | .694±.035● | .660±.029● | .746±.038● | .805±.030 | .806±.029 | .794±.030● | **.806±.029** |
| ecoli | .938±.038● | .816±.082● | .839±.050● | .768±.103● | .943±.037● | .941±.036● | .938±.035● | **.946±.035** |
| german | .754±.025 | .648±.027● | .624±.030● | .709±.028● | .750±.024● | .757±.025 | .723±.025● | **.757±.023** |
| glass | .840±.065● | .705±.102● | .816±.077● | .747±.109● | .891±.046○ | .869±.052● | **.898±.042**○ | .885±.043 |
| haberman | .640±.072● | .587±.065● | .562±.060● | .576±.070● | .664±.066 | .649±.068● | .659±.069 | **.666±.062** |
| heart | .887±.040● | .772±.056● | .729±.046● | .744±.059● | .893±.034 | .891±.036● | .884±.033● | **.894±.034** |
| hepatitis | .852±.051● | .659±.099● | .701±.094● | .676±.160● | .861±.054 | .862±.055 | .851±.073● | **.865±.061** |
| ionosphere | .959±.021● | .831±.051● | .891±.034● | .902±.031● | **.966±.014** | .963±.018● | .959±.018● | .965±.016 |
| kr-vs-kp | .995±.002● | .950±.008● | .977±.005● | .997±.002 | .996±.002● | .996±.001● | .988±.005● | **.997±.001** |
| krkopt | .917±.000● | .771±.000● | .819±.000● | .913±.000● | .920±.000● | .918±.000● | .904±.000● | **.925±.000** |
| letter | .000±.000● | .996±.004● | .000±.001● | .000±.000 | .000±.000● | .000±.000 | .000±.000● | **.000±.000** |
| liver | .668±.063● | .626±.068● | .586±.058● | .632±.060● | .681±.059● | .679±.061● | .672±.053● | **.689±.061** |
| magic | .873±.000● | .776±.000● | .784±.000● | .872±.000● | .870±.000● | **.877±.000**○ | .826±.000● | .874±.000 |
| mnist | .000±.000● | .990±.000● | .995±.000● | .000±.000 | .000±.000● | .000±.000● | .999±.000● | **.000±.000** |
| optdigits | .995±.001● | .870±.011● | .966±.005● | .996±.001● | .995±.001● | .996±.001● | .974±.005● | **.996±.001** |
| pendigits | .000±.000 | .980±.091 | .998±.002● | .998±.003● | .000±.000● | .000±.000● | .999±.001● | **.000±.000** |
| poker | .739±.000● | .504±.000● | .566±.000● | .713±.000● | .735±.000● | **.752±.000**○ | .690±.000● | .743±.000 |

（续 表）

| 数据集 | Bag | AdaBoost | MetaSVM | MetaRidge | SoftMin | ConvexDS | Agnostic | MarginWeight |
|---|---|---|---|---|---|---|---|---|
| satimage | .998±.001● | .969±.008● | .969±.006● | .998±.002● | .998±.001● | .998±.001● | .997±.001● | **.998±.001** |
| segment | .000±.001 | .827±.235● | .999±.003● | .999±.003● | .000±.001 | .000±.001 | .000±.002 | **.000±.001** |
| sonar | .843±.052● | .733±.063● | .726±.061● | .657±.094● | .845±.047● | .853±.052● | .818±.052● | **.858±.049** |
| spambase | .975±.004● | .926±.008● | .931±.008● | .976±.004● | .976±.004● | .976±.004● | .966±.005● | **.977±.004** |
| tic-tac-toe | .958±.022● | .822±.037● | .930±.025● | **.984±.010** | .969±.018● | .969±.018● | .935±.023● | .982±.011 |
| vehicle | .987±.011● | .940±.025● | .967±.018● | .981±.015● | .991±.008● | .990±.009● | .986±.008● | **.993±.007** |
| vote | .990±.007● | .923±.085● | .940±.023● | .911±.041● | .991±.006 | .991±.006● | .989±.009● | **.992±.006** |
| vowel | .923±.048● | .812±.075● | .938±.037● | .799±.090● | .943±.035● | .939±.040● | .933±.038● | **.959±.031** |
| waveform | .977±.004○ | .906±.012● | .900±.009● | .972±.004● | .977±.004 | .977±.004○ | .955±.008● | **.977±.004** |
| wine-red | .827±.016● | .729±.019● | .735±.022● | .809±.020● | .826±.017● | .830±.016● | .799±.021● | **.832±.016** |
| wine-white | .834±.010● | .702±.017● | .738±.015● | .828±.011● | .833±.010● | .835±.010 | .804±.013● | **.836±.010** |
| Average | 0.885 | 0.789 | 0.813 | 0.843 | 0.890 | 0.890 | 0.877 | **0.893** |

