

融合改进二元萤火虫算法和边界最小化测度的集成剪枝方法

朱旭辉^{1),2)} 倪志伟^{1),2)} 程美英³⁾ 李敬明⁴⁾ 金飞飞^{1),2)} 伍章俊^{1),2)}

¹⁾(合肥工业大学管理学院 合肥 230009)

²⁾(过程优化与智能决策教育部重点实验室 合肥 230009)

³⁾(湖州师范学院商学院 浙江 湖州 313000)

⁴⁾(安徽财经大学管理科学与工程学院 安徽 蚌埠 233030)

摘 要 集成剪枝是提高分类器集成性能的一种关键性技术,其通过选择较小规模的基分类器,获得更优的集成性能。目前集成剪枝方法通常单独采用基分类器间的差异性测度或元启发式算法,进行集成剪枝。基分类器的平均精度和差异性被广泛认为是集成剪枝的两个重要指标,但增大基分类器间差异性势必会减小其平均分类精度,提高基分类器的平均精度亦会降低其差异性。故在基分类器的平均精度和差异性之间存在一个平衡状态,使得集成性能最优,找到该平衡状态才是集成剪枝成功的关键。集成剪枝是一个 NP 完全问题,采用差异性测度仅能剔除集成系统中部分冗余的基分类器,难以准确地找到该平衡状态;元启发式算法在搜索该平衡状态上,具有良好的性能,但若单独采用元启发式算法,则很难穷尽搜索到该平衡状态。故该文提出了融合改进二元萤火虫算法和边界最小化测度的集成剪枝方法。首先,采用 Bootstrap 方法重复抽取训练集,获得多个训练子集,使用分类器分别进行独立训练,获得多个基分类器;其次,运用边界最小化测度对所获得的基分类器进行预剪枝,剔除综合性能较差的基分类器,显著降低集成剪枝问题的复杂度;接着,通过改进萤火虫的移动方式和搜索过程,并引入竞争行为和跳跃行为,提出了改进二元萤火虫算法;最后,利用改进二元萤火虫算法对预剪枝后的基分类器,进行二次剪枝,选择出性能最优的子集成。在 35 个 UCI 标准数据集上进行测试,实验结果表明:相较于其他方法,该文所选择的基分类器规模较小,集成分类精度更高,并验证了其有效性和显著性。

关键词 萤火虫算法;二元离散化;边界最小化;集成剪枝;差异性

中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2019.01252

Improved Binary Glowworm Swarm Optimization Combined with Margin Distance Minimization for Ensemble Pruning

ZHU Xu-Hui^{1),2)} NI Zhi-Wei^{1),2)} CHENG Mei-Ying³⁾ LI Jing-Ming⁴⁾ JIN Fei-Fei^{1),2)} WU Zhang-Jun^{1),2)}

¹⁾(School of Management, Hefei University of Technology, Hefei 230009)

²⁾(Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009)

³⁾(Business School, Huzhou University, Huzhou, Zhejiang 313000)

⁴⁾(School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui 233030)

Abstract Ensemble pruning is a key technique for a goal of achieving a better ensemble performance, using a smaller ensemble size of base classifiers, via finding a optimal sub-ensemble. Existing ensemble pruning approaches always find the optimal sub-ensemble using diversity measures among base classifiers or running heuristic searching algorithms, separately. Diversity and

收稿日期:2017-06-13;在线出版日期:2018-04-18。本课题得到国家自然科学基金(91546108,71490725)、国家重点研发计划项目(2016YFF0202604)、安徽省自然科学基金(1908085QG298,1708085MG169)、安徽省教育厅人文社科项目(JS2017AJRW0135)、湖州市科技计划自然科学基金项目(2018YZ11)、过程优化与智能决策教育部重点实验室开放课题资助。朱旭辉,博士,讲师,主要研究方向为进化计算、机器学习。E-mail: xuhuihappy@163.com。倪志伟(通信作者),博士,教授,主要研究领域为人工智能、机器学习和云计算等。E-mail: zhwnelson@163.com。程美英,博士,讲师,主要研究方向为智能计算、数据挖掘。李敬明,博士,讲师,主要研究方向为智能计算、数据挖掘。金飞飞,博士研究生,主要研究方向为智能决策、智能计算。伍章俊,博士,副教授,主要研究方向为智能算法、云计算。

accuracy of base classifiers are widely recognized as two important properties for a successful ensemble, but the increasing of the diversity of base classifiers must lead to the decrease of the average accuracy of the whole base classifiers in a constructed initial pool of classifiers, and improving the average accuracy of the whole base classifiers in the constructed initial pool of classifiers must reduce the diversity among base classifiers. Therefore, there is a tradeoff between the diversity and the accuracy of base classifiers in the constructed initial pool of classifiers, which makes the ensemble perform at its best. Finding the tradeoff between the diversity and the average accuracy of base classifiers is the key to a successful ensemble. Ensemble pruning is an NP-complete problem. Those ensemble pruning approaches based on diversity measures, using different strategies, just prune a part of redundant classifiers, and cannot exactly find the tradeoff between the diversity and the accuracy of classifiers; Those ensemble pruning methods based on heuristic searching algorithms can achieve good results when it comes to finding the tradeoff between the diversity and the accuracy, but it is hard to do an exhaustive search when the optimal sub-ensemble is achieved using heuristic algorithms. Hence, Improved Binary Glowworm Swarm Optimization combined with Margin distance minimization measure for Ensemble Pruning (IBGSOME) is proposed using a combination of the proposed improved binary glowworm swarm optimization and margin distance minimization measure. Firstly, a set of training subsets are obtained using the bootstrap sampling method, and a collection of base classifiers are achieved by independently training on the training subsets. Secondly, the generated base classifiers are pre-pruned using a margin distance minimization measure, the number of those base classifiers with worse performance is downsized significantly, which markedly reduces the computational complexity of ensemble pruning problem. Thirdly, Improved Binary Glowworm Swarm Optimization (IBGSO) is proposed by improving a moving way and searching processes of glowworms, introducing a competitive behavior and a leaping behavior. Analyzing its convergence is presented. Finally, the final ensemble is efficiently achieved from the retaining base classifiers after pre-pruning based on margin distance minimization measure using the proposed improved binary glowworm swarm optimization, and the theoretic foundation of the proposed IBGSOME is analyzed. Extensive experiments on 35 UCI machine learning benchmark datasets have been conducted. Experimental results on these UCI datasets demonstrate that the proposed IBGSOME can achieve better results than the original ensemble and other state-of-the-art ensemble pruning approaches, and the number of the members of the final ensemble achieved by the proposed IBGSOME is much less than that attained by other state-of-the-art ensemble pruning approaches. These experimental results are presented to verify the effectiveness and significance of the proposed approach.

Keywords glowworm swarm optimization; binary discretization; margin distance minimization; ensemble pruning; diversity

1 引言

近年来,集成学习已经成为一个热点问题^[1-3],被广泛应用于机器学习^[4]、模式识别^[5]和数据挖掘^[6]等领域.集成学习是先训练出多个基分类器,然后综合所有基分类器的结果,进行最终的决策^[7-8].学者们通过不断地增加基分类器的规模,以获得更

优的集成分类性能,集成系统中基分类器的差异性在提高集成分类器的泛化能力方面有重要的作用,但基分类器间差异性的增加势必会减小整体基分类器的平均精度^[9-10].差异性和平均精度之间存在一个平衡状态^[11],使得泛化能力较好. Zhou 等人^[12]通过实验研究表明,当集成系统中的基分类器间具有较大差异性时,集成部分综合性能较优的基分类器,集成分类器的泛化能力可能会比全部集成更优.即

可以通过对集成系统中的基分类器进行剪枝,以达到提高集成分类器泛化能力的目的。

集成剪枝本质上是一个组合优化问题,对于规模为 M 基分类器池,其非空子集的规模为 $2^M - 1$,故这是一个 NP 完全问题^[13-14]。为了解决该难题,众多集成剪枝方法被提出,主要被归为如下四类。

(1) 排序集成剪枝方法

该类方法先利用某种差异性测度评估各个基分类器,并计算基分类器的测度,选择大于某一阈值的基分类器进行集成。Martínez-Muñoz 等人^[14-16]先计算基分类器的差异性测度,并根据分类器的差异性大小进行有效综合集成,实验表明其集成泛化能力明显优于 Bagging; Margineantu 等人^[17]提出了 Kappa pruning (Kappa),按照基分类器的 Kappa 测度从小到大的顺序,选择集成部分 Kappa 测度较小的基分类器;Guo 等人^[18]提出了边界测度排序集成方法(MOAG),剔除部分边界测度较小的基分类器,对剩下的基分类器进行集成;Li 等人^[19]提出了差异性正则化集成剪枝方法(DREP),其给出了排序修剪在 PCA 框架下的解释,并获得了良好的效果;Dai 等人^[20]提出了反向降错排序集成剪枝方法(RRE),运用降错排序集成方法(RE)保留部分分类器,融合性能最差的分类器,取得了良好的集成分类结果;Guo 等人^[21]融合了边界和差异性测度,提出了新的组合测度,然后根据新的组合测度进行降序排列,选择部分组合测度较大的基分类器,实现集成剪枝。

(2) 优化集成剪枝方法

该类方法主要是通过使用元启发式算法,搜索到集成性能最优的子集成。Zhou 等人^[12]首次提出基于遗传算法的选择性集成方法(GASEN),通过随机赋给每个基分类器一个权重,并使用 GA 进行优化,最后选择出泛化集成误差最小的基分类器子集合,以实现集成剪枝;张琛等人^[22]提出了基于萤火虫算法的集成剪枝方法,将萤火虫算法进行二元离散化,然后采用萤火虫算法进行集成剪枝;Rokach 等人^[23]提出了集成协议集成剪枝方法(CAP),选择基分类器分类精度较高、与其他基分类器间的输出结果差异性较大的分类器。

(3) 聚类集成剪枝方法

该类方法主要是采用聚类技术^[24]进行剪枝,其主要步骤:第一,采用聚类技术将分类器归类到不同的簇,则被划分到同一簇的分类器,其差异性较小;若被划分到不同簇,则不同分类器的输出结果差别较大,即具有较大差异性。聚类技术中较为常用的

有: K -means^[25]、凝聚层次聚类方法^[26]、确定性退火方法^[27]等,上述聚类方法已经被应用于集成剪枝问题中;第二,独立剪枝每个聚类簇的基分类器,以增加基分类器间的差异性。如 Bakker 等人^[27]所选择的基分类器,通常分布在聚类簇的几何中心,并对其进行集成。

(4) 其他集成剪枝方法

该类集成剪枝方法不属于上述三类方法。Martínez-Muñoz 等人^[17]先使用 Bagging 中的 Bootstrap 抽样方法训练多个基分类器,再采用 AdaBoosting 方法进行剪枝;Lu 等人^[28]提出了基于输出不一致测度的选择性集成方法,剔除集成系统中不一致测度最大的基分类器,再采用多数投票法进行集成;An 等人^[29]提出了基于双错测度的集成方法,剔除双错测度不在统计学中单侧置信区间内的基分类器,然后进行集成;Zhou 等人^[4]提出了基于频繁模式的集成剪枝方法(EP-FP),根据基分类器的分类矩阵,挖掘其中的频繁基分类器,通过反复迭代获得最终的子集成;Cavalcanti 等人^[30]提出了组合差异性测度的集成剪枝方法(DivP),采用 GA 优化组合不同的差异性测度,运用其度量分类器的差异性,使用图着色方法选择最终的子集成;Ykhlef 等人^[31]提出了基于简单联盟博弈的集成剪枝方法(SCG-P),先通过计算班扎尼权力系数评估每个基分类器的差异性贡献,再基于最小赢得联盟获得相应的基分类器,最后进行集成;Dai 等^[32]提出了同时兼顾分类器差异性和精度的集成剪枝方法(SDAcc),同时考虑了基分类器的差异性和分类精度,然后采用贪婪集成剪枝方法(GEP)进行剪枝,并取得了良好的效果。

基于上述的分析,集成剪枝是一个 NP 难问题,而上述各类集成剪枝方法仅单独采用差异性测度或元启发式算法进行集成剪枝。单独采用差异性测度,结合不同的剪枝策略,仅能剪枝部分冗余的基分类器,难以较为精确地找到最优子集成;单独运用元启发式算法,无法在数目庞大的基分类器子集合中,进行穷尽搜索,获得性能较优的子集成。然而,元启发式算法在优化基分类器的平均精度和差异性之间的平衡状态方面,效果较为突出;但其在解决 NP 难问题时,又无法进行穷尽搜索。为了解决上述缺陷,本文在国内外集成剪枝领域内,首次提出了融合差异性测度和元启发式算法的集成剪枝方法。首先,对集成系统中的基分类器进行预剪枝,剪枝部分在性能和差异性方面表现较差的基分类器,减小集成系统中基分类器的规模,大幅降低集成剪枝问题的计算

复杂度;其次,通过改进元启发式算法,提高算法的搜索性能,对预剪枝后的基分类器进行二次剪枝。

边界最小化测度在度量分类器的差异性和预剪枝冗余分类器方面,具有明显的优势^[11,14],可以大幅度剔除性能和差异性较差的基分类器,适合作为预剪枝的标准;元启发式算法有很多,如遗传算法^[33]、粒子群算法^[34]、蚁群算法^[35]、人工鱼群算法^[36]、萤火虫算法^[37]等。遗传算法编码实现过程复杂;粒子群算法易陷入局部最优;蚁群算法搜索速度慢、计算量大、时间消耗较长;人工鱼群算法后期收敛速度慢,易陷入局部最优。基于此,考虑到萤火虫算法(Glowworm Swarm Optimization, GSO)具有易实现、鲁棒性强、搜索能力强等优势^[38-40],故选择GSO作为本文剪枝方法中的搜索策略。因此,本文采用边界最小化测度预剪枝所生成的基分类器,然后采用改进的GSO进行二次剪枝。本文提出了融合改进二元萤火虫算法和边界最小化测度的集成剪枝方法(IBGSOME)。本文先训练基分类器,构成基分类器池;再采用边界最小化测度对原始基分类器池进行预剪枝;最后采用改进二元萤火虫算法(IBGSO)进行进一步剪枝。在35个UCI数据集上的实验结果表明了本文方法的有效性、显著性和可信性。

本文的主要贡献如下:

- (1) 本文提出的融合型集成剪枝方法,显著提升了集成分类器的分类精度;边界最小化测度可以预剪枝部分综合性能较差的基分类器;
- (2) 本文提出的IBGSO,改进了萤火虫的位置更新策略,并引入竞争和跳跃行为,提高了算法在二元解空间中的收敛速度和精度;
- (3) 在35个UCI标准数据集上进行测试,实验结果表明,本文提出的IBGSOME,较其他剪枝方法,具有更优的集成性能;本文提出的IBGSO具有更优的收敛速度和精度;
- (4) 本文首次提出了基于差异性测度和二元启发式算法的组合集成剪枝的思想,为集成剪枝领域提供了一种新的研究思路。

2 边界最小化测度

差异性是提高泛化能力集成的关键因素。若集成系统中的基分类器无差异,则集成分类结果势必相同,集成学习亦无意义。文献^[14]表明了,最大化差异性并非一定可以获得最优的集成性能。文献^[11]表明了,当基分类器的平均精度一定时,增大差

异性,则系统泛化能力提高。但在基分类器的平均精度和差异性间存在一定的相关性,在增大差异性的同时,基分类器的平均精度亦会降低;增加平均精度时,差异性也会减小。因此,需要找到基分类器的平均精度和差异性间的一个平衡状态,使得集成性能最优。能否寻找存在于基分类器的平均精度和差异性之间的平衡状态?是一个未解决的难题。为了寻找该平衡状态,我们需要找到一种有效的基分类器差异性度量准则或测度,下面介绍边界最小化测度^[14],以选择具有较大差异性和较高精度的基分类器。

Martínez-Muñoz 等人在2004年提出了边界最小化测度^[15]。采用边界最小化测度选择基分类器的过程如下:

假设有 N 个样本 $X = \{x_1, x_2, \dots, x_N\}$,其实际类别 $Y = \{y_1, y_2, \dots, y_N\}$,所有样本类别 $C = \{c_1, c_2, \dots, c_l\}$, M 个基分类器集合 $F = \{f_1, f_2, \dots, f_M\}$,采用多数投票法,对 M 个基分类器集合 F 进行集成过程如式(1)。

$$E_F(x) = \arg \max_{c \in C} \left(\sum_{i=1}^M I(f_i(x) = c) \right) \quad (1)$$

其中, x 是训练或测试样本, $c(c \in C)$ 是样本类别 C 中的一类, $E_F(x)$ 是在 x 上集成分类器的输出结果, $f_i(x)$ 是 f_i 在 x 上的输出结果, $I(\cdot)$ 是一种指示函数(若“ $\cdot$ ”为真,则 $I(\cdot) = 1$;否则, $I(\cdot) = 0$), $\arg \max(f(x))$ 表示使得 $f(x)$ 最大的 x 值。

基分类器 f_i 的特征向量 $\mathbf{V}_i = (v_{i1}, v_{i2}, \dots, v_{iN})$,其中 $v_{ij} = 2I(f_i(x_j) = y_j) - 1$, $j = 1, 2, \dots, N$ 。若基分类器 f_i 在第 j 个样本上分类正确时,则 $v_{ij} = 1$;否则, $v_{ij} = -1$ 。因此,集成特征向量的均值为 $\bar{\mathbf{V}} = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_j, \dots, \bar{v}_N) = \frac{1}{M} \sum_{i=1}^M \mathbf{V}_i$ 。集成特征向量 $\bar{\mathbf{V}}$

的第 j 个元素表示第 j 个样本的边界,即 M 个基分类器中,在第 j 个样本上分类正确的基分类器数目与分类错误的数目之差。若 $\bar{v}_j > 0$ 时,则表示在第 j 个样本上的集成结果是正确的;若 $\bar{v}_j \leq 0$ 时,则表示在第 j 个样本上的集成结果是错误的。同时也说明,若 $\bar{\mathbf{V}}$ 在广义第一象限内(该象限内的向量,其所有分量均大于0),则表示在所有样本上的集成结果均正确。换言之,若集成特征向量 $\bar{\mathbf{V}}$ 与广义第一象限内的某一向量的欧式距离越近,则表示集成性能越优^[14-15]。我们取广义第一象限内的常数向量 $\mathbf{o} = (o_1, o_2, \dots, o_N)$, $o_j = \xi$, ξ 为常数(一般 $\xi = 0.075$)^[14-15]。

假设当前从集成系统中选择的子集成 S_u 的规模为 $u(1 \leq u \leq M)$,向 S_u 加入第 $u+1$ 个基分类器,

构成 S_{u+1} , 如式(2).

$$S_{u+1} = \arg \min_k \left(d \left(o, \frac{1}{M} (V_k + \sum_{i=1}^u V_i) \right) \right) \quad (2)$$

其中, $f_k \in F - S_u$, ($f_k \in F \cap f_k \notin S_u$), $d(\mathbf{x}_1, \mathbf{x}_2)$ 表示 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 之间的欧氏距离.

