ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 555－563
DOI:10. 16337/j. 1004-9037. 2018. 03. 020

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax: ＋86-025-84892742

融入差异性的帕累托集成剪枝方法

魏苗苗　杭　杰

(南京邮电大学计算机学院,南京, 210003)

摘　要: 相比于集成学习,集成剪枝方法是在多个分类器中搜索最优子集从而改善分类器的泛化性能,简化集成过程。帕累托集成剪枝方法同时考虑了分类器的精准度及集成规模两个方面,并将二者均作为优化的目标。然而帕累托集成剪枝算法只考虑了基分类器的精准度与集成规模,忽视了分类器之间的差异性,从而导致了分类器之间的相似度比较大。本文提出了融入差异性的帕累托集成剪枝算法,该算法将分类器的差异性与精准度综合为第1个优化目标,将集成规模作为第2个优化目标,从而实现多目标优化。实验表明,当该改进的集成剪枝算法与帕累托集成剪枝算法在集成规模相当的前提下,由于差异性的融入该改进算法能够获得较好的性能。

关键词: 集成剪枝;差异性;帕累托

中图分类号: TP391　　**文献标志码:** A

Pareto Ensemble Pruning with Diversity

Wei Miaomiao, Hang Jie

(College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China)

Abstract: Compared with ensemble learning, the ensemble pruning is used to search for the optimal subset among multiple classifiers to improve the generalization performance of the classifier and simplify the ensemble process. In order to improve generalization performance and simplify ensemble process, ensemble pruning is used to search an optimal subset in multiple classifiers. It has attracted widespread concern, and it is significant to reduce the complex of ensemble learning. In recent years, researchers have proposed Pareto ensemble pruning (PEP) which considers both the classification performance and the number of base learners, and solves the two goals as the bi-objective optimization. However, Pareto ensemble pruning method ignores the diversity among classifiers, which would cause relatively large similarity among classifiers. In the paper, we proposed Pareto ensemble pruning with diversity (PEPD), in which diversity among classifiers is introduced into Pareto ensemble pruning method. The first goal of the proposed method is to maximize classifiers' diversity and their classification performance. The second goal is to minimize the number of base learners. The experimental results show that the PEPD method can obtain higher performance in most cases. And the enhancement is due to diversity's combination when PEPD and PEP have the similarity number of base learners. Experiments show that the PEPD method can obtain higher performance in most cases due to diversity's combination, when PEPD and PEP have the similar number of base learners.

Key words: ensemble pruning; diversity; Pareto

基金项目: 江苏省自然科学基金(BK20131378,BK20140885)资助项目;广西高校云计算与复杂系统重点实验室(15206)资助项目。
收稿日期: 2016-09-26;**修订日期:** 2016-10-14

引 言

分类与回归是监督学习研究中的基本任务。分类与回归最终的目标是在一个由各种可能的函数组成的假设空间中搜索与实际目标分类函数最接近的分类器，使该分类器尽可能精确地分类未知实例。然而，单个学习器在某些情况下的泛化性能是有限的。为了提高学习器的泛化性能，学者们提出了集成学习的思想。1997 年 Dietterich[1] 指出集成学习将会成为机器学习领域的 4 大研究方向之首。

集成学习是使用多个学习器共同决策的过程。集成学习一般可以分为两个步骤：(1) 产生多个不同的基分类器，(2) 采用某种集成策略(如投票法)来决定最终的分类结果[2]。按照基分类器之间的种类关系，可以把集成学习分为同质集成学习和异质集成学习[3]。同质集成学习是集成多个同种类的基分类器，该同种类的代表分类器可以是人工神经网络、决策树、朴素贝叶斯和 K-近邻等。而异质集成学习是集成多个各种类别的基分类器，其中代表的基分类器有叠加法[4]和元学习法[5]等。随着机器学习领域不断深入的研究，集成学习已经应用到多个领域。在传统的集成学习中，大多数的方法是先产生多个不同的弱分类器，再由所有的基分类器构建一个强分类器。尽管这种方法可以有效地提高分类器的泛化性能，但是该方法存在一些不足：集成所有的基分类器将消耗大量的时间和空间资源；预测速度也随着基分类器的增多而急剧下降。于是，Zhou 等[6]于 2002 年首次提出了“选择性集成”的概念，选择性集成的思路是选用部分基学习器集成的效果可能比集成所有的基分类器效果更优。