从表 2 中可以看出，在绝大多数情况下，本文所提算法都以 0.05 的显著性水平优于其它算法. 并且，在全部 35 个数据集中，本文所提算法在其中的 28 个数据集上取得了最优的结果. 最后一行的平均值也同样表明本文所提算法的结果是最优的.

本文同样对 AUC 结果进行了 0.05 显著性水平上的 Friedman 检验. 检验结果如图 2 所示.

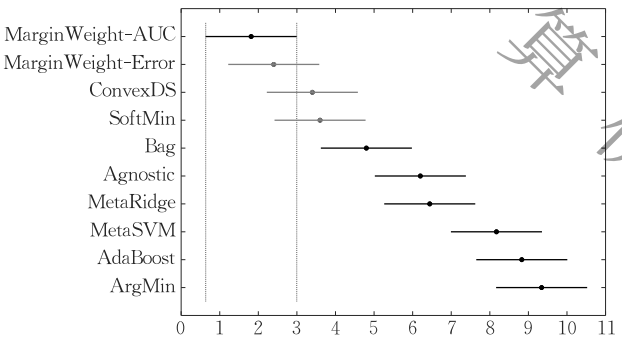



图 2 各算法 AUC 指标的 Friedman 检验结果

图 2 的结论与图 1 类似，本文所提算法同样取得了最好的效果. 此实验以 AUC 作为评价标准，而 MarginWeight-AUC 直接以 AUC 作为优化目标，因此取得了比 MarginWeight-Error 更好的效果. 同时，由于分类误差和 AUC 在衡量模型泛化能力上的一致性[30]，MarginWeight-Error 也取得了相当好的结果.

综合以上图表可以说明，与现有集成方法相比，本文所提出的权重优化算法具有更好的分类效果和泛化能力.

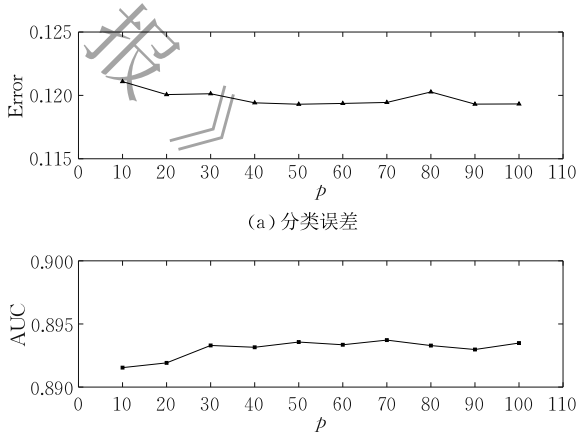
7.2.3 参数 $p$ 的影响

在算法 1 中，参数 $p$ 充当了准确性和多样性之间权衡的参数，根据其意义可知，其合理范围为 $p>0$. $p$ 值越大，表示越重视准确性，$p$ 值越小，则表示越重视多样性.

如 6.1 节所述，参数 $p$ 和 $\lambda$ 之间存在较强的关联. 在实际的应用中，可以将 $p$ 值固定为一个常数，然后使用网格搜索的方法确定 $\lambda$ 值. 在上述实验中，我们将 $p$ 固定为基分类器的一半，即 $p=50$，并对 $\lambda$ 在集合 $\{10^{-5},10^{-4},10^{-3},10^{-2},10^{-1}\}$ 内进行了网格搜索.