运用边界最小化测度所得到的基分类器子集成, 可使得基分类器具有较大的差异性和较高的分类精度. 故本文利用该测度预剪枝所生成的基分类器. 既约简了大量综合性能较差的分类器, 又保留了部分差异性和性能均较优的基分类器集合, 为基于改进二元萤火虫算法的集成剪枝方法提供了差异性大、性能优的基分类器. 上述操作明显减少了集成系统中分类器的规模, 降低了问题的复杂性, 显著提升了算法的计算效率.

3 IBGSO

3.1 GSO

GSO 是一种启发式搜索算法, 其灵感来自于萤火虫的发光行为, 发光的萤火虫是一个信号源, 可以吸引其他萤火虫^[39-40]. 在 GSO 中, 萤火虫个体被随机分布在解空间中, 荧光素高的萤火虫, 其亮度更高, 可以吸引亮度低的萤火虫向其靠拢, 从而实现算法的全局优化, 最终达到全局收敛, 基本步骤如下:

步骤 1. 初始化参数: 种群规模 $gsonum$, 荧光素挥发因子 ρ , 荧光素更新率 γ , 动态决策域更新率 β , 决策域内的萤火虫集合 $N_i(t)$, 邻域内萤火虫数目阈值 n_i , 移动步长 s , 感知半径 r_s .

步骤 2. 将萤火虫 $\mathbf{X}_i(t)$ 在第 t 次迭代的位置的目标函数值转化为荧光素值, 如式(3).

$$l_i(t) = (1 - \rho)l_i(t-1) + \gamma J(\mathbf{X}_i(t)) \quad (3)$$

步骤 3. 萤火虫 $\mathbf{X}_i(t)$ 在其动态决策半径 $r_d^i(t)$ ($0 < r_d^i(t) < r_s$) 内选择亮度比自己高的萤火虫个体组成其邻域集合 $N_i(t)$.

步骤 4. 计算萤火虫 $\mathbf{X}_i(t)$ 在其动态决策域半径内向萤火虫 $\mathbf{X}_j(t)$ 移动的概率 $p_{ij}(t)$, 如式(4).

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \quad (4)$$

步骤 5. 更新萤火虫 $\mathbf{X}_i(t)$ 的位置, 如式(5).

$$\mathbf{X}_i(t+1) = \mathbf{X}_i(t) + s \times \left[\frac{\mathbf{X}_j(t) - \mathbf{X}_i(t)}{\|\mathbf{X}_j(t) - \mathbf{X}_i(t)\|} \right] \quad (5)$$

步骤 6. 更新萤火虫 $\mathbf{X}_i(t)$ 的动态决策域半径, 如式(6).

$$r_d^i(t+1) = \min\{r_s, \max\{0, \beta \times (n_i - |N_i(t)|)\}\} \quad (6)$$

3.2 改进二元萤火虫算法 (IBGSO)

为了使 GSO 适合解决离散型问题, 对 GSO 进行二元离散化. 首先, 改进萤火虫的移动方式, 使得 GSO 可以在二元离散型空间中搜索; 其次, 改进 GSO 的搜索行为, 增加算法的随机性, 避免陷入局部最优; 最后, 引入竞争和跳跃行为, 增加种群的多样性, 提高算法的搜索效率. 基于此, 本文提出了 IBGSO.

3.2.1 移动方式的改进

基本 GSO 中固定步长不适用于二元离散型空间. 为了萤火虫个体可以在二元解空间中实现位置更新, 本文运用以一定的概率改变萤火虫个体中每一维的方式, 从而实现位置更新. 假设在第 t 次迭代中, 当前萤火虫为 $\mathbf{X}_i(t) = (x_{i1}, x_{i2}, \dots, x_{in})$, 目标萤火虫 $\mathbf{X}_j(t) = (x_{j1}, x_{j2}, \dots, x_{jn})$, 其中 n 为二元离散型解空间的维数, 则更新方式如式(7).

$$x_{ik}(t+1) = \begin{cases} x_{ik}(t), & \text{若 } r \leq p_1 \\ x_{jk}(t), & \text{若 } p_1 < r < p_2 \\ r_1, & \text{若 } r \geq p_2 \end{cases} \quad (7)$$

其中, p_1, p_2 为选择参数, $0 < p_1 < p_2 < 1$, r 为 $(0, 1)$ 间的随机数, r_1 为 0 或 1 的随机数, $k=1, 2, \dots, n$.

3.2.2 搜索行为的改进

为了增加算法在搜索过程的随机性, 避免算法陷入局部最优, 本文通过引入概率因子 τ_1, τ_2 ($0 < \tau_1 < \tau_2 < 1$), 进而实现 GSO 搜索过程的改进. 当 $rand \leq \tau_1$ 时, 当前萤火虫 $\mathbf{X}_i(t)$ 在其决策域内随机移动一步; 当 $\tau_1 < rand \leq \tau_2$ 时, 向决策域内的目标萤火虫, 按照式(7)移动一步; 当 $rand > \tau_2$ 时, 向全局最优位置, 按照式(7)移动一步.

3.2.3 竞争行为

为了增加 GSO 种群的多样性, 提高算法的搜索效率, 故将竞争行为引入到 GSO 中. 在此引入竞争因子 $comp$ ($0 < comp < 1$), 当 $rand < comp$ 时, 随机选择种群中的某一的萤火虫 $\mathbf{X}_j(t)$, 进行交叉操作, 以产生新的萤火虫个体, 从而保持种群的多样性. 当 $rand < comp$ 时, 竞争行为的执行过程如式(8)和(9).

$$\begin{cases} \mathbf{X}'_i(t) = (x_{i1}, x_{i2}, \dots, x_{i(p-1)}, x_{jp}, \dots, x_{jn}) \\ \mathbf{X}'_j(t) = (x_{j1}, x_{j2}, \dots, x_{j(p-1)}, x_{ip}, \dots, x_{in}) \end{cases} \quad (8)$$

$$\begin{cases} \mathbf{X}_i(t) = \mathbf{X}'_i(t), & \text{若 } Y_i(t) < Y'_i(t) \\ \mathbf{X}_j(t) = \mathbf{X}'_j(t), & \text{若 } Y_j(t) < Y'_j(t) \end{cases} \quad (9)$$

其中, $\mathbf{X}'_i(t), \mathbf{X}'_j(t)$ 为新产生的萤火虫个体, $Y_i(t), Y_j(t), Y'_i(t), Y'_j(t)$ 分别为 $\mathbf{X}_i(t), \mathbf{X}_j(t), \mathbf{X}'_i(t), \mathbf{X}'_j(t)$ 的适应度函数值, $p = \text{round}(rand \times n)$.

3.2.4 跳跃行为

本文将跳跃行为应用到 GSO 中, 以增加算法搜索过程中的种群多样性, 在搜索过程中萤火虫可以随机迁移至解空间中任一位置. 在每次迭代前, 若当前萤火虫与全局最优位置相同, 则迁移至解空间中任一位置; 否则, 不执行.

3.2.5 公告板

在算法中增设公告板, 记录萤火虫搜索过程中的最优位置和目标函数值. 在迭代过程中, 若萤火虫的位置优于公告板, 则更新公告板.

3.3 IBGSO 伪代码

算法 1. IBGSO.

输入: IBGSO 参数

输出: 全局最优值 $f(\mathbf{X}_{opt})$

```

1. 初始化 IBGSO 参数;
2. 随机初始化  $gsonum$  只萤火虫;
3.  $\mathbf{X}_{opt} \leftarrow \arg \max fitness(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{gsonum})$ ;
4.  $t \leftarrow 1$ ;
5. WHILE  $t < t_{max}$  DO
6.   FOR  $i \leftarrow 1$  to  $gsonum$  DO
7.     IF  $f(\mathbf{X}_i) = f(\mathbf{X}_{opt})$  THEN
8.       执行跳跃行为;
9.     END IF
10.    更新荧光素值  $l_i$  和动态决策域半径  $r_d^i$ ;
11.     $l_{max} \leftarrow l_i$ ;  $\mathbf{X}_{max} \leftarrow \mathbf{X}_i$ ;
12.    FOR  $j \leftarrow 1$  to  $gsonum$  DO
13.      IF  $\|\mathbf{X}_i - \mathbf{X}_j\| < r_d^i \wedge l_i < l_j$  THEN
14.         $l_{max} \leftarrow l_j$ ;  $\mathbf{X}_{max} \leftarrow \mathbf{X}_j$ ;
15.      END IF
16.    END FOR
17.    IF  $rand \leq \tau_1$  THEN
18.      在其决策域内随机移动一步;
19.    END IF
20.    IF  $\tau_1 < rand \leq \tau_2$  THEN
21.      向萤火虫  $\mathbf{X}_{max}$  移动一步;
22.    END IF
23.    IF  $rand > \tau_2$  THEN
24.      向萤火虫  $\mathbf{X}_{opt}$  移动一步;
25.    END IF
26.    IF  $rand < comp$  THEN
27.      执行竞争行为;
28.    END IF
29.    IF  $f(\mathbf{X}_i) > f(\mathbf{X}_{opt})$  THEN
30.       $\mathbf{X}_{opt} \leftarrow \mathbf{X}_i$ ;  $f(\mathbf{X}_{opt}) \leftarrow f(\mathbf{X}_i)$ ;
31.    END IF
32.  END FOR
33. RETURN  $f(\mathbf{X}_{opt})$ .
```

3.4 算法性能分析

为了增强 IBGSO 的理论基础, 本文分析了 IBGSO 的收敛性、时间和空间复杂度, 从理论上分析了算法的可行性和有效性.

3.4.1 收敛性分析

建立吸收态 Markov 过程模型以分析算法的全局收敛性, 本文通过萤火虫算法的改进搜索行为、竞争行为和跳跃行为, 不断地更新萤火虫个体的位置, 以便最终达到全局最优. 假设 IBGSO 中第 t 代种群 $\mathbf{X}(t)$, 则 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 为 IBGSO 的种群序列.

定义 1. 假设 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 对任意 n 个不同的 $t_1, t_2, \dots, t_n \in T$ 且 $t_1 < t_2 < \dots < t_{n-1}$, 满足下面条件 $P(\mathbf{X}(t_n) \leq \mathbf{X}_n | \mathbf{X}(t_{n-1}) = \mathbf{X}_{n-1}, \dots, \mathbf{X}(t_1) = \mathbf{X}_1) = P(\mathbf{X}(t_n) \leq \mathbf{X}_n | \mathbf{X}(t_{n-1}) = \mathbf{X}_{n-1})$, 则 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 是 Markov 过程.

定义 2. 假设 Markov 过程 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$, 其最优状态空间 $\mathbf{Y}^* \subset \mathbf{Y}$, 如果满足 $P(\mathbf{X}(t+1) \notin \mathbf{Y}^* | \mathbf{X}(t) \in \mathbf{Y}^*) = 0$, 则 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 是吸收态 Markov 过程.

定义 3. 假设吸收态 Markov 过程 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$, 其最优状态空间 $\mathbf{Y}^* \subset \mathbf{Y}$, 记 $\delta(t) = p(\mathbf{X}(t) \in \mathbf{Y}^*)$ 是在时刻 t 达到最优的概率, 若 $\lim_{t \rightarrow \infty} \delta(t) = 1$, 则 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 收敛.

引理 1. 假设 IBGSO 序列对应的随机过程为 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$, 则 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 为有限 Markov 链, 同时也是吸收态 Markov 过程.

证明. IBGSO 的搜索空间为二元离散型解空间, 每只萤火虫均为 n 维二进制码串组成的向量, 对于一个 n 维二进制解空间, 共有 $2^n - 1$ 个非零解. 因此, IBGSO 的解空间有限. 由 IBGSO 寻优过程可知, 第 t 代萤火虫 $\mathbf{X}(t)$ 的荧光素、位置、动态决策域的更新, 仅由第 $t-1$ 代萤火虫 $\mathbf{X}(t-1)$ 决定的, 与 $\mathbf{X}(t-2), \dots, \mathbf{X}(1), \mathbf{X}(0)$ 无关, 而 $\mathbf{X}(0)$ 初始化时, 其位置是随机初始化的. 故 $P(\mathbf{X}(t) | \mathbf{X}(t-1), \mathbf{X}(t-2), \dots, \mathbf{X}(0)) = P(\mathbf{X}(t) | \mathbf{X}(t-1))$, 即 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 为有限 Markov 链.

IBGSO 寻优过程包括: 萤火虫改进的搜索行为、竞争行为和跳跃行为, 且在算法中增设了公告板, 可以记录全局最优值. 若 IBGSO 搜索到的解优于全局最优解, 则更新公告板. 因此, 公告板中的最优解保持更新或不变. 基于此, 若 $\mathbf{X}(t) \in \mathbf{Y}^*$, 则 $\mathbf{X}(t+1) \in \mathbf{Y}^*$, 故而, $P(\mathbf{X}(t+1) \notin \mathbf{Y}^* | \mathbf{X}(t) \in \mathbf{Y}^*) = 0$, 即 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 是吸收态 Markov 过程. 证毕.

定理 1. 给定 IBGSO 与吸收态的 Markov

过程 $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 相对应, 其最优状态空间 $\mathbf{Z}^*$, 若 $P(\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset | \mathbf{X}(t-1) \cap \mathbf{Z}^* = \emptyset) \geq d(t-1) > 0$ 且 $\lim_{t \rightarrow \infty} \prod_{i=0}^t [1-d(i)] = 0$, 则 $\lim_{t \rightarrow \infty} \delta(t) = 1$, 其中 $\delta(t) = P(\mathbf{X}(t) \in \mathbf{Z}^*)$.

证明. 根据全概率公式知,
 $\delta(t) = [1-\delta(t-1)]P(\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset |$
 $\mathbf{X}(t-1) \cap \mathbf{Z}^* = \emptyset) + \delta(t-1)P(\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset |$
 $\mathbf{X}(t-1) \cap \mathbf{Z}^* \neq \emptyset).$

由引理 1 知, $\{\mathbf{X}(t)\}_{t=0}^{\infty}$ 是吸收态 Markov 过程, 则 $P(\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset | \mathbf{X}(t-1) \cap \mathbf{Z}^* \neq \emptyset) = 1$, 即有

$$\begin{aligned} \delta(t) &= [1-\delta(t-1)]P\{\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset | \\ &\quad \mathbf{X}(t-1) \cap \mathbf{Z}^* = \emptyset\} + \delta(t-1) \\ &\Rightarrow 1-\delta(t) = [1-P(\mathbf{X}(t) \cap \mathbf{Z}^* \neq \emptyset | \\ &\quad \mathbf{X}(t-1) \cap \mathbf{Z}^* \neq \emptyset)] \times [1-\delta(t-1)] \\ &\Rightarrow 1-\delta(t) \leq [1-d(t-1)] \cdot [1-\delta(t-1)] \\ &= [1-\delta(0)] \prod_{i=0}^{t-1} [1-d(i)] \\ &\Rightarrow \lim_{t \rightarrow \infty} \delta(t) \geq 1 - [1-\delta(0)] \lim_{t \rightarrow \infty} \prod_{i=0}^{t-1} [1-d(i)] \\ &\Rightarrow \lim_{t \rightarrow \infty} \delta(t) \geq 1. \end{aligned}$$

因为概率 $\delta(t) \leq 1$, 所以 $\lim_{t \rightarrow \infty} \delta(t) = 1$. 即 IBGSO 以概率 1 收敛到全局最优解. 证毕.