集成剪枝又称选择性集成、集成简化，它是在训练出所有基学习器之后，基于某种准则，选择一部分基学习器(所有基学习器的一个最优子集)进行集成，最终得到一个强分类器。集成剪枝的过程主要包括 3 步：产生不同的分类器；根据验证集选择最优的分类器子集；集成分类器子集。集成剪枝方法的异同主要取决于剪枝策略。剪枝策略可以根据分类器的不同划分为基于分类问题的剪枝策略和基于回归问题的剪枝策略。由于在回归问题中，集成剪枝问题研究的较少且效果不明显，所以本节主要讨论基于分类问题的剪枝策略。Tsoumakas[7] 等总结了集成剪枝的多种策略，文献[8-15]中将集成剪枝策略分为基于排序的剪枝，基于聚类的剪枝，基于优化的剪枝以及其他方法的剪枝。

剪枝过程中使用的剪枝策略决定了集成剪枝的方法。目前还没有确定的最佳剪枝策略方法，已有的代表性剪枝策略使用遗传算法进行剪枝[6]，采用人工免疫算法进行剪枝[16]，用聚类算法进行剪枝[9]，2015 年，Qian 等又采用帕累托(Pareto)占优的双目标优化思想进行剪枝[17]。然而采用遗传算法和人工免疫算法剪枝的复杂度相当高；利用聚类的中心参与集成剪枝的聚类算法也忽略了单个学习器有限的泛化性能；Pareto 占优的双目标优化集成剪枝忽略了分类器之间的差异性。因此，为了改进上述方法的不足，在 Pareto 集成剪枝的基础上，本文提出了融入差异性的 Pareto 集成剪枝方法。

1 融入差异性的 Pareto 集成剪枝

1.1 Pareto 集成剪枝方法

在工程与科学计算领域中，存在着许多多目标优化的问题(Multi-objective optimization problem，MOP)。多目标优化问题的有效解也称为 Pareto 最优解。集成剪枝方法有两个目标：最小化集成的基分类器的个数以及最大化集成后分类器的泛化性能。对于这两个目标，研究者首先想到的是将这两个目标通过某种数学模型整合为一个优化目标。虽然这种方法存在一定的道理，但是这种合二为一的方法也与最初的集成优化目标有所偏差。文献[17]方法将最小化基分类器的个数和最大化分类器的泛化性能作为两个单独的优化目标共同优化。

已知一个包含 N 个样本的数据集 $X=\{(x_i,y_i)\}N_{i=1}$ 和 n 个分类器 $H=\{h_\alpha\}n_{\alpha=1}$，其中 h_α 是一个由特征空间到类别的映射。$c\in\{0,1\}n$ 表示在 n 维向量中对分类器的选择情况，0，1 都以均等的概率随

机选取。若 $c_\alpha=1$ 表示第 α 个分类器被选中,反之,当 $c_\alpha=0$ 则表示第 α 个分类器没有被选中。将剪枝后的 $|c|$ ($0<|c|<n$)个分类结果放入集合 H_c。优化的双目标分别是能够代表分类器泛化误差的 $f(H_c)$ 以及选中的分类器的个数 $|c|=\sum_{\alpha=1}^{n} c_\alpha$。因此该双目标优化问题可以表示为

$$\arg \min_{c\in\{0,1\}^n} (f(H_c),|c|) \tag{1}$$

在该双目标优化中,$f(H_c)$ 表示选中的 $|c|$ 个分类器的泛化误差,$|c|$ 表示集成规模。定义 1 介绍两个目标之间的 Pareto 占优关系。

定义 1[17] Pareto 占优或 Pareto 支配:令存在一个双目标函数 $\varphi=(\varphi_1,\varphi_2)$,$C$ 表示所有解决方案向量的集合。若存在两个不相等的解决方案 $c,c'\in C$,则有

(a) c 弱占优于 c',当满足 $\varphi_1(c)\leq\varphi_1(c')$ 并且 $\varphi_2(c)\leq\varphi_2(c')$ 时,表示为 c 弱占优于 c'。

(b) c 占优于 c',当满足 c 弱占优于 c' 且同时满足 $\varphi_1(c)<\varphi_1(c')$ 或者 $\varphi_2(c)<\varphi_2(c')$,表示为 c 占优于 c'。