为了研究参数 $p$ 对算法 1 效果的影响，我们使用不同的 $p$ 值在 35 个数据集上进行了实验，对应的 $\lambda$ 值则通过网格搜索确定.

图 3 展示了取不同 $p$ 值时算法 1 在全部 35 个数据集上的平均分类误差和 AUC 结果. 从图中可以看出，$p$ 值对分类效果影响不大，尤其在 $p\geq 40$ 之后，分类效果几乎不再产生变化. 这主要是由于 $\lambda$ 是通过网格搜索自动的适应不同的数据集和不同的 $p$ 值的，因此 $p$ 值的变化没有对最终结果产生明显影响.




(a) 分类误差

(b) AUC

图 3 分类误差和 AUC 随参数 $p$ 的变化曲线

因此，本文所提算法对参数 $p$ 并不敏感，从而具有较强的应用价值.

## 8 结 论

本文从泛化误差的分解理论出发，首先给出了 AUC 的分解定理，然后基于这两种分解理论，讨论

了泛化误差、AUC、多样性与间隔之间的关系. 本文的理论结果表明，最大化间隔在降低了集成模型的经验误差的同时，也会降低集成模型的多样性，这使得通过最大化间隔方法得到的集成模型往往会过拟合.

在理论结果的基础上，本文提出了两种新的基分类器权重优化算法，来分别最优化集成模型的分类误差和 AUC. 算法通过求解一个二次优化问题，来实现准确性和多样性之间的最佳平衡. 本文所提出的算法不仅简单高效，而且在 35 个公开实验数据集上的结果表明，在 0.05 的显著性水平上，本文算法在绝大多数情况下都优于其它常用的集成方法.

## 参　考　文　献


[1] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review, 2010, 33(1): 1-39

[2] Zhou Z H. Ensemble Methods: Foundations and Algorithms. Boca Raton, USA: CRC Press, 2012

[3] Mendes-Moreira J, Soares C, Jorge A M, et al. Ensemble approaches for regression: A survey. ACM Computing Surveys, 2012, 45(1): 10

[4] Li N, Yu Y, Zhou Z H. Diversity regularized ensemble pruning. Machine Learning and Knowledge Discovery in Databases, 2012, 7523: 330-345

[5] Yang Chang-Sheng, Tao Liang, Cao Zhen-Tian, et al. Pair wise diversity measures based selective ensemble method. Pattern Recognition and Artificial Intelligence, 2010, 23(4): 565-571(in Chinese)
(杨长盛，陶亮，曹振田等. 基于成对差异性度量的选择性集成方法. 模式识别与人工智能，2010，23(4)：565-571)

[6] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1): 119-139

[7] Schapire R E, Freund Y, Bartlett P, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics, 1998, 26(5): 1651-1686

[8] Gao W, Zhou Z H. On the doubt about margin explanation of boosting. Artificial Intelligence, 2013, 203: 1-18

[9] Brown G, Kuncheva L I. "Good" and "bad" diversity in majority vote ensembles. Multiple Classifier Systems, 2010, 5997: 124-133

[10] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 1994: 231-238

[11] Ueda N, Nakano R. Generalization error of ensemble estimators//Proceedings of the IEEE International Conference on Neural Networks. Washington, USA. 1996: 90-95

[12] Brown G, Wyatt J, Harris R, et al. Diversity creation methods: A survey and categorisation. Information Fusion, 2005, 6(1): 5-20

[13] Jiang Z, Liu H, Fu B, et al. Generalized ambiguity decompositions for classification with applications in active learning and unsupervised ensemble pruning//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2073-2079

[14] Wolpert D H. Stacked generalization. Neural Networks, 1992, 5(2): 241-259

[15] Yin X C, Huang K, Yang C, et al. Convex ensemble learning with sparsity and diversity. Information Fusion, 2014, 20: 49-59