3.4.2 时间复杂度

假设 IBGSO 中种群规模是 $gsonum$, 萤火虫空间维数是 n , 最大迭代次数是 $t_{\max}$, IBGSO 的时间复杂度分析如下:

初始化算法参数, 复杂度是 $O(gsonum)$; 初始化算法种群, 包括 $gsonum$ 只萤火虫, 并计算其适应度函数值, 复杂度是 $O(gsonum \times n)$; 判断并执行跳跃行为, 复杂度最高为 $O(gsonum \times n)$; $gsonum$ 只萤火虫荧光素值、动态决策域半径更新, 并选择目标萤火虫的复杂度为 $O(gsonum^2)$; 选择目标萤火虫并移动, 复杂度均为 $O(gsonum \times n)$; 执行竞争行为的复杂度为 $O(gsonum \times n)$; 更新公告板的复杂度为 $O(gsonum)$.

通过上述分析: 在 $t_{\max}$ 次迭代后, IBGSO 的整体时间复杂度为 $O(t_{\max} \times gsonum^2)$.

3.4.3 空间复杂度

存储 GSO 中萤火虫的每个参数所需空间均为 $gsonum$; 存储每一只萤火虫所需空间为 n , 则存储 $gsonum$ 只萤火虫所需空间为 $gsonum \times n$; 公告板存储所需空间为 n ; 其他参数的存储所需空间均为常数; 经过上述分析, 算法搜索过程中的空间复杂度为 $O(gsonum \times n)$.

4 IBGSOME

4.1 基分类器池

采用 Bootstrap 方法对训练集进行抽取, 反复抽取 M 次, 可以获得 M 个训练子集, 运用基分类器对所获得的 M 个训练子集进行独立训练, 则所获得基分类器的数目为 M . 基于此, 本文使用所获得的基分类器构建基分类器池.

4.2 预剪枝

在构建基分类器池后, 集成系统中基分类器的规模为 M , 其有 $2^M - 1$ 个非空子集, 此为 NP 难问题. 当基分类器池的规模较大时, 则采用二元启发式算法无法进行穷尽搜索. 因此, 预剪枝集成系统中部分冗余基分类器是非常重要的, 剔除集成系统中大量综合性能较差的基分类器, 可以显著降低基分类器的规模, 提高 IBGSO 在搜索最优子集成过程中的计算效率. 通过运用边界最小化测度, 预剪枝基分类器池中部分基分类器, 以达到提高二元启发式算法搜索效率的目的. 本文采用边界最小化测度保留 M' 个基分类器, 至于 M' 的确定, 将在第 5 节的 5.2 小节中进行讨论. 采用边界最小化测度对基分类器池进行预剪枝后, 本文运用 IBGSO 进行进一步精确剪枝, 搜索出集成分类精度最优的基分类器子集合.

4.3 剪 枝

原始基分类器池, 在预剪枝后, 保留了 M' 个基分类器. 再采用 IBGSO 对其进行二次剪枝, 选择出集成性能更优的基分类器子集成.

4.3.1 适应度函数的构造

为了计算基分类器子集成的适应度函数值, 将式(10)^[12,22] 作为集成剪枝问题的适应度函数.

$$Fitness = A \quad (10)$$

其中, A 是集成精度, $A = \frac{1}{m} \sum_{j=1}^m Acc(\hat{y}_j, y_j)$,

$$Acc(\hat{y}_j, y_j) = \begin{cases} 1, & \hat{y}_j = y_j \\ 0, & \hat{y}_j \neq y_j \end{cases}, m \text{ 是测试样本的规模,}$$

$\hat{y}_j$ 和 y_j 是在第 j 个样本上, 集成分类器的输出类别和原始类别. $Fitness$ 越高, 则集成分类器的集成性能越优.

4.3.2 编码方式

基分类器池的规模为 M , 即 $F = \{f_1, f_2, \dots, f_M\}$, 采用边界最小化测度进行预剪枝后, 保留了 M' 个综合性能较优的基分类器, 记为 $F' = \{f'_1, f'_2, \dots, f'_{M'}\}$.

基分类器是否选取分别采用 1 或 0 表示. 萤火虫 $\mathbf{X}=(x_1, x_2, \dots, x_{M'})$, $x_i=1 (i=1, 2, \dots, M')$ 表示选择第 i 个基分类器; $x_i=0$ 表示不选择第 i 个基分类器. 例如: $\mathbf{X}=(1, 0, 1, 1, 0, 0)$ 表示选择了第 1、3、4 个基分类器, 不选择第 2、5、6 个基分类器.

4.3.3 不可行解处理方式

在算法搜索基分类器最优子集成的过程中, 若萤火虫的所有分量均为 1 或 0, 表示全部基分类器均被选择或不选择. 本文认为其为不可行解, 则该不可行解被重新初始化.

4.4 理论分析

下面将从理论的角度分析本文方法的有效性.

定理 2. 在基分类器池中, 集成边界最小化测度较小的前 M' ($M' \leq M$) 个基分类器, 可以提高集成性能.

证明. 本证明过程中基本符号说明如下:

N : 样本数目;

M : 基 ELM 数目;

c : 样本的类别, $c=1, 2, \dots, C$;

$\mathbf{V}_i$: 基分类器 f_i 的特征向量, $\mathbf{V}_i=(v_{i1}, v_{i2}, \dots, v_{iN})$, $v_{ij}=2 \times I(f_i(x_j)=y_j)-1, j=1, 2, \dots, N$.

$\bar{\mathbf{V}}$: 集成特征向量的均值, $\bar{\mathbf{V}}=(\bar{v}_1, \bar{v}_2, \dots,$

$$\bar{v}_j, \dots, \bar{v}_N)=\frac{1}{M} \sum_{i=1}^M \mathbf{V}_i;$$

$\mathbf{o}$: 广义第一象限内的某一常向量, $\mathbf{o}=(o_1, o_2, \dots, o_N)$, $o_j=\xi, \xi$ 为常数.

通过 Bootstrap 抽样方法训练出 M 个基分类器集合 $F=\{f_1, f_2, \dots, f_M\}$.

当基分类器 f_i 在第 j 个样本上分类正确时, 则 $I(f_i(x_j)=y_j)=1$, 故基分类器 f_i 的特征向量 $\mathbf{V}_i$ 的第 j 个分量 $v_{ij}=2 \times I(f_i(x_j)=y_j)-1=1$;

当基分类器 f_i 在第 j 个样本上分类错误时, 则 $I(f_i(x_j)=y_j)=0$, 故基分类器 f_i 的特征向量 $\mathbf{V}_i$ 的第 j 个分量 $v_{ij}=2 \times I(f_i(x_j)=y_j)-1=-1$.

基分类器池中 M 个基分类器集成特征向量均值为 $\bar{\mathbf{V}}=(\bar{v}_1, \bar{v}_2, \dots, \bar{v}_j, \dots, \bar{v}_N)=\frac{1}{M} \sum_{i=1}^M \mathbf{V}_i$.

若集成特征向量 $\bar{\mathbf{V}}$ 的第 j 个分量 $\bar{v}_j=\frac{1}{M} \sum_{i=1}^M v_{ij} > 0$, 则表示基分类器在第 j 个样本上, 分类正确的数目大于分类错误的数目. 根据多数投票算法 (参照式(1)) 可得, 在第 j 个样本上的集成分类结果是正确的; 若集成特征向量 $\bar{\mathbf{V}}$ 的第 j 个分量 $\bar{v}_j \leq 0$, 则表示在第 j 个样本上分类正确的基分类器数目, 小于或等于分类错误的基分类器数目. 根据多数投票算

法可得, 在第 j 个样本上集成分类结果是错误的.

若 $\bar{\mathbf{V}} > \mathbf{0}$, 则表示在所有样本上的集成分类结果全部正确, 即集成特征向量 $\bar{\mathbf{V}}$ 落入广义第一象限内. M 个基分类器的集成性能越高, 则其在 N 个样本上, 集成分类正确的样本个数就越多. 即表示其集成特征向量 $\bar{\mathbf{V}}$ 中大于 0 的分量的个数越多, 其距离广义第一象限就越近. 因此, 基分类器的集成性能与平均特征向量和广义第一象限的距离成正相关.

为了便于计算, 从广义第一象限内的某一常向量 $\mathbf{o}=(o_1, o_2, \dots, o_N)$, $o_j=\xi, \xi$ 为常数. 本文通过计算平均特征向量 $\bar{\mathbf{V}}$ 和 $\mathbf{o}$ 之间的欧式距离 $d(\bar{\mathbf{V}}, \mathbf{o})$, 衡量基分类器的集成性能. 因此, $d(\bar{\mathbf{V}}, \mathbf{o})$ 越小, 则集成性能越优.

首先从基分类器集合 F 中, 选择一个性能最优的基分类器; 其次按照式(2)逐个加入基分类器, 使得平均特征向量 $\bar{\mathbf{V}}$ 和 $\mathbf{o}$ 之间的距离 $d(\bar{\mathbf{V}}, \mathbf{o})$ 达到最小. 如此循环, 会使得原基分类器集合 $F=\{f_1, f_2, \dots, f_M\}$, 形成一个新的顺序 $f'_1, f'_2, \dots, f'_M$. 假设选择前 M' ($M' \leq M$) 个基分类器 $f'_1, f'_2, \dots, f'_{M'}$, 使得 $d(\bar{\mathbf{V}}, \mathbf{o})$ 最小. 即其集成性能最优. 因此, 集成边界最小化测度最小的前 M' 个基分类器, 可以提高集成性能. 证毕.

推论 1. 采用 IBGSO, 对预剪枝后的基分类器集合 $\{f'_1, f'_2, \dots, f'_{M'}\}$ 进行二次剪枝后, 其集成性能可以进一步提高.

证明. 原始基分类器池 $F=\{f_1, f_2, \dots, f_M\}$ 在经过预剪枝后, 得到基分类器集合 $\{f'_1, f'_2, \dots, f'_{M'}\}$. 由定理 2 知, 其预剪枝后的集成性能优于原始集成性能.

然而, 预剪枝后基分类器集合 $\{f'_1, f'_2, \dots, f'_{M'}\}$ 有 $2^{M'}-1$ 个非空子集合. 基分类器集合 $\{f'_1, f'_2, \dots, f'_{M'}\}$ 的集成性能是否优于其 $2^{M'}-1$ 个非空子集的集成性能? 换言之, 是否存在集成性能更优的基分类器子集合? 为了解决这个难题, 本文采用 IBGSO 对基分类器集合 $\{f'_1, f'_2, \dots, f'_{M'}\}$ 进行进一步剪枝.

在 IBGSO 中, 我们设置了公告板, 记录了集成性能最优的基分类器集合. 通过采用 IBGSO 的搜索机制, 不断地搜索基分类器子集合, 其搜索过程见 4.5 小节. 若出现集成性能更优的子集合, 则更新公告板; 否则, 不更新. 因此, 可以获得集成性能更优的基分类器子集合. 即二次剪枝后的集成性能, 亦可以提高. 证毕.

综上所述, 由定理 2 和推论 1 可知, IBGSOME 在进行预剪枝和二次剪枝后, 可以显著提高分类器的集成性能.

4.5 IBGSOMEF 伪代码

算法 2. IBGSOMEF.

输入: 训练集, 测试集, IBGSOMEF 参数

输出: 集成剪枝结果 $f(\mathbf{X}_{opt})$

```
1. 构建原始基分类器池  $F \leftarrow \{f_1, f_2, \dots, f_M\}$ ;
2.  $F' \leftarrow \emptyset$ ;
3.  $f'_1 \leftarrow \arg \max_{acc} (f_1, f_2, \dots, f_M)$ ;
4.  $F' \leftarrow \{f'_1\}$ ;
5.  $F \leftarrow F - \{f'_1\}$ ;
6. FOR  $i \leftarrow 2$  TO  $M'$  DO
7.   从  $F$  中按照式(2)选取一个基分类器  $f'_i$ ;
8.    $F' \leftarrow F' \cup \{f'_i\}$ ;
9.    $F \leftarrow F - \{f'_i\}$ ;
10. END FOR
11. 随机初始化  $gsonum$  只萤火虫;
12.  $\mathbf{X}_{opt} \leftarrow \arg \max_{fitness} (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{gsonum})$ ;
13.  $t \leftarrow 1$ ;
14. WHILE  $t < t_{max}$  DO
15.   FOR  $i \leftarrow 1$  TO  $gsonum$  DO
16.     IF  $f(\mathbf{X}_i) = f(\mathbf{X}_{opt})$  THEN
17.       执行跳跃行为;
18.     END IF
19.   更新荧光素值  $l_i$  和动态决策域半径  $r_d^i$ ;
20.    $l_{max} \leftarrow l_i$ ;  $\mathbf{X}_{max} \leftarrow \mathbf{X}_i$ ;
21.   FOR  $j \leftarrow 1$  TO  $gsonum$  DO
22.     IF  $\|\mathbf{X}_i - \mathbf{X}_j\| < r_d^i \wedge l_i < l_j$  THEN
23.        $l_{max} \leftarrow l_j$ ;  $\mathbf{X}_{max} \leftarrow \mathbf{X}_j$ ;
24.     END IF
25.   END FOR
26.   IF  $rand \leq \tau_1$  THEN
27.     在其决策域内随机移动一步;
28.   END IF
29.   IF  $\tau_1 < rand \leq \tau_2$  THEN
30.     向萤火虫  $\mathbf{X}_{max}$  移动一步;
31.   END IF
32.   IF  $rand > \tau_2$  THEN
33.     向萤火虫  $\mathbf{X}_{opt}$  移动一步;
34.   END IF
35.   IF  $rand < comp$  THEN
36.     执行竞争行为;
37.   END IF
38.   IF  $sum(\mathbf{X}_i) = M' \vee sum(\mathbf{X}_i) = 0$  THEN
39.     重新初始化  $\mathbf{X}_i$ ;
40.   END IF
41.   IF  $f(\mathbf{X}_i) = f(\mathbf{X}_{opt})$  THEN
42.      $\mathbf{X}_{opt} \leftarrow \mathbf{X}_i$ ;
43.   END IF
44. END WHILE
45. RETURN  $f(\mathbf{X}_{opt})$ .
```

5 实验结果及分析

为了检验 IBGSOMEF 的一般性,在 35 个 UCI 标准数据集上进行测试,如表 1. 本文选择的基分类器是分类回归树(CART)^[11]. 结合 k -折交叉验证技术^[41],其中 $k=5$,原始数据集被随机划分为 5 份,训练集选择其中 4 份,测试集选择剩下 1 份.

5.1 实验环境及参数设置

本文运用 Matlab R2012a 编写实验中所涉及的算法代码,PC 机配置:Windows 7 操作系统、英特尔奔腾 E7500 处理器、CPU 主频为 2.93 GHz、内存 4 GB DDR3. 根据 Krishanand 和 Ghose^[42-44] 的建议参数,设置如下:荧光素挥发因子 $\rho=0.4$,荧光素更新率 $\gamma=0.6$,动态决策域更新率 $\beta=0.08$,邻域阈值 $n_i=5$. 本文实验中结果均为独立重复实验 30 次,最后取平均值.