若在所有解决方案的集合 C 中,若没有任一个方案占优于 c,那么 c 就是 Pareto 最优的解决方案。而初始化 解决方案 c 的方法是随机产生的一个由 0,1 组成的向量,并将此向量放入候选方案集合 C 中;然后通过迭代更新 C 中的解决方案。每次迭代都将在 C 中随机挑选一个解决方案 c,给 c 一定的扰动使之生成 c';若 C 中没有任何一个方案能够占优于 c',则将 c' 加入集合 C 中,同时将集合 C 中被 c' 弱占优的方案去除。而对于每次迭代过程中生成的 c',文献中采用可变深度搜索[17](Variable depth search,VDS)的方法(见算法 2),搜索出与 c' 相距一个汉明距离的所有解决方案。在这些解决方案中若能搜索出可占优于 c' 的解决方案,则用该占优的解决方案代替 c'。将 c' 从 C 中去除。深度优先搜索的方法是有序的局部贪心搜索,每次搜索都选择局部最优解。为了避免重复搜索,算法中引入一个 L 变量,用于记录已被搜索过了的路径。最终从候选集合 C 中选出泛化误差最低的解决方案作为双目标的问题的解。

算法 1 Pareto 集成剪枝

输入:

一系列已训练的分类器 $H=\{h_\alpha\}_{\alpha=1}^{n}$,第一个目标函数 $f(H_c)$,第二个目标 $|c|$,评价准则 evaluation

输出:$\arg\min_{c\in C}\text{evaluation}(c)$

1. 令 $F(c)=(f(H_c),|\boldsymbol{c}|)$ 为待优化的双目标函数。
2. 初始化 $\boldsymbol{c}$ 为一个 n 维向量,元素由 0,1 随机组成。
3. Repeat
4. 在集合 C 中随机抽取一个方案 c
5. 以 $\frac{1}{n}$ 的概率随机翻转 c 中任一元素,生成方案 c'
6. if$\nexists z\in C$ 使得 z 占优于 c'
7. $C=(C-\{z\in C\mid c' \text{ 弱占优于 } z\})\cup\{c'\}$
8. $Q=\text{VDS}(e,c')$
9. for $q\in Q$
10. if$\nexists z\in C$ 使得 z 占优于 q
11. $C=(C-\{z\in C\mid q \text{ 弱占优于 } z\})\cup\{q\}$

算法 2 可变深度搜索(Variable depth search,VDS)

输入:一个伪布尔函数 e,一个解决方案 c

输出:Q

(1) $Q=\varnothing, L=\varnothing$

(2) 令 $N(\cdot)$ 为与解决方案 c 相距一个汉明距离的所有解决方案的集合

(3) While

$V_c=\{y\in N(c)\mid (y_k\neq c_k\Rightarrow k\in L)\}=\varnothing$

(4) 选择 使得 e 达到最小值的 $y\in V_c$

(5) $Q=Q\cup\{y\}$

(6) $L=L\cup\{k\mid y_k\neq c_k\}$

(7) $c=y$

在该双目标优化中,第 1 个目标是最大化分类准确率,第 2 个目标是最小化剪枝后分类器的个数,即集成规模。算法 1 的(6~7)和(9~11)步已经决定了搜索的路径方向,因此不会存在只优化其中一个目标的问题。VDS 只是搜索局部最优值,不影响最优解决方案的搜索方向。评价准则 evaluation 可以根据侧重点来选取,更加侧重分类准确则在剪枝后的集合 C 中根据 $f(H_c)$ 函数选取最优的解决方案,若更加侧重集成分类器的规模,则根据分类器的个数 $|c|$ 选取。采用已选取的分类结果在验证集上的泛化误差来衡量泛化性能 $f(H_c)$,$f(H_c)$ 越小代表泛化性能越高;将差异性和泛化误差都归一化到(0,1)之间,使得目标结果 $g(c)$ 满足越接近 0 泛化性能越好、差异性越大。剪枝过程完成后,采用熵度量[18]衡量分类器之间的差异性,该差异性也被归一化到(0,1)之间,且越接近 0 差异性越小。算法 1 中的(3)循环的次数设为 $|-n^2\log n-|$ [17],实验选出最优的分类器子集之后,采用多数投票的方法进行集成。