[16] Yin X C, Huang K, Hao H W, et al. A novel classifier ensemble method with sparsity and diversity. Neurocomputing, 2014, 134: 214-221

[17] Zhang Xin-Yu, Zou Guo-Hua. Model averaging method and its application in forecast. Statistical Research, 2011, 28(6): 97-102(in Chinese)
(张新雨，邹国华. 模型平均方法及其在预测中的应用. 统计研究，2011，28(6)：97-102)

[18] Valiant L G. A theory of the learnable. Communications of the ACM, 1984, 27(11): 1134-1142

[19] Collins M, Schapire R E, Singer Y. Logistic regression, AdaBoost and Bregman distances. Machine Learning, 2002, 48(1): 253-285

[20] Gao W, Jin R, Zhu S, et al. One-pass AUC optimization//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 906-914

[21] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent//Proceedings of the International Conference on Machine Learning. Bonn, Germany, 2005: 89-96

[22] Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, 4(11): 933-969

[23] Zhong W, Kwok J T. Accurate probability calibration for multiple classifiers//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 1939-1945

[24] Dua D, Karra Taniskidou E. UCI Machine Learning Repository. Irvine, USA: University of California, 2017

[25] Qian C, Yu Y, Zhou Z H. Pareto ensemble pruning//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2935-2941

[26] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees. Belmont, USA: Wadsworth International Group, 1984

[27] Breiman L. Bagging predictors. Machine learning, 1996, 24(2): 123-140

[28] Lacoste A, Marchand M, Laviolette F, et al. Agnostic Bayesian learning of ensembles//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 611-619

[29] Zhang Y, Duchi J C, Wainwright M J. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Journal of Machine Learning Research, 2015, 16(1): 3299-3340

[30] Ling C X, Huang J, Zhang H. AUC: A statistically consistent and more discriminating measure than accuracy//Proceedings of the International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003, 3: 519-524

## 附录 1.

### 1 泛化误差分解中各项与间隔的关系

在分类问题中,最常用的损失函数包括对数损失和指数损失.对于一个集成模型 $H(\boldsymbol{x})$,以这两种损失函数评价的泛化误差的形式分别为 $E_D\{\log(1+e^{-y\bar{f}})\}$ 和 $E_D\{e^{-y\bar{f}}\}$.以对数损失为例,可以得到下面结论

$$\min E \Leftrightarrow \min E_D\{\log(1+e^{-y\bar{f}})\}$$

容易看出,间隔 $y\bar{f}$ 越大,集成模型的误差 $E$ 越低.

参考文献[13]已经证明,对于二分类问题,在基分类器的输出 $f_\alpha$ 满足 $f_\alpha\in\{-1,+1\}$ 时,$E$、$\bar{E}$ 和 $\bar{A}$ 三项均只与间隔有关,并且

$$\min \bar{E} \Leftrightarrow \max E_D\{y\bar{f}\}$$

文章还证明,使用对数损失作为损失函数时,

$$\max \bar{A} \Leftrightarrow \min E_D\{\log(e^{\frac{y\bar{f}}{2}}+e^{-\frac{y\bar{f}}{2}})\}$$

使用指数损失作为损失函数时,

$$\max \bar{A} \Leftrightarrow \min E_D\left\{\frac{e-e^{-1}}{2}y\bar{f}+e^{-y\bar{f}}\right\}$$

无论是哪种情况,在 $y\bar{f}$ 过大时,都会导致分歧项 $\bar{A}$ 降低,从而使得多样性变低.

由此可见,对于二分类问题,在 $f_\alpha\in\{-1,+1\}$ 的情况下,正文第 5 节所述的三个结论均成立.

### 2 AUC 分解中各项与间隔的关系

AUC 是一种成对损失,使用 AUC 衡量集成模型性能时,$\Delta f_{ij}^{\text{ens}}=H(\boldsymbol{x}_i^+)-H(\boldsymbol{x}_j^-)$ 为集成模型的间隔.