表 1 UCI 数据集

数据集	实例个数	属性个数	类别
Iris	150	4	2
Hayes	160	5	3
Wine	178	13	3
Seeds	210	7	3
Heart	270	13	2
Cleveland	303	13	5
Haberman	306	3	2
Column	310	6	2
Ecoli	336	8	8
Bupa	345	6	2
Ionosphere	351	34	2
Dermatology	366	34	6
Wholesale	440	7	2
Forest	523	27	4
Wdbc	569	30	2
ILP	583	10	2
Balance	625	4	3
Australian	690	14	2
BCW	699	10	2
Blood	748	4	2
Diabetes	768	8	2
Vehicle	846	18	4
Tic-tac-toe	958	9	2
German	1000	20	2
QSAR	1055	41	2
Diabetic-r	1151	19	2
CMC	1473	9	3
Yeast	1484	9	10
Wineq-r	1599	11	6
Car	1728	6	4
Segment	2310	19	7
Abalone	4177	8	3
Spambase	4601	57	2
Wineq-w	4898	11	7
Landsat	6435	36	6

5.2 预剪枝

本小节的目的：通过实验确定预剪枝后，应保留的基分类器的数目 M' ，为 IBGSOME_P 提供参数 M' 的选取依据。

本文在上述 35 个数据集上均进行了分析，鉴于篇幅的限制，故在此选择了两个具有代表性的数据集 Diabetes、Spambase 进行展示。在 Diabetes、Spambase 数据集上，在基分类器规模为 100、200 和 300 的条件下，在进行预剪枝后，剩下的分类器规模与集成分类性能间的关系，见图 1。由图 1 可知，在集成系统中基分类器规模的不断增加的情况下，集成分类准确率先呈上升趋势，然后下降。由于集成初期，基分类器数目较少，集成精度不高；但随着基分类器数目的不断增加，基分类器中存在一些冗余的基分类器，影响了集成精度，故出现了较为明显的下降现象。同时从图 1 中亦可发现，在基分类器数目达到 25 之前，集成精度可以达到最优。换言之，保留了前 25 个基分类器后，继续增加的基分类器的差异性和性能均较差，已经影响了集成精度。故保留前 25 个基分类器，已经达到为二次剪枝提供性能和差异性较好的基分类器的目的。基于边界最小化测度的预剪枝，只需保留 25 个基分类器即可，即 $M'=25$ 。因此，本文在采用边界最小化测度保留 25 个基分类器后，再使用 IBGSO 进行二次剪枝。结论：运用边界最小化测度进行预剪枝，保留前 25 个基分类器。

5.3 结果分析

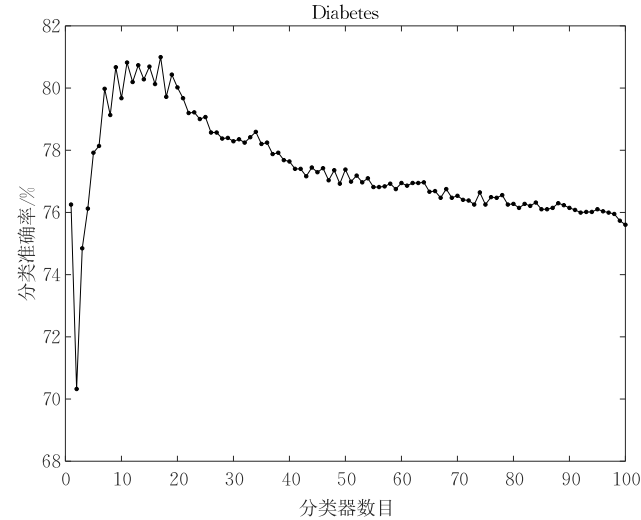
本小节目的：通过对比实验分析，评估 IBGSOME_P 的性能。

本文提出的 IBGSOME_P 在不同规模基分类器下的集成结果，与基分类器池中单个基分类器的最优、平均、最差精度的对比分析，如表 2、表 3 所示，其中“Max”，“Mean”，“Min”分别表示基分类器的最优、平均和最差的精度。同时从表 2、表 3 可以看出，IBGSOME_P 的集成精度在 35 个 UCI 数据集上均优于集成系统中单个分类器的最优和平均分类准确率。表 4、表 5 是在不同基分类器规模的条件下，本文方法的集成分类结果及最终子集成规模，与 Bagging 的对比分析，其中“ n ”表示最终子集成规模，表明了本文提出的 IBGSOME_P 所选择的最后子集成的规模，明显少于 Bagging，其集成分类性能也得到显著提升。

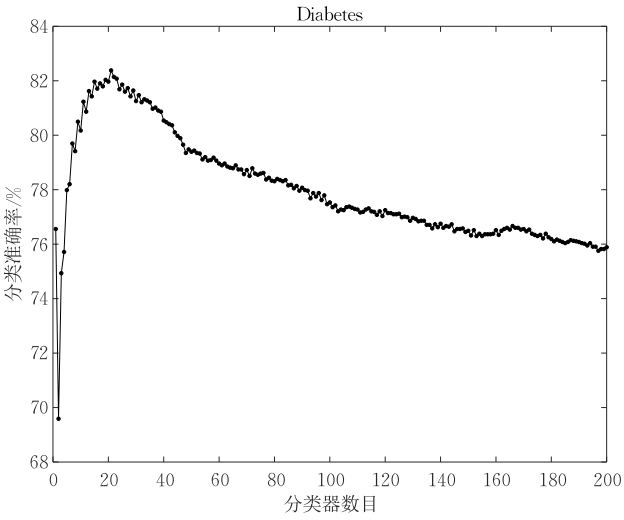
为了进一步评估本文提出方法的性能，下面与 Bagging^[45]，Kappa^[17]，AGOB^[15]，POBE^[16]，DREP^[19]，D-D^[28]，DF-D^[29]，GASEN^[12]，GSOSEN^[22]，MOAG^[18]，

RRE^[20]，DivP^[30]，SCG-P^[31]，SDAcc^[32]，RE^[14]，COM^[14]，MDM^[14]，FP-EP^[4]，DFTwo^[32]，AccRein^[32]，MDEP^[21] 进行了对比分析。Bagging 可以等概率地抽取训练数据集中的部分样本进行训练，可以增加基分类器间的差异性；Kappa，AGOB，POBE 和 MOAG 均采用了某种差异性测度对基分类器进行了排序集成，表明了排序集成的重要性；DREP 先选择性能最优的基分类器，再通过不断增加新的基分类器，使得新的基分类器集合在测试数据集上的集成误差最小化；D-D 通过计算基分类器的不一致测度，然后剔除不一致测度最大的基分类器，最后进行集成；DF-D 计算基分类器的双错测度，剔除双侧测度在置信区间之外的基分类器，最后进行集成；GASEN 通过随机赋给每个基分类器一个权重，然后采用遗传算法进行优化，使得集成误差最小化；GSOSEN 采用离散型萤火虫上算法，对基分类器进行集成剪枝；RRE 选择了集成剪枝过程中被遗弃的基分类器，进行综合集成，充分利用了资源；DivP 采用遗传算法将多个差异性测度融合成一个组合测度，并使用图着色理论选择部分基分类器进行集成；SCG-P 采用班扎夫权力系数衡量分类器的差异性贡献度，获得最优子集成，使其满足最小赢得联盟；SDAcc，DFTwo 和 AccRein 同时考虑了基分类器的差异性和精度，并采用 GEP 进行集成剪枝；RE，COM 和 MDM 采用了不同差异性测度，结合 GEP，进行了集成剪枝；MDEP 基于边界测度和差异性测度形成了新的组合测度，然后进行排序集成。

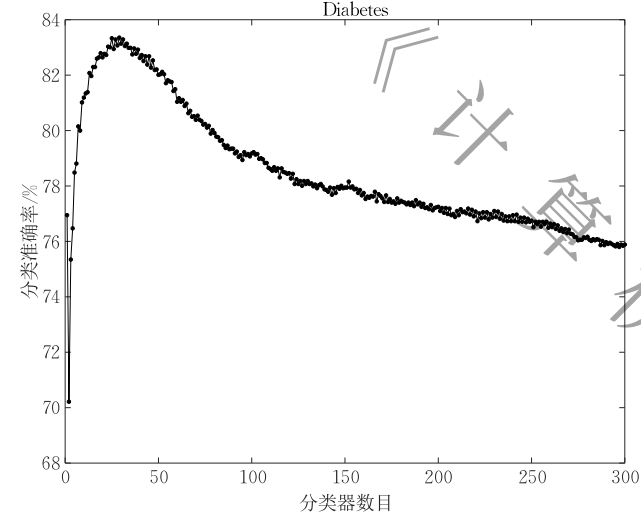
由表 2、表 3 可知，在原始分类器数目取 200 时，本文方法的集成分类准确率最高，进一步增加分类器规模，其集成分类精度仅能获得较小提升，所以分类器的数目宜取 200。为了保证 IBGSOME_P 的稳定性，本文在基分类器规模为 100 和 200 的条件下，与其他集成剪枝方法进行对比分析，如表 6~表 11 所示，其中“Win/Tie/Loss”分别表示在 35 个 UCI 数据上，集成精度优于、相同于、差于其他方法的个数。从表 6~表 11 可以看出，在 35 个 UCI 数据集上，当基分类器规模为 100 时，IBGSOME_P 的集成结果优于其他方法的数据集数目达到 23 个；当基分类器规模为 200 时，优于其他方法的数据集数目达到 24 个。总体来说，IBGSOME_P 的集成性能明显优于其他集成剪枝方法。由表 6~表 11 中的数据可得，各种集成剪枝方法所选取基分类器平均规模与平均精度之间的关系，如图 2 所示。由图 2 可知，本文方法可以使用较少数目的基分类器，获得了较优的集成性



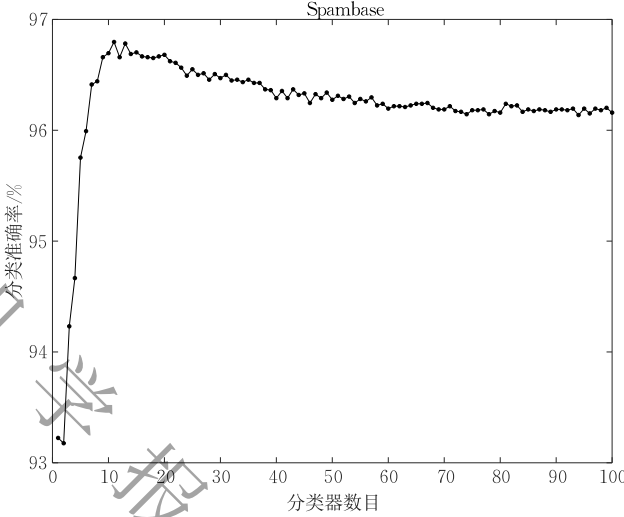
(a) 在Diabetes上基分类器池规模为100的排序集成性能分析



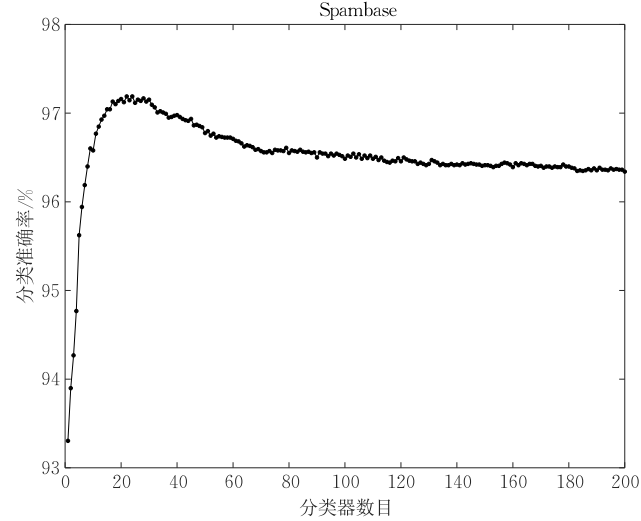
(b) 在Diabetes上基分类器池规模为200的排序集成性能分析



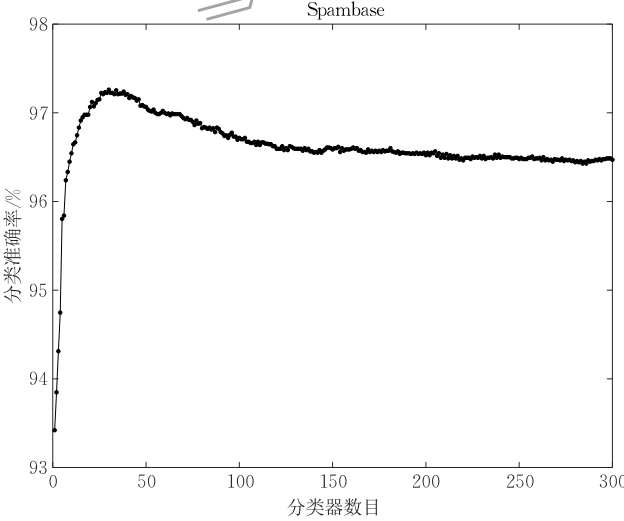
(c) 在Diabetes上基分类器池规模为300的排序集成性能分析



(d) 在Spambase上基分类器池规模为100的排序集成性能分析



(e) 在Spambase上基分类器池规模为200的排序集成性能分析



(f) 在Spambase上基分类器池规模300的排序集成性能分析

图 1 在 Diabetes,Spambase 上不同分类器池规模条件下按照边界最小化测度排序集成精度趋势分析

表 2 在不同规模基分类器(50,100,150)下的集成精度

数据集	50				100				150			
	IBGSOME _P	Best	Mean	Worst	IBGSOME _P	Best	Mean	Worst	IBGSOME _P	Best	Mean	Worst
Iris	97.84	97.30	93.84	86.22	98.38	97.30	93.70	83.78	98.92	97.84	93.79	83.51
Hayes	99.39	90.00	67.70	43.03	100.00	92.42	68.05	40.91	100.00	94.85	67.34	40.00
Wine	100.00	99.77	89.61	76.36	100.00	100.00	89.60	75.23	100.00	100.00	89.38	72.27
Seeds	100.00	99.42	89.88	77.69	100.00	98.65	90.04	77.69	100.00	99.23	90.04	75.96
Heart	94.03	85.07	75.80	63.98	95.92	86.47	75.80	62.49	97.11	87.06	75.92	61.00
Cleveland	65.32	58.38	48.16	37.12	68.56	58.92	48.19	35.05	70.81	60.18	48.21	35.50
Haberman	80.39	76.05	61.53	26.05	81.71	76.71	60.94	25.79	82.89	76.84	60.73	25.53
Column	96.10	88.92	80.42	71.17	98.61	89.87	80.44	69.78	98.70	90.13	80.46	70.30
Ecoli	88.89	82.06	73.55	63.41	90.48	83.02	73.22	61.51	90.79	83.33	73.06	59.92
Bupa	81.86	71.40	61.29	51.01	85.27	72.79	61.53	49.46	88.29	73.18	61.59	48.84
Ionosphere	99.90	95.90	88.58	79.14	100.00	97.24	88.48	76.29	100.00	97.14	88.47	76.86
Dermatology	100.00	99.58	92.10	79.58	100.00	100.00	92.19	77.50	100.00	100.00	92.31	78.75
Wholesale	94.92	91.74	86.41	79.32	95.98	92.73	86.50	78.03	96.44	92.27	86.38	77.12
Forest	88.38	81.24	71.73	61.71	89.81	81.90	72.11	59.90	91.43	82.76	72.29	60.00
Wdbc	100.00	97.28	90.95	81.40	100.00	97.98	91.06	80.70	100.00	97.89	90.99	79.21
ILP	86.55	79.49	70.63	61.37	89.34	80.17	70.77	60.97	90.26	80.51	70.58	58.86
Balance	88.96	80.59	73.60	66.08	90.93	80.96	73.84	65.81	91.31	81.71	73.71	64.96
Australian	76.47	66.52	57.96	48.70	79.86	67.63	57.98	47.25	81.35	68.55	57.93	47.15
BCW	100.00	98.00	93.91	88.62	100.00	98.43	93.86	88.10	100.00	98.52	93.98	87.86
Blood	76.93	72.53	66.14	59.60	77.27	73.27	66.10	59.27	78.87	73.53	66.21	59.27
Diabetes	83.38	75.54	68.01	60.04	85.41	76.28	68.08	59.31	86.32	76.54	67.99	58.96
Vehicle	84.79	74.62	67.68	61.30	87.04	76.21	67.82	58.99	89.23	76.09	67.87	59.53
Tic-tac-toe	99.10	90.28	81.36	71.25	99.86	90.87	81.07	70.69	100.00	91.15	81.17	69.58
German	83.67	74.50	66.88	59.87	85.33	74.30	66.68	59.43	86.80	74.40	66.82	57.97
QSAR	90.08	83.19	77.44	71.91	91.60	83.73	77.42	70.46	92.29	83.76	77.35	69.98
Diabetic-r	75.94	67.51	60.36	52.93	79.42	67.83	60.42	52.61	80.64	69.16	60.52	52.43
CMC	61.92	53.63	47.47	40.99	62.92	54.12	47.44	40.32	64.88	54.76	47.51	40.09
Yeast	67.88	54.35	48.18	41.49	69.80	55.05	48.11	41.33	70.41	55.47	48.03	40.41
Wineq-r	75.60	62.85	57.06	51.65	77.77	63.60	57.18	50.42	78.46	63.42	57.28	50.65
Car	99.77	96.03	92.87	88.23	99.96	96.38	92.86	88.52	100.00	96.71	92.92	87.53
Segment	99.24	96.53	93.44	89.42	99.60	96.47	93.52	89.16	99.82	96.91	93.53	88.73
Abalone	60.56	52.81	47.88	42.97	62.65	53.32	47.89	42.68	63.41	53.86	47.84	42.54
Spambase	97.23	92.72	89.88	86.99	97.52	93.10	89.88	86.77	97.86	93.10	89.90	86.10
Wineq-w	74.14	60.33	55.69	51.09	75.65	60.94	55.70	50.91	76.49	61.05	55.75	50.69
Landsat	92.46	84.97	82.39	79.68	93.20	85.73	82.45	79.45	93.47	85.58	82.41	78.95