1.2 融入差异性的集成剪枝

分类器集成中的差异性学习途径通常可以分为两种,隐性差异性和显性差异性。隐性差异性是指通过不同的数据训练不同的基分类器,隐性地使得各分类器具有差异性,如 Bagging[19],Boosting[20];显性差异性是指最大化某个与差异性相关的目标函数来集成不同的分类器,如半定规划[21]。本文研究的差异性是隐性差异性与显性差异性的组合,实验前期使用 Bagging[19]训练不同的基分类器,后期使用融入差异性的目标函数对分类器进行剪枝,使之为集成提供差异较大的基分类器的集合。差异性是提高集成泛化能力的必要条件,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。Ali 等[22]指出只有当分类器集合中各个分类器具有显著的互补性,它们的集成效果才能充分体现。

Pareto 集成剪枝方法可以提高分类精确度、缩减集成规模,然而在分类器差异性这方面的工作仍是空白。集成学习需要有差异性,分类器之间的差异性可以确保分类器之间的相互独立性,若一系列分类器的集成效果突出,那么分类器之间的差异应足够包含分类的错分类型。在差异性研究的基础上,本文提出了融入差异性的帕累托集成剪枝方法(Pareto ensemble pruning with diversity, PEPD)。该方法将差异性的度量融入 Pareto 集成剪枝算法的第 1 个目标中,第 2 个目标仍是集成规模。因此,PEPD 算法可以同时优化精准度,差异性以及集成规模这 3 个目标。

差异性学习和分类器准确率度量在集成学习中有着不同目的和算法处理过程。因此,实施这些不同的学习策略算法最初是分开的、独立的。Yin 等[23]提出将差异性与稀疏性线性相加为一个优化目标的方法。而本文是在该 Pareto 占优的双目标优化的基础上,在泛化目标中线性增加了差异性度量,以增加差异性对于剪枝策略的影响。第 1 个优化目标更改为泛化误差和差异性的线性结合体:$\min_c(\mu f(H_c)+\lambda d(H_c))$,第 2 个优化目标仍是集成规模 $\min|c|$ 。那么融入差异性的帕累托集成剪枝方法的目标函数可以表示成

$$\arg\min_{c\in\{0,1\}^n}(\mu f(H_c)+\lambda d(H_c),|c|) \tag{2}$$

其中,融入差异性的集成 $d(H_c)$ 表示该 $|c|$ 个分类器之间的差异性。$\mu(0<\mu<1)$ 和 $\lambda(0<\lambda<1)$ 是调节泛化误差和差异性之间所占比重的参数,$\mu+\lambda=1$。根据数据集的不同对参数 μ、λ 进行调节。为

了减少 $f(H_c)$ 和 $d(H_c)$ 函数值本身对优化目标即线性函数的影响,实验中将 $f(H_c)$ 和 $d(H_c)$ 都进行了归一化处理。

算法 3 融入差异性的 Pareto 集成剪枝算法

输入:

一系列已训练的分类器 $H=\{h_\alpha\}n_{\alpha=1}$,第 1 个目标函数 $g(c)=\mu f(H_c)+\lambda d(H_c)$,第 2 个目标 $|c|$,评价准则 evaluation

输出:$\text{argmin}_{c\in C}\text{evaluation}(c)$

1. 令 $F(c)=(g(c),|c|)$ 为待优化的双目标函数。

2. 初始化 $\boldsymbol{c}$ 为一个 n 维向量,元素由 0,1 随机组成。

3. Repeat

4. 在集合 C 中随机抽取一个方案 c

5. 以 $\frac{1}{n}$ 的概率随机翻转 c 中任一元素,生成方案 c'

6. if $\nexists z\in C$ 使得 z 占优于 c'

7. $C=(C-\{z\in C\mid c' \text{ 弱占优于 } z\})\cup\{c'\}$

8. $Q=\text{VDS}(e,c')$

9. for $q\in Q$

10. if $\nexists z\in C$ 使得 z 占优于 q

11. $C=(C-\{z\in C\mid q \text{ 弱占优于 } z\})\cup\{q\}$

将该融入差异性的改进算法与算法 1 相比较,首先将输入的第一个优化目标改为 $g(c)=\mu f(H_c)+\lambda d(H_c)$;其次,在每一次的占优、弱占优的比较中,目标函数 $\varphi_1=\mu f(H_c)+\lambda d(H_c)$,$\varphi_2=|c|$,即在每 1 次的迭代比较中,都会将分类器子集的分类准确率,各分类器间的差异性以及集成规模做对比。若某解决方案 c 的第 1 个目标优于另一解决方案 c' 的第 1 个优化目标,而 c 的第 2 个优化目标劣于 c' 的第 2 个优化目标,即 $|c|>|c'|$。它们都不满足任一方占优或弱占优另一方的情况,那么将这两个解决方案均加入解决方案集合 C 中。直到某次迭代中存在其他方案能够占优于 c 或 c',将 c 或 c' 顶替出集合。若不存在其他解决方案可以占优 c、c',则在挑选最终解决方案时,按照 evaluation 进行筛选。evaluation 是按照相同集成规模选择具有最优第 1 目标的剪枝方案,为确保集成剪枝方法不降低集成分类精准度,因此在不同集成规模时也选择最优第 1 目标的剪枝方案。