根据定理 3 和定理 4 中 $\mathcal{L}^{\text{ens}}$ 的定义,以对数损失为例,可以得到

$$\min \mathcal{L}^{\text{ens}} \Leftrightarrow \min \frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\log(1+e^{-\Delta f_{ij}^{\text{ens}}})$$

可以看出,间隔 $\Delta f_{ij}^{\text{ens}}$ 越大,集成模型的误差 $\mathcal{L}^{\text{ens}}$ 越低.

为了说明间隔 $\Delta f_{ij}^{\text{ens}}$ 越大,基分类器的平均损失 $\bar{\mathcal{L}}$ 越低,下面给出一个关于 $\bar{\mathcal{L}}$ 的定理.

**定理 5.** 在定理 3 和定理 4 中,若损失函数 $l(\cdot)$ 为单调下降的凸函数(如对数损失和指数损失等),且基分类器的输出 $h(\boldsymbol{x})$ 满足 $h(\boldsymbol{x})\in\{0,1\}$,那么平均间隔越大,则基分类器的平均损失 $\bar{\mathcal{L}}$ 越低.

证明. 根据泰勒定理,函数 $l(\Delta f_{ij}^{\alpha})$ 可以在 0 点附近展开为

$$l(\Delta f_{ij}^{\alpha}) = l(0)+l'(0)\Delta f_{ij}^{\alpha}+\sum_{n=2}^{+\infty}\frac{l^{(n)}(0)}{n!}(\Delta f_{ij}^{\alpha})^n$$

在 $h(\boldsymbol{x})\in\{0,1\}$ 的情况下,$\Delta f_{ij}^{\alpha}\in\{-1,0,1\}$,因此有

$$(\Delta f_{ij}^{\alpha})^n=\begin{cases}\Delta f_{ij}^{\alpha}, & n=1,3,5,\cdots\\ |\Delta f_{ij}^{\alpha}|, & n=2,4,6,\cdots\end{cases}$$

对 $l(\Delta f_{ij}^{\alpha})$ 加权平均,可得

$$\sum_{\alpha}\omega_\alpha l(\Delta f_{ij}^{\alpha})=\left[l(0)+\sum_{n=2,4,6,\cdots}\frac{l^{(n)}(0)}{n!}\sum_{\alpha}\omega_\alpha|\Delta f_{ij}^{\alpha}|\right]+\left[l'(0)+\sum_{n=3,5,7,\cdots}\frac{l^{(n)}(0)}{n!}\right]\Delta f_{ij}^{\text{ens}}$$

令

$$A=\sum_{n=2,4,6,\cdots}\frac{l^{(n)}(0)}{n!},\ B=\sum_{n=3,5,7,\cdots}\frac{l^{(n)}(0)}{n!}$$

考虑到

$$l(1)=l(0)+l'(0)+A+B$$
$$l(-1)=l(0)-l'(0)+A-B$$

可知

$$A=\frac{l(1)+l(-1)}{2}-l(0)$$
$$B=\frac{l(1)-l(-1)}{2}-l'(0)$$

因此

$$\sum_{\alpha}\omega_\alpha l(\Delta f_{ij}^{\alpha})=l(0)+\left[\frac{l(1)+l(-1)}{2}-l(0)\right]\cdot\sum_{\alpha}\omega_\alpha|\Delta f_{ij}^{\alpha}|+\frac{l(1)-l(-1)}{2}\Delta f_{ij}^{\text{ens}}$$

由于 $l(\cdot)$ 为单调下降的凸函数,因此

$$\frac{l(1)+l(-1)}{2}-l(0)\geq 0,\ \frac{l(1)-l(-1)}{2}<0$$

再对所有正负样本对求和,即可得到

$$\bar{\mathcal{L}}=\sum_{\alpha}\omega_\alpha\mathcal{L}^{\alpha}\propto-\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\Delta f_{ij}^{\text{ens}}$$

注意到 $\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\Delta f_{ij}^{\text{ens}}$ 为集成模型的平均间隔,因此平均间隔越大,集成模型的平均损失 $\bar{\mathcal{L}}$ 越低. 证毕.