表 3 在不同规模基分类器(200,250,300)下的集成精度

数据集	200				250				300			
	IBGSOME _P	Best	Mean	Worst	IBGSOME _P	Best	Mean	Worst	IBGSOME _P	Best	Mean	Worst
Iris	99.19	97.84	93.86	82.97	98.92	97.57	93.78	82.43	99.73	98.38	93.86	82.97
Hayes	100.00	95.45	67.30	38.48	100.00	95.45	67.53	36.97	100.00	95.76	67.58	36.97
Wine	100.00	100.00	89.34	72.95	100.00	100.00	89.43	71.36	100.00	100.00	89.46	71.82
Seeds	100.00	99.23	90.14	75.00	100.00	100.00	90.05	74.81	100.00	99.62	90.10	75.38
Heart	98.11	87.26	75.81	60.20	98.21	88.06	75.94	60.10	98.21	88.56	75.97	58.71
Cleveland	72.52	60.09	48.13	34.59	73.15	60.54	48.20	34.77	73.15	60.81	48.25	34.77
Haberman	85.13	77.89	60.78	24.74	83.82	77.37	60.79	24.34	83.68	77.63	60.81	24.61
Column	99.39	90.74	80.35	68.05	99.57	90.65	80.35	68.14	99.57	90.91	80.34	67.45
Ecoli	91.19	82.78	73.17	60.40	91.75	83.73	73.16	59.37	91.51	84.05	73.14	59.37
Bupa	88.60	73.80	61.45	49.38	89.85	73.41	61.55	47.83	90.00	73.80	61.47	48.22
Ionosphere	100.00	97.33	88.43	75.90	100.00	97.62	88.42	75.52	100.00	97.52	88.45	75.43
Dermatology	100.00	100.00	92.21	75.83	100.00	100.00	92.38	77.50	100.00	100.00	92.25	75.42
Wholesale	96.67	92.80	86.42	77.58	96.82	93.11	86.45	77.58	96.82	93.11	86.38	76.74
Forest	91.62	83.14	72.15	57.33	92.29	83.24	72.07	57.24	91.90	83.90	72.03	57.71
Wdbc	100.00	97.81	90.89	77.72	100.00	98.07	90.98	78.07	100.00	98.33	90.90	77.46
ILP	91.85	80.34	70.82	59.09	91.79	80.80	70.65	59.09	92.31	80.51	70.74	58.63
Balance	92.27	82.13	73.78	64.96	92.69	82.56	73.77	64.85	92.75	81.92	73.72	64.43
Australian	82.22	68.12	58.03	46.57	82.37	69.08	58.05	47.20	83.77	69.32	57.98	46.57
BCW	100.00	98.67	93.94	87.90	100.00	98.67	93.93	87.76	100.00	98.62	93.93	87.29
Blood	79.20	73.60	66.15	57.87	79.93	74.13	66.12	58.27	79.60	73.87	66.09	57.60
Diabetes	87.23	76.80	68.09	57.97	87.71	76.93	68.04	58.14	88.01	76.97	68.02	57.97
Vehicle	89.70	76.39	67.87	59.23	89.23	76.39	67.75	58.34	89.94	76.45	67.79	58.22
Tic-tac-toe	100.00	91.67	81.23	69.27	100.00	92.26	81.14	68.99	100.00	91.63	81.10	68.54
German	88.17	74.80	66.79	58.30	88.33	75.10	66.79	57.70	88.53	75.23	66.80	57.40
QSAR	93.14	84.08	77.35	68.85	93.43	84.04	77.36	68.66	93.14	84.74	77.33	68.69
Diabetic-r	81.48	68.90	60.43	52.32	82.00	69.36	60.45	52.09	82.70	69.19	60.48	52.09
CMC	64.36	55.57	47.48	40.07	65.76	55.68	47.42	39.32	66.03	55.64	47.50	39.77
Yeast	71.33	55.25	48.06	40.54	71.89	55.63	48.10	40.02	71.55	55.86	48.14	40.29
Wineq-r	78.92	63.75	57.22	50.83	79.73	63.92	57.24	50.54	79.29	64.15	57.24	49.98
Car	100.00	96.67	92.86	87.69	100.00	96.92	92.91	87.61	100.00	96.82	92.90	87.15
Segment	99.84	96.93	93.50	88.76	99.98	97.09	93.54	88.87	100.00	97.11	93.53	88.51
Abalone	64.53	54.26	47.89	42.01	65.18	54.10	47.89	42.01	65.90	54.29	47.87	41.72
Spambase	97.99	93.35	89.89	85.93	98.10	93.38	89.91	85.87	98.09	93.51	89.88	85.75
Wineq-w	76.98	61.16	55.73	49.97	76.95	61.54	55.78	50.03	77.06	61.65	55.74	49.95
Landsat	93.67	85.84	82.37	78.72	93.92	85.64	82.40	78.79	93.96	85.72	82.41	78.74

表 4 在不同规模基分类器(50,100,150)下 IBGSOME_P 与 Bagging 对比分析

数据集	50				100				150			
	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>
Iris	94.59	50	97.84	13	95.41	100	98.38	12	95.68	150	98.92	13
Hayes	81.52	50	99.39	13	83.03	100	100.00	14	79.39	150	100.00	16
Wine	99.09	50	100.00	12	98.41	100	100.00	13	99.32	150	100.00	12
Seeds	96.92	50	100.00	12	97.88	100	100.00	12	97.88	150	100.00	13
Heart	87.76	50	94.03	11	88.26	100	95.92	12	88.66	150	97.11	12
Cleveland	51.89	50	65.32	9	51.35	100	68.56	11	52.07	150	70.81	11
Haberman	71.84	50	80.39	11	71.71	100	81.71	12	71.84	150	82.89	12
Column	87.19	50	96.10	12	89.18	100	98.61	12	88.23	150	98.70	13
Ecoli	84.13	50	88.89	12	84.44	100	90.48	11	84.68	150	90.79	12
Bupa	68.29	50	81.86	11	68.84	100	85.27	13	69.69	150	88.29	14
Ionosphere	95.62	50	99.90	11	95.71	100	100.00	13	95.90	150	100.00	13
Dermatology	97.64	50	100.00	12	97.36	100	100.00	13	97.22	150	100.00	13
Wholesale	90.08	50	94.92	12	90.45	100	95.98	11	90.53	150	96.44	12
Forest	81.43	50	88.38	11	81.52	100	89.81	10	82.00	150	91.43	13
Wdbc	97.81	50	100.00	13	98.07	100	100.00	13	97.89	150	100.00	13
ILP	76.35	50	86.55	11	76.70	100	89.34	12	76.41	150	90.26	14
Balance	83.15	50	88.96	13	84.32	100	90.93	13	83.84	150	91.31	13
Australian	66.33	50	76.47	11	67.00	100	79.86	12	67.25	150	81.35	14
BCW	98.48	50	100.00	12	98.52	100	100.00	12	98.52	150	100.00	14
Blood	71.00	50	76.93	10	72.33	100	77.27	11	71.93	150	78.87	12
Diabetes	75.80	50	83.38	11	76.41	100	85.41	11	75.97	150	86.32	13
Vehicle	76.09	50	84.79	12	75.86	100	87.04	13	76.33	150	89.23	14
Tic-tac-toe	96.01	50	99.10	14	96.42	100	99.86	15	96.84	150	100.00	15
German	76.63	50	83.67	11	75.43	100	85.33	14	75.77	150	86.80	14
QSAR	84.80	50	90.08	11	84.87	100	91.60	12	84.64	150	92.29	13
Diabetic-r	66.41	50	75.94	12	67.54	100	79.42	13	67.77	150	80.64	13
CMC	52.90	50	61.92	13	52.99	100	62.92	13	53.49	150	64.88	13
Yeast	61.51	50	67.88	13	61.71	100	69.80	14	61.73	150	70.41	15
Wineq-r	70.75	50	75.60	12	70.92	100	77.77	14	71.19	150	78.46	14
Car	98.02	50	99.77	14	98.21	100	99.96	13	98.17	150	100.00	14
Segment	97.62	50	99.24	11	97.62	100	99.60	11	97.69	150	99.82	11
Abalone	54.43	50	60.56	12	54.35	100	62.65	13	54.45	150	63.41	15
Spambase	95.94	50	97.23	13	96.19	100	97.52	13	96.17	150	97.86	14
Wineq-w	70.11	50	74.14	14	70.78	100	75.65	16	70.60	150	76.49	14
Landsat	90.68	50	92.46	14	90.75	100	93.20	15	91.06	150	93.47	15

表 5 在不同规模基分类器(200,250,300)下 IBGSOME_P 与 Bagging 对比分析

数据集	200				250				300			
	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>	Bagging	<i>n</i>	IBGSOME _P	<i>n</i>
Iris	95.95	200	99.19	12	96.49	250	98.92	11	95.95	300	99.73	12
Hayes	80.00	200	100.00	14	78.79	250	100.00	12	79.70	300	100.00	14
Wine	98.86	200	100.00	12	99.32	250	100.00	13	98.86	300	100.00	13
Seeds	97.88	200	100.00	14	97.88	250	100.00	13	98.08	300	100.00	13
Heart	88.96	200	98.11	11	88.56	250	98.21	13	88.76	300	98.21	11
Cleveland	51.62	200	72.52	10	52.16	250	73.15	12	51.62	300	73.15	12
Haberman	71.84	200	85.13	13	71.45	250	83.82	12	72.11	300	83.68	13
Column	88.14	200	99.39	14	88.66	250	99.57	14	88.23	300	99.57	12
Ecoli	84.37	200	91.19	12	85.00	250	91.75	11	85.24	300	91.51	11
Bupa	68.76	200	88.60	14	68.84	250	89.85	14	68.99	300	90.00	14
Ionosphere	95.81	200	100.00	13	95.71	250	100.00	13	95.71	300	100.00	12
Dermatology	97.22	200	100.00	11	97.22	250	100.00	13	97.22	300	100.00	11
Wholesale	90.30	200	96.67	13	90.30	250	96.82	11	90.15	300	96.82	12
Forest	81.90	200	91.62	13	82.10	250	92.29	12	81.81	300	91.90	11
Wdbc	97.81	200	100.00	14	97.98	250	100.00	13	97.98	300	100.00	14
ILP	76.52	200	91.85	13	76.01	250	91.79	13	76.47	300	92.31	14
Balance	84.00	200	92.27	14	83.84	250	92.69	13	83.36	300	92.75	14
Australian	67.92	200	82.22	13	67.97	250	82.37	13	68.02	300	83.77	12
BCW	98.52	200	100.00	13	98.57	250	100.00	12	98.57	300	100.00	14
Blood	72.33	200	79.20	13	72.27	250	79.93	11	72.13	300	79.60	11
Diabetes	75.93	200	87.23	13	75.97	250	87.71	12	76.28	300	88.01	13
Vehicle	76.27	200	89.70	15	76.33	250	89.23	13	76.15	300	89.94	12
Tic-tac-toe	96.88	200	100.00	16	96.84	250	100.00	15	96.81	300	100.00	14
German	75.77	200	88.17	14	75.87	250	88.33	14	75.90	300	88.53	14
QSAR	85.02	200	93.14	13	85.02	250	93.43	14	84.90	300	93.14	13
Diabetic-r	67.45	200	81.48	14	66.93	250	82.00	15	67.28	300	82.70	14
CMC	52.95	200	64.36	13	52.79	250	65.76	13	53.08	300	66.03	13
Yeast	61.78	200	71.33	14	61.94	250	71.89	14	61.89	300	71.55	15
Wineq-r	70.77	200	78.92	15	71.04	250	79.73	14	71.08	300	79.29	14
Car	98.17	200	100.00	13	98.17	250	100.00	14	98.23	300	100.00	14
Segment	97.73	200	99.84	13	97.76	250	99.98	12	97.73	300	100.00	14
Abalone	54.48	200	64.53	14	54.77	250	65.18	15	54.61	300	65.90	15
Spambase	96.26	200	97.99	13	96.45	250	98.10	13	96.35	300	98.09	12
Wineq-w	71.02	200	76.98	16	71.07	250	76.95	16	71.03	300	77.06	15
Landsat	90.96	200	93.67	15	91.02	250	93.92	15	91.01	300	93.96	15

表 6 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 100)

数据集	IBGSOME	<i>n</i>	Bagging	<i>n</i>	Kappa	<i>n</i>	AGOB	<i>n</i>	POBE	<i>n</i>	DREP	<i>n</i>	D-D	<i>n</i>
Iris	98.38	12	95.41	100	95.14	20	97.84	13	95.68	14	96.22	50	95.14	99
Hayes	100.00	14	83.03	100	96.57	20	96.77	26	97.17	21	89.49	50	79.60	99
Wine	100.00	13	98.41	100	100.00	20	100.00	16	100.00	28	99.39	50	99.24	99
Seeds	100.00	12	97.88	100	100.00	20	98.72	17	100.00	23	98.46	50	97.56	99
Heart	95.92	12	88.26	100	89.65	20	92.54	18	92.84	20	91.24	50	87.96	99
Cleveland	68.56	11	51.35	100	55.41	20	61.98	13	58.02	22	56.85	50	52.07	99
Haberman	81.71	12	71.71	100	72.77	20	76.93	15	75.35	31	75.70	50	71.67	99
Column	98.61	12	89.18	100	92.55	20	94.81	12	93.94	21	92.64	50	87.71	99
Ecoli	90.48	11	84.44	100	83.89	20	87.30	22	86.67	17	86.59	50	85.08	99
Bupa	85.27	13	68.84	100	70.70	20	78.29	16	79.69	21	75.35	50	69.61	99
Ionosphere	100.00	13	95.71	100	99.14	20	99.62	13	99.90	21	97.43	50	95.81	99
Dermatology	100.00	13	97.36	100	100.00	20	99.07	16	100.00	20	98.43	50	97.31	99
Wholesale	95.98	11	90.45	100	91.89	20	94.02	18	93.94	20	92.42	50	90.53	99
Forest	89.81	10	81.52	100	84.25	20	87.37	15	86.10	22	84.25	50	81.71	99
Wdbc	100.00	13	98.07	100	100.00	20	100.00	16	100.00	23	99.65	50	98.30	99
ILP	89.34	12	76.70	100	81.14	20	82.91	18	83.76	21	81.42	50	76.52	99
Balance	90.93	13	84.32	100	86.19	20	87.79	18	86.93	22	86.19	50	82.83	99
Australian	79.86	12	67.00	100	65.07	20	72.75	24	72.61	21	71.35	50	67.29	99
BCW	100.00	12	98.52	100	99.86	20	99.33	20	99.62	21	98.86	50	98.38	99
Blood	77.27	11	72.33	100	71.02	20	75.02	17	75.29	21	73.91	50	71.78	99
Diabetes	85.41	11	76.41	100	78.61	20	80.48	20	80.48	22	80.04	50	75.58	99
Vehicle	87.04	13	75.86	100	80.47	20	82.56	17	83.43	21	80.55	50	75.86	99
Tic-tac-toe	99.86	15	96.42	100	99.24	20	98.65	29	99.48	22	98.30	50	96.74	99
German	85.33	14	75.43	100	78.10	20	80.33	20	81.80	22	79.67	50	75.87	99
QSAR	91.60	12	84.87	100	88.50	20	89.86	16	88.78	21	87.65	50	84.96	99
Diabetic-r	79.42	13	67.54	100	70.84	20	73.74	20	74.75	22	72.87	50	67.48	99
CMC	62.92	13	52.99	100	53.79	20	59.41	16	58.53	21	56.50	50	53.04	99
Yeast	69.80	14	61.71	100	62.07	20	65.90	18	64.50	22	65.79	50	61.98	99
Wineq-r	77.77	14	70.92	100	71.60	20	74.46	23	73.60	22	73.67	50	71.02	99
Car	99.96	13	98.21	100	99.38	20	99.73	21	99.23	21	99.04	50	98.07	99
Segment	99.60	11	97.62	100	98.78	20	99.07	15	98.78	22	98.18	50	97.60	99
Abalone	62.65	13	54.35	100	54.23	20	58.79	19	59.28	21	57.97	50	54.83	99
Spambase	97.52	13	96.19	100	96.75	20	97.03	14	97.01	20	96.80	50	96.30	99
Wineq-w	75.65	16	70.78	100	71.41	20	73.32	22	73.35	21	72.73	50	70.56	99
Landsat	93.20	15	90.75	100	91.31	20	91.80	25	91.67	20	91.74	50	90.93	99
Win/Tie/Loss	23/7/5		0/0/35		0/4/31		0/1/34		0/4/31		0/0/35		0/0/35	

表 7 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 100)