1.3 差异性度量

单分类器之间的差异性在集成学习中起着至关重要的作用。然而到目前为止,在学术界还没有一个可被公认的差异性的定义,因此,明确定义分类器之间的差异性比较困难。Kuncheva 等对比了不同差异性度量的方法并且分析了它们与集成准确率之间的关系[24]。这些度量差异性的方法可以分为成对的差异性度量(Q 统计量,相关系数,不一致度量以及双错度量)和非成对的差异性度量(熵度量,KW 方差以及难点度量)两个类别。为了验证融入差异性的集成剪枝方法确实对集成泛化性能有提升作用,可以在差异性度量的两类方法中分别选取一个度量方法。

已知剪枝后的分类器个数 $\theta=|c|$,则剪枝后的分类器可表示为 $H_c=\{h_1,h_2,\cdots,h_\theta\}$,$N$ 表示测试集中的样本数。根据剪枝后的分类器对测试样本的预测,可构造一个矩阵 $\boldsymbol{W}=[w_{\alpha,j}]_{\theta\times N}$ 用于度量分类器之间的差异性。当 $w_{\alpha,j}=1$ 时,表示第 α 个分类器对测试集中第 j 个样本的正确预测,反之,$w_{\alpha,j}=0$ 表示错误预测。两种差异性度量的方法详述如下。

(1) 不一致度量

$$\mathrm{DIS}=\frac{2}{\theta(\theta-1)}\sum_{p=1}^{\theta-1}\sum_{q=p+1}^{\theta}\mathrm{dis}_{pq} \tag{3}$$

其中

$$\mathrm{dis}_{pq}=\frac{\sum_{j=1}^{N}w_{pj}(1-w_{qj})+\sum_{j=1}^{N}(1-w_{pj})\cdot w_{qj}}{N} \tag{4}$$

$$d(H_c)=1-\mathrm{DIS} \tag{5}$$

(2) 熵度量

$$\mathrm{ENT}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\theta-|\theta/2|}\mathrm{value} \tag{6}$$

$$\mathrm{value}=\min\left\{\sum_{\alpha=1}^{\theta}w_{\alpha j},\theta-\sum_{\alpha=1}^{\theta}w_{\alpha j}\right\} \tag{7}$$

$$d(H_c)=1-\mathrm{ENT} \tag{8}$$

其中,不一致度量为成对的差异性度量方法,熵度量为非成对的差异性度量方法。两种度量方法都将差异性 $d(H_c)$ 限定在(0,1)范围之间,$d(H_c)$ 越接近于 0 代表差异性越大,反之 $d(H_c)$ 越接近于 1 代表差异性越小,与泛化误差函数 $f(H_c)$ 恰好保持一致收敛。

2 实验评价与分析

2.1 数据集

为验证提出的融入差异性的 Pareto 剪枝方法的有效性,选择 8 个公开的 UCI 数据集进行实验。数据集的详细信息如表 1 所示。若实例个数为 r,根据数据集的大小将数据集划分为 3 种规模:小规模数据($0<r\leqslant 1\ 000$),中等规模数据($1\ 000<r\leqslant 10\ 000$)以及大规模数据($r>10\ 000$)。每个规模的数据都选取 2-3 个数据集,则数据集的大小可以从 270 涵盖到 19 020。实验将对比多个数据集在不同剪枝方法下的泛化能力以及分类器之间的差异性,如表 1 所示每个数据集对应一个编号,图 1~6 中的横坐标对应的编号也如此。

表 1 实验中的 8 个真实数据集

Tab. 1 Eight real datasets in the proposed experiments

编号	数据集	样本数	特征数
1	German number	1 000	24
2	Australian	690	14
3	Diabetes	768	8
4	Madelon	2 000	500
5	EGG	14 980	14
6	Heart scale	270	13
7	Wilt	4 839	5
8	Magic	19 020	10