下面将以对数损失为例说明,间隔越大,基分类器之间的分歧 $\bar{\mathcal{A}}$ 越低.

使用对数损失函数时,

$$l(\Delta f_{ij}^{\text{ens}}) = \log(1+e^{-\Delta f_{ij}^{\text{ens}}})$$

根据定理 5 的证明,可知

$$\sum_{\alpha}\omega_\alpha l(\Delta f_{ij}^{\alpha})=-\frac{1}{2}\Delta f_{ij}^{\text{ens}}+F$$

其中 $F=l(0)+\left[\frac{l(1)+l(-1)}{2}-l(0)\right]\sum_{\alpha}\omega_\alpha|\Delta f_{ij}^{\alpha}|$.因此

$$\bar{\mathcal{A}}\propto\sum_{\alpha}\omega_\alpha l(\Delta f_{ij}^{\alpha})-l(\Delta f_{ij}^{\text{ens}})=-\log(e^{\Delta f_{ij}^{\text{ens}}/2}+e^{-\Delta f_{ij}^{\text{ens}}/2})+F$$

可以看出,间隔 $\Delta f_{ij}^{\text{ens}}$ 越大,基分类器之间的分歧 $\bar{\mathcal{A}}$ 越低.

同理可以验证类似的结论对指数损失的情况也是成立的.

**JIANG Zheng-Shen**, born in 1990, Ph. D. candidate. His research interests include machine learning and recommender systems.

**LIU Hong-Zhi**, born in 1982, Ph. D., associate professor. His research interests include recommender systems, information fusion and ensemble learning.

**FU Bin**, born in 1990, Ph. D. candidate. His research interest focuses on recommender systems.

**WU Zhong-Hai**, born in 1968, Ph. D., professor. His research interests include context aware services, cloud security and privacy protection, embedded intelligence, big data and information fusion.

**Background**

In this paper, we research on weight optimization of the base classifiers, which is a fundamental and challenging problem in the ensemble learning domain.

The main difficulty of this problem is that it is easy to over-fit. At present, the approaches that have been proposed to determine the weight of the base classifiers could be divided into two categories: Bayesian model averaging methods and optimization methods. Bayesian model averaging methods typically do not suffer from over-fitting problems. However, for a certain learning problem, it is usually difficult to determine the prior distribution. Moreover, such methods requires a lot of training data to estimate the posterior distribution. As a result, the application of such methods is limited. The optimization methods usually suffer from over-fitting problem, and a term that is related to diversity is typically used to avoid over-fitting. However, due to the unclear definition of diversity and the difficulty of parameter tuning, such methods usually could not achieve satisfactory performance.

In this paper, we first analyze the generalization error and AUC of the ensemble model, then we discuss the relationship between generalization error, AUC, diversity and margin. In order to analyze the AUC criterion, we present the AUC decomposition theorem based on the Ambiguity Decomposition, which is one of the most important generalization error decomposition theory. Based on our theoretical analysis, we point out that the margin maximization method not only reduces the empirical error, but also reduces the diversity among the base classifiers, which leads to the problem of over-fitting.

Based on the theoretical results, we propose two new weight optimization algorithms to combine the base classifiers, and the targets of the two algorithms are classification error and AUC, respectively. In both of our proposed algorithms, we make use of the margin of the ensemble model. Moreover, the objective functions are quadratic functions of the margin, thus the learning procedure can be guaranteed to be convergence. By introducing a trade-off parameter $p$, the margin is optimized to a proper level, and the optimal balance between accuracy and diversity is achieved. Since the parameter $p$ is highly related to the regularization parameter, in practice, we could fix $p$ and determine the regularization parameter using grid search. Thus there is only one parameter that needs to be tuned. And therefore the proposed algorithms are highly applicable.

The work of this paper is supported by the National Natural Science Foundation of China (61232005) and the National High Technology Research and Development Program(863 Program) of China (2015AA016009).