数据集	DF-D	<i>n</i>	GA SEN	<i>n</i>	GS SEN	<i>n</i>	MOAG	<i>n</i>	RRE	<i>n</i>	DivP	<i>n</i>	SCG-P	<i>n</i>	SDAcc	<i>n</i>
Iris	97.30	51	97.30	51	97.30	50	96.94	17	97.84	11	97.48	3	96.40	22	98.20	6
Hayes	95.15	55	95.96	45	97.37	45	89.49	22	99.60	12	98.79	5	88.28	20	100.00	12
Wine	100.00	58	100.00	51	99.85	51	100.00	18	100.00	11	100.00	2	100.00	7	100.00	22
Seeds	99.23	58	99.49	41	98.97	49	98.85	16	100.00	11	99.87	3	97.95	9	100.00	15
Heart	90.95	57	92.09	51	92.54	47	89.55	17	94.23	15	92.04	3	91.65	27	92.74	23
Cleveland	56.04	55	60.54	45	58.78	48	56.40	27	66.85	14	63.51	2	63.33	48	65.09	8
Haberman	73.42	64	76.75	36	77.28	46	75.26	20	81.05	12	78.25	5	74.30	19	76.23	33
Column	92.21	56	93.77	50	92.99	49	90.91	14	96.88	16	94.29	3	92.39	27	93.51	21
Ecoli	86.83	55	87.38	48	87.26	47	86.11	16	89.37	16	87.54	5	85.48	53	87.70	37
Bupa	74.96	55	78.84	44	75.47	48	72.79	28	84.03	15	79.92	7	76.33	34	75.58	23
Ionosphere	97.05	56	97.43	50	97.71	48	96.95	19	99.62	11	98.95	1	96.57	7	100.00	9
Dermatology	99.44	55	99.63	41	98.89	48	98.61	19	100.00	12	100.00	2	97.59	24	100.00	25
Wholesale	92.35	55	92.84	46	93.18	42	91.89	22	95.30	11	94.24	3	93.67	14	93.79	7
Forest	83.94	56	85.21	45	85.97	45	83.81	17	89.27	14	86.03	10	83.92	21	85.46	13
Wdbc	99.30	58	99.94	45	99.82	46	98.95	18	100.00	11	99.88	9	99.12	16	100.00	23
ILP	81.03	56	83.42	45	81.62	42	79.49	32	87.24	16	85.24	6	83.75	19	85.82	21
Balance	86.29	57	88.56	48	87.68	49	84.75	26	89.92	25	87.52	2	85.63	74	86.35	41
Australian	69.66	56	73.77	45	72.25	46	69.52	15	77.92	17	75.51	9	72.37	34	75.92	26
BCW	98.81	57	99.00	47	99.29	47	98.71	20	99.81	13	100.00	5	98.67	5	100.00	8
Blood	73.87	56	74.84	47	75.33	46	73.16	22	78.13	12	76.53	11	77.76	40	74.40	25
Diabetes	79.65	56	81.10	47	80.00	46	77.45	27	84.16	20	82.12	6	81.27	52	82.31	23
Vehicle	80.83	54	81.14	47	82.52	49	78.78	20	86.19	22	83.59	15	83.75	41	79.61	26
Tic-tac-toe	98.78	57	99.84	51	99.53	48	97.40	14	99.65	24	98.99	11	98.92	28	99.92	30
German	79.20	58	81.65	47	80.00	47	77.87	27	84.97	21	81.80	3	81.87	37	81.40	23
QSAR	87.01	57	88.53	43	87.82	48	86.73	34	91.94	14	89.51	2	87.35	29	88.36	23
Diabetic-r	71.25	56	75.22	46	72.78	47	69.65	31	77.94	22	76.49	4	75.36	33	75.61	31
CMC	56.16	55	58.54	46	56.92	46	54.87	20	62.71	19	60.86	7	57.03	77	59.77	28
Yeast	63.83	57	66.86	48	65.24	48	63.18	24	69.48	29	66.67	13	65.57	76	66.96	55
Wineq-r	73.04	57	75.38	49	74.16	52	72.35	28	77.10	23	74.79	6	73.81	67	75.94	34
Car	98.81	56	99.25	47	99.10	49	98.44	16	99.77	15	99.09	8	98.40	48	99.44	36
Segment	98.13	57	98.40	49	98.40	45	98.09	23	99.24	12	98.71	3	97.84	27	99.38	12
Abalone	57.54	56	59.98	47	58.68	48	56.09	30	62.89	25	61.10	6	59.53	82	61.59	51
Spambase	96.86	57	97.13	47	96.96	47	96.51	10	97.45	22	96.87	7	96.55	54	96.87	40
Wineq-w	72.34	56	74.22	48	73.18	47	71.47	20	75.48	29	73.99	8	73.20	74	73.69	50
Landsat	91.57	56	92.60	48	92.02	48	91.22	21	92.87	33	91.88	9	91.16	71	91.38	56
Win/Tie/Loss	0/1/34		0/1/34		0/0/35		0/1/34		3/3/29		0/3/32		0/1/34		1/6/28	

表 8 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 100)

数据集	RE	<i>n</i>	COMP	<i>n</i>	MDM	<i>n</i>	EP-FP	<i>n</i>	DFTwo	<i>n</i>	AccRein	<i>n</i>	MDEP	<i>n</i>
Iris	95.20	4	98.02	7	98.56	12	95.86	58	98.20	4	98.20	4	97.68	30
Hayes	98.60	11	96.36	16	99.60	11	95.05	52	99.19	11	99.60	15	99.43	30
Wine	100.00	3	100.00	3	100.00	10	99.09	58	100.00	2	100.00	2	99.09	30
Seeds	99.62	8	98.59	11	100.00	10	97.44	57	100.00	10	100.00	10	99.44	30
Heart	93.13	13	91.84	17	92.64	15	89.76	69	92.24	25	92.24	25	94.66	30
Cleveland	65.48	13	62.34	11	60.36	12	59.70	52	65.09	8	65.09	8	68.25	30
Haberman	80.35	14	78.25	7	77.81	18	75.84	54	77.23	10	77.23	11	80.96	30
Column	93.06	14	95.06	13	94.81	11	91.23	59	93.73	25	93.73	25	97.97	30
Ecoli	87.44	11	87.14	17	87.30	15	85.40	58	88.10	27	88.10	26	88.08	30
Bupa	83.65	17	80.23	15	80.08	14	75.63	56	76.05	20	76.74	14	83.46	30
Ionosphere	99.62	10	99.71	11	100.00	10	95.90	57	99.62	11	99.62	11	100.00	30
Dermatology	100.00	2	99.91	1	100.00	10	97.50	57	100.00	2	100.00	2	99.41	30
Wholesale	93.77	12	93.94	14	93.71	11	90.53	58	93.86	9	93.86	9	93.45	30
Forest	84.71	17	87.56	19	87.11	11	85.90	95	85.33	24	85.33	24	89.95	30
Wdbc	100.00	10	99.94	19	100.00	10	97.95	55	100.00	16	100.00	16	97.95	30
ILP	86.29	16	82.11	19	83.08	14	81.47	58	86.65	16	86.60	16	86.18	30
Balance	88.08	19	87.89	21	87.04	17	85.20	58	86.81	46	86.87	50	86.04	30
Australian	75.50	18	73.09	17	73.04	15	73.65	50	76.50	33	76.16	28	77.05	30
BCW	97.71	10	99.38	16	99.90	11	98.52	48	100.00	13	100.00	13	98.52	30
Blood	77.64	13	74.53	15	74.31	16	74.04	58	75.69	26	75.69	26	76.00	30
Diabetes	82.76	13	80.65	18	80.56	14	79.10	57	83.61	32	83.57	28	82.80	30
Vehicle	83.96	17	82.01	15	82.33	17	78.37	59	80.25	21	80.25	21	83.09	30
Tic-tac-toe	98.58	22	98.09	29	98.72	17	96.98	58	98.68	50	98.65	45	98.81	30
German	82.17	19	81.57	13	81.37	15	80.13	58	82.50	36	83.00	36	83.90	30
QSAR	90.41	15	89.28	12	89.26	15	86.96	59	90.17	17	91.02	17	90.83	30
Diabetic-r	76.86	16	73.62	16	73.54	17	73.80	60	76.43	22	76.58	27	77.28	30
CMC	59.67	16	58.01	14	58.46	15	55.38	51	60.23	19	60.20	17	61.68	30
Yeast	66.94	20	65.36	24	65.14	25	64.23	52	68.10	41	68.74	38	67.46	30
Wineq-r	74.21	23	74.58	23	73.40	24	71.15	59	75.48	29	75.48	27	73.94	30
Car	97.75	16	99.65	16	99.52	13	98.15	59	98.94	35	98.94	35	98.15	30
Segment	97.31	13	98.76	15	98.71	11	97.58	58	98.91	22	98.91	23	97.53	30
Abalone	60.78	25	59.03	15	58.66	16	54.69	56	61.62	41	61.68	44	61.23	30
Spambase	93.55	16	96.74	18	96.84	14	96.14	59	96.65	42	96.65	42	96.10	30
Wineq-w	72.61	32	72.69	22	72.37	28	70.57	57	74.55	58	73.46	54	74.33	30
Landsat	90.30	25	92.20	29	92.01	23	91.07	59	92.58	66	91.94	63	92.02	30
Win/Tie/Loss	0/3/32		0/1/34		0/5/30		0/0/35		0/5/30		0/5/30		1/1/33	

表 9 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 200)

数据集	IBGSOME	<i>n</i>	Bagging	<i>n</i>	Kappa	<i>n</i>	AGOB	<i>n</i>	POBE	<i>n</i>	DREP	<i>n</i>	D-D	<i>n</i>
Iris	99.19	12	95.95	200	95.50	40	98.56	24	96.04	24	96.40	100	96.04	199
Hayes	100.00	14	80.00	200	98.79	40	97.58	40	98.18	43	90.91	100	81.01	199
Wine	100.00	12	98.86	200	100.00	40	100.00	23	100.00	55	100.00	100	99.09	199
Seeds	100.00	14	97.88	200	100.00	40	98.21	24	100.00	44	98.21	100	97.82	199
Heart	98.11	11	88.96	200	90.65	40	92.64	28	93.43	40	92.24	100	88.86	199
Cleveland	72.52	10	51.62	200	55.32	40	62.07	23	58.11	43	56.49	100	51.71	199
Haberman	85.13	13	71.84	200	72.19	40	78.60	21	75.61	60	76.84	100	71.84	199
Column	99.39	14	88.14	200	94.37	40	96.45	24	93.94	42	93.25	100	88.14	199
Ecoli	91.19	12	84.37	200	85.08	40	87.46	33	87.30	40	86.90	100	84.76	199
Bupa	88.60	14	68.76	200	75.19	40	82.17	24	80.93	43	76.51	100	69.53	199
Ionosphere	100.00	13	95.81	200	99.90	40	99.90	21	100.00	40	97.43	100	95.62	199
Dermatology	100.00	11	97.22	200	100.00	40	99.63	45	100.00	41	98.70	100	97.22	199
Wholesale	96.67	13	90.30	200	93.03	40	94.32	27	94.55	42	92.73	100	90.53	199
Forest	91.62	13	81.90	200	85.33	40	87.43	23	86.22	41	84.25	100	81.46	199
Wdbc	100.00	14	97.81	200	100.00	40	100.00	32	100.00	45	99.42	100	98.25	199
ILP	91.85	13	76.52	200	82.56	40	83.30	27	85.58	42	81.14	100	76.13	199
Balance	92.27	14	84.00	200	87.20	40	89.23	35	88.00	43	87.41	100	84.00	199
Australian	82.22	13	67.92	200	69.08	40	74.98	24	74.44	41	71.21	100	67.54	199
BCW	100.00	13	98.52	200	100.00	40	99.38	32	99.90	41	98.62	100	98.57	199
Blood	79.20	13	72.33	200	72.49	40	75.29	28	75.33	42	74.62	100	71.96	199
Diabetes	87.23	13	75.93	200	78.74	40	81.39	29	82.34	42	80.48	100	76.02	199
Vehicle	89.70	15	76.27	200	81.07	40	83.75	25	83.23	42	81.18	100	76.13	199
Tic-tac-toe	100.00	16	96.88	200	99.79	40	98.78	44	99.79	44	98.65	100	97.08	199
German	88.17	14	75.77	200	79.10	40	82.40	28	83.13	42	79.97	100	76.20	199
QSAR	93.14	13	85.02	200	88.85	40	90.08	27	89.38	41	87.68	100	85.37	199
Diabetic-r	81.48	14	67.45	200	70.99	40	75.04	25	75.51	41	72.61	100	67.62	199
CMC	64.36	13	52.95	200	54.28	40	59.07	27	57.74	42	56.84	100	52.86	199
Yeast	71.33	14	61.78	200	63.69	40	67.25	29	65.70	42	65.38	100	61.58	199
Wineq-r	78.92	15	70.77	200	72.75	40	75.35	25	74.98	43	74.42	100	70.96	199
Car	100.00	13	98.17	200	99.73	40	99.79	29	99.65	46	99.13	100	98.17	199
Segment	99.84	13	97.73	200	98.87	40	99.16	25	98.73	42	98.24	100	97.78	199
Abalone	64.53	14	54.48	200	55.09	40	60.37	31	60.49	42	58.76	100	54.29	199
Spambase	97.99	13	96.26	200	97.00	40	97.19	33	97.20	40	96.97	100	96.36	199
Wineq-w	76.98	16	71.02	200	72.39	40	74.16	39	74.04	43	73.59	100	70.90	199
Landsat	93.67	15	90.96	200	92.33	40	92.55	39	92.74	43	92.16	100	90.99	199
Win/Tie/Loss	24/7/4		0/0/35		0/5/30		0/2/33		0/5/30		0/1/34		0/0/35	

表 10 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 200)

数据集	DF-D	<i>n</i>	GASEN	<i>n</i>	GSOSEN	<i>n</i>	MOAG	<i>n</i>	RRE	<i>n</i>	DivP	<i>n</i>	SCG-P	<i>n</i>	SDAcc	<i>n</i>
Iris	97.30	84	97.30	103	97.12	100	97.30	9	99.10	21	97.84	3	96.94	25	99.46	14
Hayes	96.16	105	94.55	95	95.96	95	90.71	23	99.80	26	99.60	8	89.90	23	100.00	10
Wine	100.00	113	100.00	101	100.00	98	99.70	21	100.00	21	100.00	5	100.00	27	100.00	20
Seeds	99.74	110	98.72	82	98.33	101	98.59	35	100.00	21	100.00	2	98.08	61	100.00	37
Heart	91.34	110	91.49	98	91.64	100	90.05	26	95.12	22	92.34	6	91.45	62	94.43	26
Cleveland	56.76	106	58.78	96	57.84	66	57.66	30	69.10	21	64.95	12	63.79	67	65.82	5
Haberman	72.72	128	76.49	72	76.49	97	75.18	36	82.11	24	79.67	7	77.82	29	77.28	8
Column	92.55	109	93.38	96	92.99	98	90.13	46	97.58	22	95.41	6	92.65	72	96.02	26
Ecoli	86.90	106	86.90	102	86.90	94	86.11	34	90.08	21	88.17	11	85.71	74	88.65	33
Bupa	76.05	107	76.40	98	75.23	93	73.64	47	86.36	25	81.16	17	76.25	70	76.51	29
Ionosphere	97.24	106	96.86	101	97.43	91	96.86	32	99.81	25	98.86	3	96.86	6	100.00	7
Dermatology	99.72	105	99.54	87	97.96	76	98.43	26	100.00	22	100.00	3	97.50	31	100.00	21
Wholesale	92.65	108	92.73	95	93.07	72	91.89	36	95.61	22	94.70	5	93.74	17	94.24	12
Forest	84.00	107	84.70	90	84.70	97	83.30	38	89.71	24	87.24	7	83.24	25	85.90	17
Wdbc	99.47	113	99.54	84	99.42	92	99.06	17	100.00	21	99.94	5	99.06	17	100.00	12
ILP	80.97	108	82.14	95	81.62	81	79.49	42	88.77	25	86.27	11	84.03	35	85.82	22
Balance	86.72	108	87.76	103	87.04	97	85.07	34	90.45	25	88.96	22	86.32	39	87.09	74
Australian	70.43	107	72.61	97	71.23	89	69.42	49	81.45	22	77.10	12	72.81	44	76.61	23
BCW	98.86	111	98.57	100	98.86	71	98.67	39	99.90	21	100.00	4	98.76	5	100.00	8
Blood	74.13	108	74.58	97	74.93	76	73.51	26	78.62	26	76.89	5	77.73	41	74.39	12
Diabetes	79.61	107	81.04	97	80.26	97	77.66	46	85.41	26	83.33	15	81.92	54	81.65	32
Vehicle	79.84	106	80.24	88	81.74	75	79.13	55	87.57	30	84.02	15	82.67	81	86.61	35
Tic-tac-toe	99.24	108	99.64	96	99.27	97	98.06	36	99.69	24	99.41	34	99.30	61	100.00	26
German	80.13	109	80.85	92	79.60	94	77.87	60	86.63	34	84.43	20	81.57	37	81.50	30
QSAR	87.17	108	88.01	98	87.39	89	86.41	47	92.32	25	89.79	11	87.41	39	87.99	35
Diabetic-r	71.88	109	73.43	97	71.70	93	69.62	54	80.23	32	77.36	20	75.43	62	75.68	48
CMC	56.20	108	57.29	95	56.24	93	54.82	67	64.60	24	61.60	13	58.65	82	59.81	17
Yeast	64.01	108	65.64	97	64.80	97	62.84	62	71.45	32	68.40	20	65.98	88	66.48	62
Wineq-r	73.56	107	75.09	96	74.10	95	72.71	54	78.73	32	76.25	34	73.56	72	76.15	45
Car	98.77	111	99.13	96	98.99	90	98.61	32	99.77	26	99.23	11	98.34	86	99.87	55
Segment	98.20	109	98.33	94	98.33	98	97.98	52	99.53	24	98.78	12	97.91	77	99.84	22
Abalone	58.41	109	59.78	94	57.92	89	55.98	59	64.69	40	62.54	27	56.12	86	61.73	55
Spambase	96.86	109	97.02	99	97.02	95	96.62	30	97.77	28	97.01	24	96.68	73	96.93	56
Wineq-w	72.97	109	73.94	94	73.16	96	71.95	32	77.05	49	75.02	33	73.35	96	74.46	72
Landsat	91.91	110	92.44	99	92.07	100	91.45	58	92.82	30	92.55	35	91.35	94	91.81	81
Win/Tie/Loss	0/1/34		0/1/34		0/1/34		0/0/35		4/4/27		0/4/31		0/1/34		0/7/28	

表 11 与其他方法在集成精度和集成规模方面对比分析(基分类器规模 200)