2.2 评价度量

在评价模型的性能时,评价指标起着至关重要的作用。对于分类器的集成,通常的评价指标有分类精准度、错误率或差异性来衡量。本实验将泛化误差和差异性综合为一个评价度量,即

$$g(c)=\mu f(H_c)+\lambda d(H_c) \tag{9}$$

其中 $f(H_c)$ 为分类器子集的泛化误差函数,$d(H_c)$ 为熵度量(非成对的差异性度量方法之一)或者不一致度量(成对的差异性度量方法之一)。由于泛化误差函数和差异性函数都归一化到(0,1)之间,而且值越小泛化性能及差异性能越好,参数 μ,λ 满足 $\mu+\lambda=1$,可以分别定义为 0.5 和 0.5,也可以根据数据集的不同做适当的调整。第 1 个优化目标分别融入两种不同差异性度量方法(熵度量和不一致度量)作为评价准则,将该改进算法剪枝后的分类器之间的差异性与 Pareto 集成剪枝后的分类器之间的差异性做对比实验,使用熵度量方法衡量剪枝后分类器间的差异性。

2.3 实验结果及分析

2.3.1 实验结果

本实验主要对比 Bagging 集成、Pareto 集成剪枝方法以及融入差异性的 Pareto 集成剪枝方法,分别

记为 Bagg，Bagg_Prun 和 Bagg_Div_Prun。选择对数据扰动比较敏感的 k 近邻分类器[25]和 C4.5 决策树分类器[26]分类器用于实验。随机抽取数据集的 60%用于训练分类模型(Train)，余下 40%的数据中一半作为验证集(Validation)，另一半作为测试集(Test)。根据数据集的大小训练出多个基分类器，数据集较大则训练的基分类器个数较多，反之训练个数较少。将基分类器在测试集上的预测结果转化为一个预测矩阵 $\boldsymbol{P}=[p_1,p_2,\cdots,p_n]$，代表 n 个分类器对同一个测试集的不同预测结果向量，其中 $p_\alpha=\{p1_\alpha,p2_\alpha,\cdots,pT_\alpha\}$ 表示第 $\alpha(\alpha\in[1,n])$ 个分类器在 T 个测试样本中的预测结果，pt_α 表示第 α 个分类器对第 t 个测试样本的预测值，预测正确则 $pt_\alpha=1$，反之 $pt_\alpha=-1$。

2.3.2 结果分析

实验对比了集成学习和集成剪枝的预测效果，以及融入差异性的集成剪枝和未融入差异性的集成剪枝的预测效果。如表 2,3 所示，表中粗体数字代表融入差异性的集成剪枝方法得到更优的分类效果。如图 1,2 所示，集成剪枝后的预测能力比集成学习的预测能力稍强，加入差异性的集成剪枝比没有加入差异性的剪枝策略泛化性能高。如图 3,4 所示，集成学习中基分类器的规模较大，而两种集成剪枝的基分类器规模相当且都比集成学习的规模要小许多，剪枝后的泛化性能与集成学习的泛化性能相当或者稍高，表明剪枝策略在不降低泛化性能的基础上有效减少集成学习的时间和空间资源的消耗，融入差异性的集成剪枝与 Pareto 集成剪枝方法集成规模相当，但能获得更高的泛化性能(图 1,2 所示)。分别采用不一致度量(成对)和熵度量(非成对)的差异性度量方法融入目标 1 中，再综合泛化误差对剪枝模型进行筛选。最后统一使用熵度量的方法分别衡量两种剪枝方法(PEP 与 PEPD)的分类器之间的差异性，实验结果如图 5,6 所示。图 5,6 表明，融入差异性的集成剪枝策略确实能够提升分类器之间的差异性。

表 2 基于 KNN 的不同集成剪枝方法准确率对比

Tab. 2 Comparison of accuracy in different ensemble pruning methods based on KNN classifier

数据集编号	Bagg-accu	Bagg-prun-accu	Bagg-div-prun-accu
1	0.770	0.785	0.785
2	0.812	0.812	0.848
3	0.740	0.734	0.760
4	0.578	0.580	0.608
5	0.811	0.815	0.824
6	0.833	0.852	0.889
7	0.930	0.932	0.932
8	0.833	0.832	0.84