数据集	RE	<i>n</i>	COMP	<i>n</i>	MDM	<i>n</i>	EP-EP	<i>n</i>	DFTwo	<i>n</i>	AccRein	<i>n</i>	MDEP	<i>n</i>
Iris	96.46	14	98.56	10	99.10	21	96.22	78	99.10	11	99.10	11	97.22	30
Hayes	99.39	20	97.37	26	100.00	20	95.62	66	100.00	9	100.00	9	99.20	30
Wine	100.00	1	100.00	1	100.00	20	99.39	94	100.00	2	100.00	2	99.39	30
Seeds	99.87	9	99.23	9	100.00	20	97.82	86	100.00	5	100.00	5	99.82	30
Heart	93.23	23	92.84	28	93.83	24	89.15	97	95.23	33	95.03	31	94.96	30
Cleveland	65.74	20	62.16	12	62.34	23	59.34	93	66.45	13	66.54	14	67.98	30
Haberman	80.72	21	78.60	10	79.30	21	75.28	74	77.28	8	77.28	8	81.05	30
Column	93.92	21	96.02	25	96.36	24	91.92	98	93.77	15	93.77	15	97.74	30
Ecoli	86.92	20	87.54	37	87.54	21	85.32	98	88.73	48	88.73	47	87.16	30
Bupa	83.05	20	81.40	24	82.40	23	77.70	92	75.66	32	76.81	33	83.99	30
Ionosphere	99.71	20	100.00	24	100.00	20	95.71	97	99.71	9	99.71	9	100.00	30
Dermatology	100.00	1	100.00	1	100.00	20	97.22	93	100.00	2	100.00	2	99.22	30
Wholesale	93.91	20	94.24	27	94.70	20	90.61	95	94.09	15	94.09	15	93.61	30
Forest	84.90	20	87.62	24	87.81	23	87.16	96	85.97	26	86.03	26	89.16	30
Wdbc	100.00	20	100.00	24	100.00	20	98.36	86	100.00	19	100.00	19	98.19	30
ILP	87.00	24	83.82	28	84.84	23	81.52	96	83.28	22	85.62	33	86.52	30
Balance	88.61	29	89.81	32	89.23	28	85.27	97	86.67	96	86.56	98	87.11	30
Australian	75.34	23	75.17	21	75.41	25	73.89	30	75.74	32	75.69	30	77.97	30
BCW	97.67	17	99.52	31	99.95	20	98.57	92	100.00	10	100.00	10	98.57	30
Blood	76.98	24	75.38	27	75.42	27	74.22	96	74.13	36	74.53	36	76.22	30
Diabetes	82.84	24	81.73	36	81.99	27	79.41	97	81.65	31	81.65	30	82.15	30
Vehicle	83.01	24	83.98	27	83.87	27	79.13	97	86.29	41	86.25	41	83.06	30
Tic-tac-toe	97.83	21	98.82	40	99.79	25	97.08	95	99.13	93	99.13	93	98.98	30
German	82.47	29	81.87	32	83.00	23	80.13	96	81.60	32	81.60	32	83.87	30
QSAR	90.29	24	90.05	25	89.86	23	87.02	95	87.58	38	87.58	38	90.02	30
Diabetic-r	76.09	27	75.48	29	76.12	28	73.14	58	76.30	21	76.39	26	77.39	30
CMC	60.18	25	59.73	24	59.75	25	56.33	61	60.56	18	60.61	18	61.88	30
Yeast	66.17	31	66.85	28	67.25	32	64.30	51	64.39	52	66.48	48	67.44	30
Wineq-r	74.88	28	75.50	28	75.48	32	71.40	73	76.10	50	76.13	62	74.96	30
Car	97.83	23	99.77	33	99.67	20	98.21	98	99.21	52	99.21	52	98.21	30
Segment	97.49	23	99.18	23	98.93	20	97.76	95	99.18	26	99.18	26	97.76	30
Abalone	60.55	35	60.69	31	60.51	25	54.48	99	60.22	55	60.22	58	60.37	30
Spambase	93.71	24	97.23	35	97.14	20	96.30	98	96.91	57	96.91	57	96.28	30
Wineq-w	72.83	40	73.88	37	74.01	32	70.71	95	73.89	101	73.82	82	74.59	30
Landsat	90.79	32	92.42	41	92.33	31	91.16	98	91.67	112	91.65	113	93.08	30
Win/Tie/Loss	0/3/32		0/4/31		0/6/29		0/0/35		0/6/29		0/6/29		0/1/34	

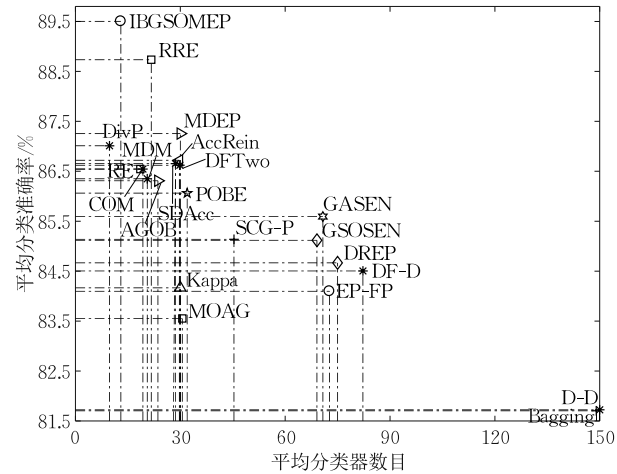


图 2 平均集成规模与平均集成精度之间的关系分析

能,相较于其他集成剪枝方法,具有明显的优势.

本文提出的 IBGSOMEPE 先采用边界最小化测度进行预剪枝,再运用 IBGSO 进行二次剪枝,获得了集成性能更优的基分类器子集合.而边界最小化测度(MDM)单独采用边界最小化测度进行剪枝,只能剔除部分性能和差异性较差的基分类器,却无法实现较为精确的剪枝,故影响了最后的集成性能;GASEN、GSSEN 分别采用了 GA 和 GSO 进行选择集成,由于集成系统中分类器的数目较多,其子集成的规模十分庞大(规模为 100 的基分类器池,有 1.27×10^{30} 个子集合),仅单纯采用 GA 或 GSO 无法找到性能较优的基分类器子集合.故本文采用预剪枝和二次剪枝相结合的方式,为集成剪枝问题提供了新的解决思路.

Bagging 集成了基分类器池中全部分类器,未能剔除性能和差异性较差的分类器,影响了集成性能;Kappa、AGOB、POBE、DREP、D-D、DF-D、MOAG、RRE、SCG-P、RE、COMP、MDM、SDAcc、DFTwo、AccRein、MDEP 均采用了某一种差异性测度或多种测度的组合测度,结合不同的剪枝策略,仅能剪枝集成系统中部分冗余分类器,无法实现精确剪枝,故影响了集成性能;EP-FP 剪枝后的基分类器数目太多,包含了大量性能和差异性较差的基分类器,影响了集成性能;DivP 采用 GA 优化多种差异性测度,形成了组合测度,再采用图着色理论保留的分类器数目太少,分类器间缺乏差异性,最终集成分类器性能较差;GASEN、GSSEN 难以从规模如此庞大的子集成中,搜索到集成分类精度最高的子集成,即无法实现精确剪枝.

为了进一步挖掘本文方法的新颖性,对 IBGSOMEPE 进行进一步分析.通过表 6~表 11 可知,本文提出的 IBGSOMEPE 相较于仅单独采用元启发式算

法的 GASEN 和 GSSEN 以及单独采用边界最小化测度的 MDM,使用了更少数目的基分类器,获得了更优的集成性能.从而说明,本文采用基于差异性测度和元启发式算法的组合剪枝方法的有效性和显著性.同时,由表 6~表 11 亦可知,当基分类器池的规模由 100 增加到 200 时,IBGSOMEPE 和 MDM 的集成性能均有一定的提升,但 GASEN 和 GSSEN 的集成性能反而下降了.这说明了 IBGSOMEPE 和 MDM 均适用于集成规模较大的集成剪枝问题;但元启发式搜索算法不适用于集成规模较大的集成剪枝问题,需要对集成规模较大的剪枝问题进行预剪枝.同时,也说明了本文方法中预剪枝操作的合理性.故而,本文先采用差异性测度预剪枝部分综合性能较差的基分类器,减少基分类器的整体规模,大幅提升集成剪枝问题的计算效率;再采用元启发式搜索算法进行二次剪枝,以搜索到最优子集成.基于差异性测度和元启发式搜索算法的融合型集成剪枝的方法为集成剪枝领域开辟了一条新的研究思路.这也是本文最重要的原创性和创新性.

为了检验本文提出的 IBGSOMEPE 与其他集成剪枝方法间差异性是否显著.本文采用 Wilcoxon 秩检验方法进行检验^[30](一般显著水平取 0.05),即当 p -值小于 0.05 时,两方法间的差异性显著;否则,差异性不显著.由表 9~表 11 中数据,计算出检验分析结果如表 12 所示,通过表 12 可知,所有 p -值均小于 0.05,表明 IBGSOMEPE 与其他集成剪枝方法之间具有显著性差异.从而验证了 IBGSOMEPE 的显著性和有效性.

表 12 Wilcoxon 符号秩检验结果分析

方法比较	p -值	是否显著
IBGSOMEPE vs Bagging	0.0003	是
IBGSOMEPE vs Kappa	0.0017	是
IBGSOMEPE vs AGOB	0.0005	是
IBGSOMEPE vs POBE	0.0017	是
IBGSOMEPE vs DREP	0.0004	是
IBGSOMEPE vs D-D	0.0003	是
IBGSOMEPE vs DF-D	0.0004	是
IBGSOMEPE vs GASEN	0.0004	是
IBGSOMEPE vs GSSEN	0.0004	是
IBGSOMEPE vs MOAG	0.0003	是
IBGSOMEPE vs RRE	0.0072	是
IBGSOMEPE vs DivP	0.0012	是
IBGSOMEPE vs SCG-P	0.0004	是
IBGSOMEPE vs SDAcc	0.0105	是
IBGSOMEPE vs RE	0.0008	是
IBGSOMEPE vs COM	0.0012	是
IBGSOMEPE vs MDM	0.0026	是
IBGSOMEPE vs EP-FP	0.0003	是
IBGSOMEPE vs DFTwo	0.0026	是
IBGSOMEPE vs AccRein	0.0026	是
IBGSOMEPE vs MDEP	0.0004	是

5.4 IBGSO 参数分析

本小节的目的：优化 IBGSO 的主要参数，为 IBGSOMEF 中所运用的 IBGSO 的参数选取提供依据。

在本文提出的 IBGSOMEF 中，使用了 IBGSO 进行二次剪枝，为了提高算法的搜索效率，下面分析算法的主要参数：参数 p_1 、 p_2 、概率因子 τ_1 、 τ_2 、最大迭代次数、种群规模、决策域半径初始值和最大值。鉴于篇幅的限制，本节实验仅在 Diabetes 和 Spambase 数据集上进行测试，原始基分类器池规模设置成 200，在基于边界最小化测度的预剪枝后，剩下分类器规模为 25，并运用 IBGSO 搜索最优子集成，具体参数分析如下：

为了评估 IBGSO 的搜索效率，本文将 IBGSO 分别与二元萤火虫算法 (BGSO)^[46]、离散型萤火虫算法

(DGSO)^[47]，改进离散型萤火虫算法 (IDGSO)^[39]、扩散二元萤火虫算法 (SBGSO)^[40]、二元人工鱼群算法 (BAFSA)^[48]，改进二元人工鱼群算法 (IBAFSA)^[49]、改进人工鱼群算法 (IAFSA)^[50] 和 GA^[12] 进行了对比分析，如图 3 所示。图 3 展示了在 Diabetes、Spambase 数据集上，对预剪枝后的剩下的分类器，运用不同的二元启发式算法进行二次剪枝的结果分析。上述所有算法的种群规模设置为 25，其余参数均参照原文献设置。通过图 3 可以看出，上述二元启发式搜索算法的收敛速度和精度均与迭代次数成正相关，而 IBGSO 的收敛速度和精度，明显优于 BGSO、DGSO、IDGSO、SBGSO、BAFSA、IBAFSA、IAFSA 和 GA。在迭代次数达到 500 时，若进一步增加迭代次数，算法的性能趋于平稳，建议最大迭代次数设置为 500。

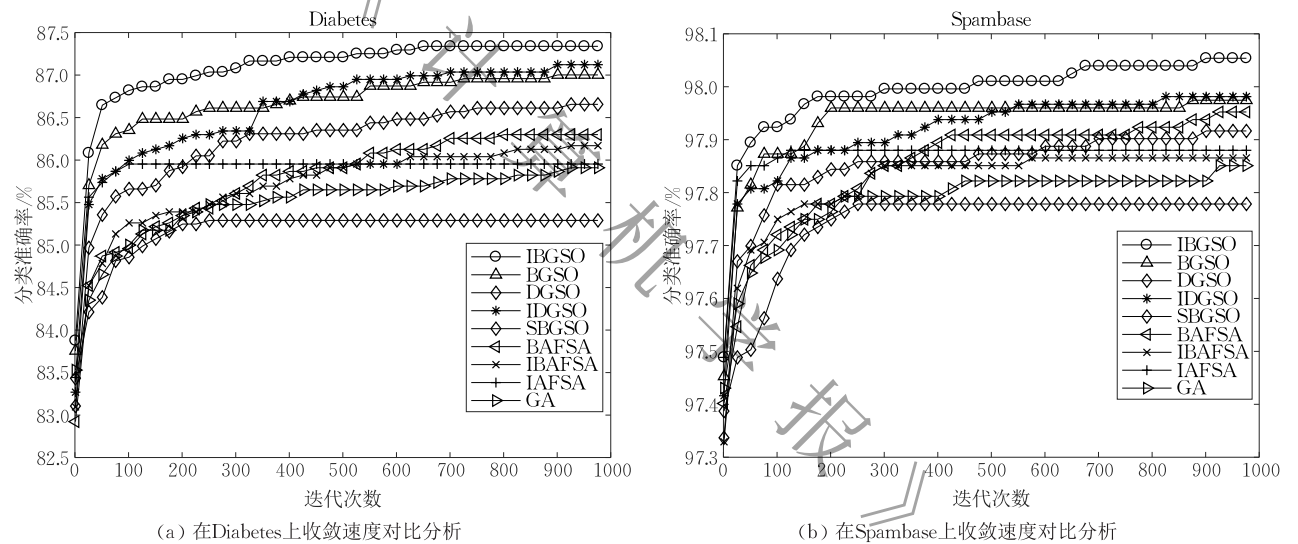


图 3 在 Diabetes, Spambase 上不同二元启发式算法的性能分析

图 4(a)对参数 p_1 、 p_2 进行了分析。我们针对参数 p_1 、 p_2 选取了五组具有代表性的组合：(0.05, 0.95)，(0.15, 0.85)，(0.25, 0.75)，(0.35, 0.65)，(0.45, 0.55)，代表性组合的选择方式借鉴了文献[22]。通过图 4(a)可知，当参数 p_1 、 p_2 选择组合 (0.15, 0.85) 时，IBGSO 的性能最优。建议参数 p_1 、 p_2 取 (0.15, 0.85)。

图 4(b)分析了概率因子 τ_1 、 τ_2 对算法性能的影响。针对概率因子 τ_1 、 τ_2 选取了 7 组具有代表性的组合：(0.05, 0.95)，(0.10, 0.90)，(0.15, 0.85)，(0.20, 0.80)，(0.25, 0.75)，(0.30, 0.70)，(0.35, 0.65)，代表性组合的选择方式借鉴了文献[49]。由图 4(b)可知，当概率因子 τ_1 、 τ_2 取 (0.10, 0.90) 时，IBGSO 的性能最佳。建议概率因子 τ_1 、 τ_2 取 (0.10, 0.90)。

图 4(c)分析了种群规模对 IBGSO 性能的影响。通过图 4(c)可知，随着种群规模的增加，IBGSO 的性能逐步提升；当种群规模超过 25 时，算法性能趋于平稳，进一步增加种群规模，对算法性能提升作用不大，反而会大幅增加算法的计算复杂度。建议种群规模设置为 25。

图 4(d)分析了 IBGSO 的初始决策域半径对 IBGSO 性能的影响。采用边界最小化测度进行预剪枝后的基分类器数目为 25，初始决策域半径范围设置为 1 到 25。若初始决策域半径较小，则决策域内的萤火虫个体较少，算法的收敛速度较慢；若初始决策域半径较大，算法易陷入局部最优。从图 4(d)可知，当初始决策域半径取 11 时，IBGSO 的收敛精度达到最高。建议初始决策域半径设置为 11。

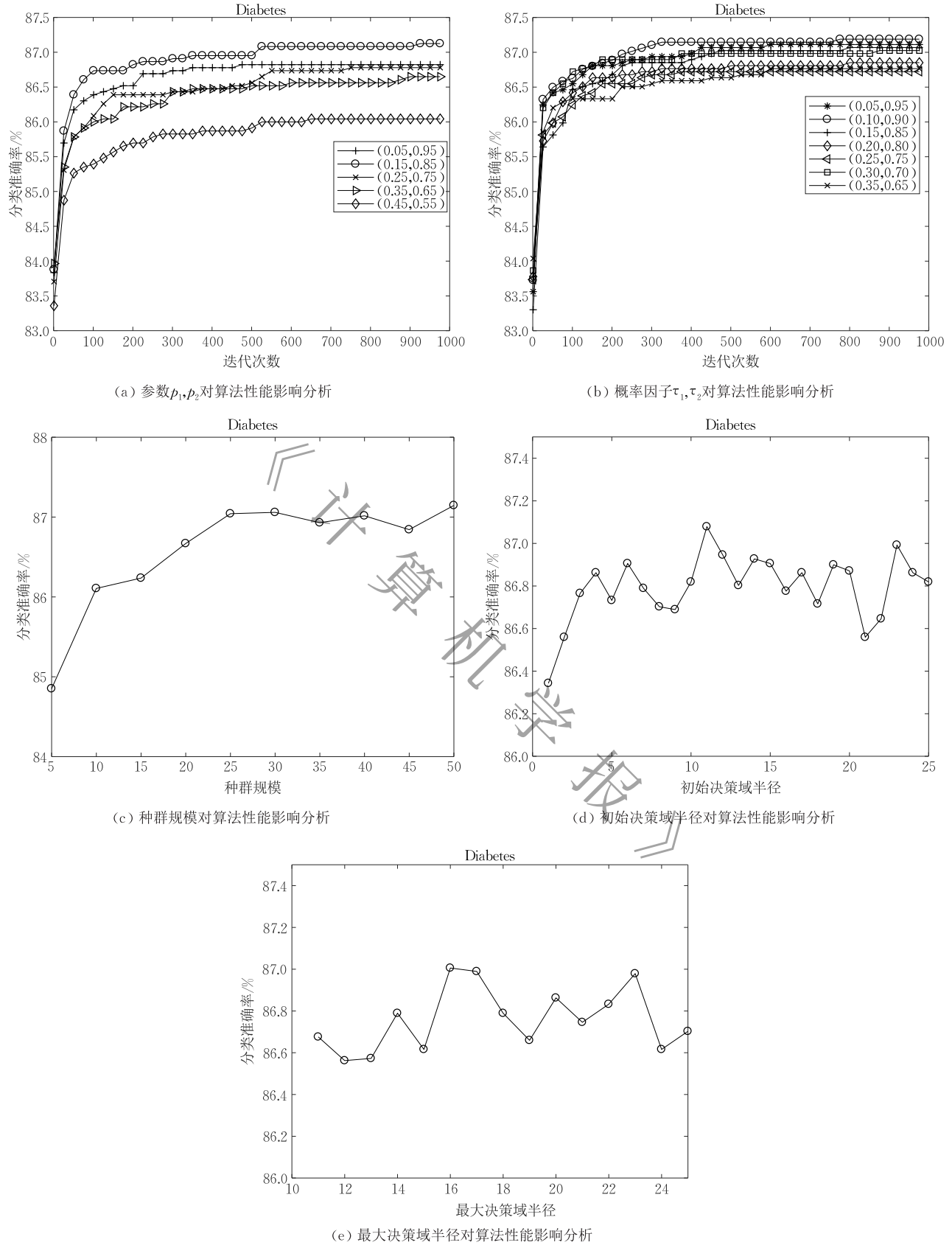


图 4 不同参数对 IBGSO 的性能影响分析

图 4(e)分析了最大决策域半径对 IBGSO 性能的影响. 在 IBGSO 中, 最大决策域半径应大于或等于初始半径, 因此, 最大半径的范围应设置为 11 到 25. 由图 4(e)可知, 在最大决策域半径为 16 时, 算法

的性能达到最优,建议最大决策域半径设置为 16。

6 结束语

分类器的差异性和平均精度对于提升集成分类器的性能十分关键。找到基分类器的平均精度和差异性之间的平衡点,才是集成系统是否成功的关键。而集成剪枝是一个 NP 难问题。为了搜索到集成性能更优的子集成,本文将边界最小化测度和 IBGSO 进行有效融合,提出了 IBGSOMEIP。其为先运用边界最小化测度预剪枝集成系统中部分综合性能较差的分类器,减少了集成系统中基分类器的数目,降低了集成剪枝的计算复杂度;其次,对 GSO 进行了改进,提出了 IBGSO,提高了算法的收敛速度和精度;最后,融合了边界最小化测度和 IBGSO,对预剪枝后的基分类器,进行二次剪枝,较为显著地提高了集成剪枝的性能。在 35 个 UCI 数据集上的实验结果,验证了 IBGSOMEIP 的有效性和显著性,为集成学习中集成剪枝问题的解决提供了较为新颖的研究方向。下一步研究重点是运用其他测度进行预剪枝,为基于二元启发式算法的集成剪枝技术提供综合性能更优的分类器。

致 谢 感谢合肥工业大学管理学院杨善林院士和刘业政教授提供了宝贵的建议,在此向他们表示深深的感谢!

参 考 文 献

- [1] Han H, Liu B. Ensemble of extreme learning machine for remote sensing image classification. *Neurocomputing*, 2015, 149(PA): 65-70
- [2] Mashhoori A. Block-wise two-directional 2DPCA with ensemble learning for face recognition. *Neurocomputing*, 2013, 108(5): 111-117
- [3] Kavitha B, Karthikeyan S, Maybell P S. An ensemble design of intrusion detection system for handling uncertainty using neutrosophic logic classifier. *Knowledge-Based Systems*, 2012, 28(2): 88-96
- [4] Zhou H, Zhao X, Wang X. An effective ensemble pruning algorithm based on frequent patterns. *Knowledge-Based Systems*, 2014, 56(1): 79-85
- [5] Lanitis A, Draganova C, Christodoulou C. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, 2004, 34(1): 621-628
- [6] He D, Wang Z, Yang B, et al. Genetic algorithm with ensemble learning for detecting community structure in complex networks//*Proceedings of the Computer Sciences and Convergence Information Technology*. Seoul, South Korea, 2009: 702-707
- [7] Galar M, Fernández A, Barrenechea E, et al. Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. *Information Sciences*, 2016, 354(C): 178-196
- [8] Zhang Chun-Xia, Zhang Jiang-She. A survey of selective ensemble learning algorithms. *Chinese Journal of Computers*, 2011, 35(8): 1399-1410(in Chinese)
(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)
- [9] 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
- [10] Tang E K, Suganthan P N, Yao X. An analysis of diversity measures. *Machine Learning*, 2006, 65(1): 247-271
- [11] Yang Chun, Yin Xu-Cheng, Hao Hong-Wei, et al. Classifier ensemble with diversity: Effectiveness analysis and ensemble optimization. *Acta Automatica Sinica*, 2014, 40(4): 660-674 (in Chinese)
(杨春, 殷绪成, 郝红卫等. 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)
- [12] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 2002, 137(1): 239-263
- [13] Partalas I, Tsoumakas G, Vlahavas I. Pruning an ensemble of classifiers via reinforcement learning. *Neurocomputing*, 2009, 72(7-9): 1900-1909
- [14] Martínez-Muñoz G, Hernández-Lobato D, Suárez A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(2): 245-259
- [15] Martínez-Muñoz G, Suárez A. Aggregation ordering in bagging //*Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*. Innsbruck, Austria, 2004: 258-263
- [16] Martínez-Muñoz G, Suárez A. Pruning in ordered bagging ensembles //*Proceedings of the 23rd International Conference on Machine Learning*. New York, USA, 2006: 609-616
- [17] Margineantu D D, Dietterich T G. Pruning adaptive boosting //*Proceedings of the 14th International Conference on Machine Learning*. New York, USA, 1997: 211-218
- [18] Guo L, Boukir S. Margin-based ordered aggregation for ensemble pruning. *Pattern Recognition Letters*, 2013, 34(6): 603-609
- [19] Li N, Yu Y, Zhou Z H. Diversity regularized ensemble pruning//*Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany, 2012: 330-345
- [20] Dai Q, Zhang T, Liu N. A new reverse reduce-error ensemble pruning algorithm. *Applied Soft Computing*, 2015, 28(3): 237-249

- [21] Guo H, Liu H, Li R, et al. Margin & diversity based ordering ensemble pruning. *Neurocomputing*, 2018, 275(2): 237-246
- [22] Ni Zhi-Wei, Zhang Chen, Ni Li-Ping. Haze forecast method of selective ensemble based on glowworm swarm optimization algorithm. *Pattern Recognition and Artificial Intelligence*, 2016, 29(2): 143-153(in Chinese)
(倪志伟, 张琛, 倪丽萍. 基于萤火虫群优化算法的选择性集成雾霾天气预测方法. *模式识别与人工智能*, 2016, 29(2): 143-153)
- [23] Rokach L. Collective-agreement-based pruning of ensembles. *Computational Statistics and Data Analysis*, 2009, 53(4): 1015-1026
- [24] de Oliveira J V, Szabo A, de Castro L N. Particle swarm clustering in clustering ensembles; Exploiting pruning and alignment free consensus. *Applied Soft Computing*, 2017, 55(6): 141-153
- [25] Lazarevic A, Obradovic Z. Effective pruning of neural network classifier ensembles//*Proceedings of the International Joint Conference on Neural Networks*. Washington, USA, 2001: 796-801
- [26] Giacinto G, Roli F, Fumera G. Design of effective multiple classifier systems by clustering of classifiers//*Proceedings of the International Conference on Pattern Recognition*. Barcelona, Spain, 2000: 160-163
- [27] Bakker B, Heskes T. Clustering ensembles of neural network models. *Neural Network*, 2003, 16(2): 261-269
- [28] Lu Hui-Juan, An Chun-Lin, Ma Xiao-Ping, et al. Disagreement measure based ensemble of extreme learning machine for gene expression data classification. *Chinese Journal of Computers*, 2013, 36(2): 341-348(in Chinese)
(陆慧娟, 安春霖, 马小平等. 基于输出不一致测度的极限学习机集成的基因表达数据分类. *计算机学报*, 2013, 36(2): 341-348)
- [29] Lu H J, An C L, Zheng E H, Lu Y. Dissimilarity based ensemble of extreme learning machine for gene expression data classification. *Neurocomputing*, 2014, 128(5): 22-30
- [30] Cavalcanti G D C, Oliveira L S, Moura T J M, et al. Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, 2016, 74(C): 38-45
- [31] Ykhlef H, Bouchaffra D. An efficient ensemble pruning approach based on simple coalitional games. *Information Fusion*, 2017, 34(C): 28-42
- [32] Dai Q, Ye R, Liu Z. Considering diversity and accuracy simultaneously for ensemble pruning. *Applied Soft Computing*, 2017, 58(9): 75-91
- [33] Altiparmak F, Gen M, Lin L, et al. A genetic algorithm approach for multi-objective optimization of supply chain networks. *Computers and Industrial Engineering*, 2006, 51(1): 196-215
- [34] Bonyadi M R, Michalewicz Z. Analysis of stability, local convergence, and transformation sensitivity of a variant of the particle swarm optimization algorithm. *IEEE Transactions on Evolutionary Computation*, 2016, 20(3): 370-385
- [35] Wang Z, Xing H, Li T, et al. A modified ant colony optimization algorithm for network coding resource minimization. *IEEE Transactions on Evolutionary Computation*, 2016, 20(3): 325-342
- [36] Chen Y, Zhu Q, Xu H. Finding rough set reducts with fish swarm algorithm. *Knowledge-Based Systems*, 2015, 81(C): 22-29
- [37] Jayakumar D N, Venkatesh P. Glowworm swarm optimization algorithm with topsis for solving multiple objective environmental economic dispatch problem. *Applied Soft Computing*, 2014, 23(5): 375-386
- [38] Ding S, An Y, Zhang X, et al. Wavelet twin support vector machines based on glowworm swarm optimization. *Neurocomputing*, 2017, 225(C): 157-163
- [39] Ni Zhi-Wei, Xiao Hong-Wang, Wu Zhang-Jun, et al. Attribute selection method based on Improved discrete glowworm swarm optimization and fractal dimension. *Pattern Recognition and Artificial Intelligence*, 2013, 26(12): 1169-1178(in Chinese)
(倪志伟, 肖宏旺, 伍章俊等. 基于改进离散型萤火虫群优化算法和分形维数的属性选择方法. *模式识别与人工智能*, 2013, 26(12): 1169-1178)
- [40] Cheng Mei-Ying, Ni Zhi-Wei, Zhu Xu-Hui. Spread binary glowworm optimization combined with rough set for attribute reduction problem. *Systems Engineering and Electronics*, 2016, 38(10): 2449-2456(in Chinese)
(程美英, 倪志伟, 朱旭辉. 融合粗糙集和扩散二元萤火虫算法的属性约简方法. *系统工程与电子技术*, 2016, 38(10): 2449-2456)
- [41] Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 2011, 21(2): 137-146
- [42] Krishnanand K N, Ghose D. Glowworm swarm optimisation: A new method for optimising multi-modal functions. *International Journal of Computational Intelligence Studies*, 2009, 1(1): 93-119
- [43] Krishnanand K N, Ghose D. Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. *Swarm Intelligence*, 2009, 3(2): 87-124
- [44] Krishnanand K N, Ghose D. Glowworm swarm based optimization algorithm for multimodal functions with collective robotics applications. *Multiagent and Grid Systems*, 2006, 2(3): 209-222
- [45] Breiman L. Bagging predictors. *Machine Learning*, 1996, 24(2): 123-140
- [46] Li M W, Wang X, Gong Y. Binary glowworm swarm optimization for unit commitment. *Journal of Modern Power Systems and Clean Energy*, 2014, 2(4): 357-365
- [47] Zhou Yong-Quan, Huang Zheng-Xin, Liu Hong-Xia. Discrete glowworm swarm optimization algorithm for TSP problem. *Acta Electronica Sinica*, 2012, 40(6): 1164-1170(in Chinese)
(周永权, 黄正新, 刘洪霞. 求解 TSP 问题的离散型萤火虫群优化算法. *电子学报*, 2012, 40(6): 1164-1170)

[48] Singhal P K, Naresh R, Sharma V. Binary fish swarm algorithm for profit-based unit commitment problem in competitive electricity market with ramp rate constraints. *IET Generation, Transmission & Distribution*, 2015, 9(13): 1697-1707

[49] Azad M A K, Rocha A M A C, Fernandes E M G P.

Improved binary artificial fish swarm algorithm for the 0-1 multidimensional knapsack problems. *Swarm and Evolutionary Computation*, 2014, 14(2): 66-75

[50] Luan X Y, Li Z P, Liu T Z. A novel attribute reduction algorithm based on rough set and improved artificial fish swarm algorithm. *Neurocomputing*, 2016, 174(2): 522-529



ZHU Xu-Hui, Ph. D. , lecturer. His research interests include evolution computation and machine learning.

NI Zhi-Wei, Ph. D. , professor. His research interests include artificial intelligence, machine learning and cloud computing.

CHENG Mei-Ying, Ph. D. , lecturer. Her research interests include intelligence computation and data mining.

LI Jing-Ming, Ph. D. , lecturer. His research interests include intelligence computation and data mining.

JIN Fei-Fei, Ph. D. candidate. His research interests include intelligence decision and intelligence computation.

WU Zhang-Jun, Ph. D. , associate professor. His research interests include intelligence algorithm and cloud computing.

Background

This work is partially supported by the National Natural Science Foundation of China “Research on Key Techniques in Collaborative Business Intelligence Construction in Big Data Environment” (91546108), “Research on Business Analytics, Computing Techniques and Platforms in the Context of Big Data” (71490725), the National Key Research and Development Plan (2016YFF0202604), the Anhui Provincial Natural Foundation (1908085QG298, 1708085MG169), the Social Science Research Project of Education Department of Anhui Province (JS2017AJRW0135), and the Natural Science Foundation of Huzhou (2018YZ11).

As we know, diversity among base classifiers is very important to a successful ensemble. A collection of base classifiers which perform badly in an ensemble can lead to the decrease of the ensemble performance in classification. In addition, as the increasing of the size of data sets, an ensemble composed of massive classifiers incurs the increase in memory space and computational cost, which motivates the appearance of ensemble pruning. Ensemble pruning is an NP-complete problem. To address this issue, margin distance minimization performs well when it comes to measuring the diversity of classifiers and pre-pruning the classifiers with worse performance and less diversity, which can downsize the classifiers in a constructed initial pool, and significantly reduce the computational complexity of ensemble pruning. The final ensemble extracted from the retaining classifiers after pre-pruning is efficiently achieved using a heuristic algorithm.

Considering that Glowworm Swarm Optimization (GSO) is a heuristic algorithm with many good advantages, such as less parameters, easy implementation, strong robustness. GSO can be used as a strategy to search for the optimal sub-ensemble extracted from the retaining classifiers after pre-pruning. To make GSO search for the optimal solution in a binary space, the Improved Binary Glowworm Swarm Optimization (IBGSO) is proposed by improving the moving way of glowworms and the searching processes. The proposed IBGSO is employed to further prune the retaining classifiers after pre-pruning, and the final ensemble is achieved.

Based on the above analysis, Improved Binary Glowworm Swarm Optimization combined with Margin distance minimization for Ensemble Pruning (IBGSOMEPP) is proposed using the combination of IBGSO and MDM. Experimental results on 35 UCI datasets demonstrate that the proposed approach obviously outperforms other state-of-the-art ensemble pruning approaches with less number of base classifiers.

The key contributions of this work are described as follows: (1) A novel approach for ensemble pruning is proposed using the combination of IBGSO and MDM. (2) We find that MDM can do well with respect to pre-pruning the classifiers with less diversity and worse predictive ability. (3) The proposed IBGSO can achieve a good result when it comes to searching in a binary solution space. (4) Experiments on 35 UCI datasets demonstrate the effectiveness and efficiency of the proposed IBGSOMEPP. (5) It provides a novel research idea for ensemble pruning.