表 3 基于 C4.5 的不同集成剪枝方法准确率对比

Tab. 3 Comparison of accuracy in different ensemble pruning methods based on C4.5 classifier

数据集编号	Bagg-accu	Bagg-prun-accu	Bagg-div-prun-accu
1	0.720	0.720	0.720
2	0.855	0.870	0.877
3	0.772	0.760	0.773
4	0.745	0.750	0.758
5	0.907	0.892	0.910
6	0.778	0.789	0.833
7	0.975	0.979	0.985
8	0.878	0.871	0.875

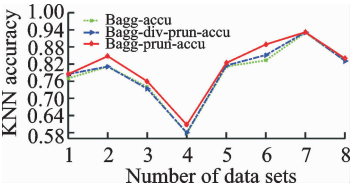

图 1 基于 KNN 的不同集成方法准确率

Fig. 1 Accuracy in different ensemble methods based on KNN classifier

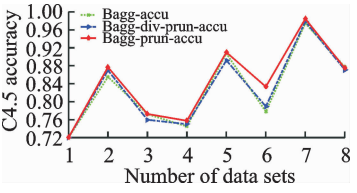

图 2 基于 C4.5 的不同集成方法准确率

Fig. 2 Accuracy in different methods based on C4.5 classifier

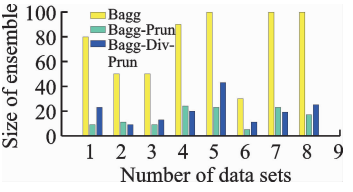

图 3 基于 KNN 的不同集成剪枝方法集成规模

Fig. 3 Ensemble size in different ensemble pruning methods based on KNN classifier

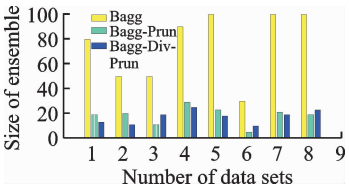

图 4 基于 C4.5 的不同集成剪枝方法集成规模

Fig. 4 Ensemble size in different ensemble pruning methods based on C4.5 classifier

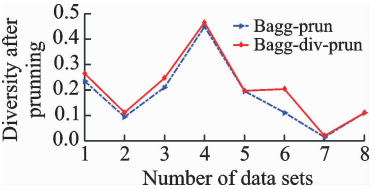

图 5 融入不一致度量后 Bagg-Div-Prun, Bagg-Prun 的差异性

Fig. 5 Diversity between Bagg-Div-Prun and Bagg-Prun with disagreement measure

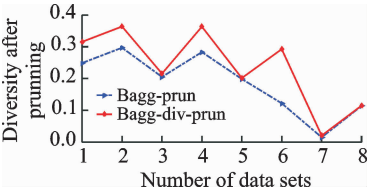

图 6 融入熵度量后 Bagg-Div-Prun, Bagg-Prun 的差异性

Fig. 6 Diversity between Bagg-Div-Prun and Bagg-Prun with entropy measure

3 结束语

集成剪枝方法一般以获得高泛化性能和低集成规模为目标,传统的集成剪枝根据某种数学变换将泛化性能和集成规模转化为一个待优化目标。该方法虽然可以改善集成学习的泛化性能,但是与集成剪枝的初衷有所偏离。目前,研究者们根据经济学原理中的 Pareto 思想提出了采用双目标优化的方法解决分类器的子集筛选问题。该方法更加直观地解决原始问题,而不是将原问题转化。本文提出的融入差异性的集成剪枝方法是基于 Pareto 集成剪枝的思想,并在原双目标的基础上增加了一个目标即分类器间的差异性。多个角度剪枝,不仅考虑了分类器的分类准确率、集成规模,还考虑了分类器之间的差异性对集成系统的影响。该过程的优势展现在:(1)融入差异性的 Pareto 集成剪枝策略确实比没有融入差异性的 Pareto 集成剪枝策略更优;(2)融入差异性的剪枝策略在泛化性能上的优势来自于差异性,具有相当规模的集成分类器个数时,融入差异性的剪枝策略比没有融入差异性的剪枝策略的泛化性能高。此外,将该融入差异性的集成剪枝策略应用于更多的分类器以及多标签问题中是一个值得研究的方向。

参考文献:

[1] Dietterich T G. Machine learning research[J]. AI Magazine, 1997, 18(4): 97.

[2] Dietterich T G. Ensemble methods in machine learning[C]// Multiple Classifier Systems. Springer Berlin Heidelberg: [s. n.], 2000: 1-15.

[3] Yu Shixin. Feature selection and classifier ensembles: A study on hyperspectral remote sensing data[D]. Antwerpen, België: Universiteit Antwerpen, 2003.

[4] Wolpert D H. Stacked generalization, neural networks[M]. Oxford, UK: Pergamon Press, 1992: 241-259.

[5] Ricardo V, Youssef D. A perspective view and survey of meta-learning[J]. Artificial Intelligence Review, 2002, 18(2): 77-95.

[6] Zhou Zhihua, Wu Jianxin, Tang Wei, et al. Ensembling neural networks: Many could be better than all[J]. Artif Intell, 2002, (137): 239-263.

[7] Tsoumakas G, Partalas I, Vlahavas I. An ensemble pruning primer[C]//Applications of Supervised and Unsupervised Ensemble Methods. Berlin, Heidelberg: Springer, 2009: 1-13.

[8] 张春霞,张讲社. 选择性集成学习算法综述[J]. 计算机学报, 2011, 34(8): 1399-1410.
Zhang Chunxia, Zhang Jiangshe. A survey of selective ensemble learning algorithms[J]. Chinese Journal of Computers, 2011, 34(8): 1399-1410.

[9] Li Kai, Huang Houkuan, Ye Xiuchen, et al. A selective approach to neural network ensemble based on clustering technology//Procof the International Conference on Machine Learning and Cybernetics. Banff, Canada: [s. n.], 2004: 3229-3233.

[10] Zhou Z H, Tang W. Selective ensemble of decision trees[C]//Rough Sets, Fuzzy Sets, Data Mining, and Granular

Computing. Berlin, Heidelberg:Springer, 2003: 476-483.
[11] Kim M J, Kang D K. Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction[J]. Expert Systems with Applications, 2012, 39(10): 9308-9314.
[12] Zhang Y, Burer S, Street W N. Ensemble pruning via semi-definite programming[J]. The Journal of Machine Learning Research, 2006, 7: 1315-1338.
[13] Partalas I, Tsoumakas G, Vlahavas I. A study on greedy algorithms for ensemble pruning[R]. Technical Report TR-LPIS-360-12, Department of Informatics. Aristotle University of Thessaloniki, Greece, 2012.
[14] Tsoumakas G, Angelis L, Vlahavas I. Selective fusion of heterogeneous classifiers[J]. Intelligent Data Analysis, 2005, 9(6): 511-525.
[15] Partalas I, Tsoumakas G, Vlahavas I. Pruning an ensemble of classifiers via reinforcement learning[J]. Neurocomputing, 2009, 72(7): 1900-1909.
[16] Castro Pablo A, Dalbem d, et al. Designing ensembles of fuzzy classification systems: An immune-inspired approach[C]// Proceedings of the 4th International Conference on Artificial Immune System. Berlin, Heidelberg:Springer, 2005:469-482.
[17] Qian C, Yu Y, Zhou Z H. Pareto ensemble pruning[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, Palo Alto, USA: AAAI Press, 2015:2935-2941.
[18] Kuncheva, L I, Christopher J W. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine Learning, 2003, 51: 181-207.
[19] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.
[20] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The Annals of Statistics, 2000, 28(2): 337-407.
[21] Zhang Y, Burer S, Street W N. Ensemble pruning via semi-definite programming[J]. The Journal of Machine Learning Research, 2006, 7: 1315-1338.
[22] Ali K M, Pazzani M J. On the link between error correlation and error reduction in decision tree ensembles[M]. Irvine: University of California, 1995.
[23] Yin X C, Huang K, Yang C, et al. Convex ensemble learning withsparsity and diversity[J]. Information Fusion, 2014, 20: 49-59.
[24] Kuncheva L I, Christopher J W. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine Learning, 2003, 51: 181-207.
[25] Liao Y, Vemuri V R. Use of k-nearest neighbor classifier for intrusion detection[J]. Computers & Security, 2002, 21(5): 439-448.
[26] Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data[J]. Remote Sensing of Environment, 1997, 61(3): 399-409.

作者简介:

魏苗苗(1991-),女,硕士研究生,研究方向:集成学习,E-mail: wmm7374@163.com。

杭杰(1994-),男,硕士研究生,研究方向:集成学习,E-mail: 1216043136@njupt.edu.cn。

(编辑:刘彦东)