

进化集成学习算法综述

胡毅¹, 瞿博阳², 梁静¹, 王杰¹, 王艳丽¹

(1. 郑州大学电气工程学院, 河南 郑州 450001; 2. 中原工学院电子信息学院, 河南 郑州 450007)

摘要: 进化集成学习结合了集成学习和进化算法两方面的优势, 并在机器学习、数据挖掘和模式识别等领域被广泛应用。首先对进化集成学习算法的理论基础、组成结构及分类情况进行了概述。然后根据进化算法在集成学习中的优化任务, 从样本选择、特征选择、集成模型参数组合优化、集成模型结构优化以及集成模型融合策略优化几个方面对当前进化集成学习领域的一些代表性研究成果进行了详细的综述, 并分析和总结了各种进化集成学习算法的特点。最后对现有的进化集成研究的优缺点进行探讨, 并给出了未来的研究方向。

关键词: 集成学习; 机器学习; 进化算法; 分类; 回归; 聚类

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202103

A survey on evolutionary ensemble learning algorithm

HU Yi¹, QU Boyang², LIANG Jing¹, WANG Jie¹, WANG Yanli¹

1. School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China

2. School of Electronic and Information, Zhongyuan University of Technology, Zhengzhou 450007, China

Abstract: Evolutionary ensemble learning integrates advantages of ensemble learning and evolutionary algorithm and is widely used in machine learning, data mining, and pattern recognition. Firstly, the theoretical basis, formation, and taxonomy are introduced. Secondly, according to the optimization task of evolutionary algorithm in ensemble learning, some representative researches on evolutionary ensemble learning field were analysed from the aspects of instance selection, feature selection, parameter optimization, structure optimization, and fusion strategy optimization, and the characteristics of different evolutionary ensemble learning methods were summarized. Finally, the pros and cons of the current researches on evolutionary ensemble learning were analysed, and research directions in the future work were given.

Key words: ensemble learning, machine learning, evolutionary algorithm, classification, regression, clustering

1 引言

作为当前人工智能 (artificial intelligence, AI) 发展的主要方向之一, 机器学习 (machine learning, ML) 越来越多地受到人们的关注。由于单个学习器的泛化能力或鲁棒性往往较差, 因此一些研究将

多个学习器以一定的策略组合, 构成集成模型, 以提高学习器解决问题的能力。许多研究表明^[1-2], 新构建的集成学习 (ensemble learning, EL) 系统往往可以获得比单一学习器更优越的泛化性能或鲁棒性, 并且可以在保证集成学习系统性能的前提下, 选择一组最佳子学习器来提高学习系统的效

收稿日期: 2020-08-14; 修回日期: 2020-08-25

通信作者: 梁静, liangjing@zzu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61922072, No.61976237, No.61876169, No.61673404, No.61806179); 河南省高等学校重点科研项目 (No.18A470015, No.20B120002); 中国博士后科学基金资助项目 (No.2017M622373, No.2018M630835)

Foundation Items: The National Natural Science Foundation of China (No.61922072, No.61976237, No.61876169, No.61673404, No.61806179), The Key Scientific Research Projects of Colleges of Henan Province (No.18A470015, No.20B120002), China Postdoctoral Science Foundation (No.2017M622373, No.2018M630835)

率, 这些优势使集成学习逐渐成为机器学习领域的研究热点, 美国人工智能协会(American Association Artificial Intelligence, AAAI)前主席 Thomas G Dietterich 教授还曾将集成学习、可扩展机器学习(例如对大数据集、高维数据的学习等)、强化学习、概率网络列为机器学习的四大研究方向^[3]。近些年, 随着集成学习的不断发展, 其已被广泛应用于数据挖掘、自然语言处理、搜索引擎、图像处理、模式识别等多个领域。

集成学习方法最早被应用于解决分类问题, 之后这种整合多个子学习器、利用群体学习器的输出来补偿误差、提高整个学习系统的泛化能力或鲁棒性的学习方法, 逐渐被拓展到解决回归和聚类等问题中。然而, 集成学习中不同子学习器的多样性与求得解的准确性或鲁棒性之间往往存在冲突, 这种冲突一般可以通过调整子学习器自身结构参数或群体学习器的融合权值等方法解决, 这明显是一个优化问题。与传统优化方法相比, 遗传算法(genetic algorithm, GA)、粒子群优化(particle swarm optimization, PSO)算法和差分进化(differential evolution, DE)算法等一系列进化算法(evolutionary algorithm, EA)具有适用性广、鲁棒性强、可以实现全局优化等特点。因此, 一些学者将集成学习与进化算法结合, 组成进化集成学习(evolutionary ensemble learning, EEL)算法。

进化算法很早就被应用于单个学习器的优化任务中, 如优化学习器的参数、内部连接权值、网络结构等。随后, 一些学者逐渐将进化算法引入集成学习中以平衡学习系统的多样性和精度等性能,

并以此来构建进化集成学习系统。进化集成学习算法示意如图1所示。在进化集成学习算法中, 进化算法通常被用于集成学习系统中的样本选择^[4-5]、特征选择^[6-10]或对集成模型包含的参数^[11-21]、网络结构^[22-25]等进行优化, 也有部分工作使用进化优化算法选择一组最佳子学习器^[1,26]或对多个子学习器的融合策略^[27-38]进行优化。

以“ensemble learning”和“evolutionary algorithm”为关键词在 Web of Science 数据库中搜索相关文章, 数目变化趋势如图2所示, 由图2可知, 相关文章数量总体呈增长趋势。随着集成学习处理问题的维度、复杂度的增加(如大数据、云计算和深度网络集成等)以及进化算法的突破和发展(如多模态优化、大规模优化、超多目标优化等), 进化集成学习算法具有越来越广阔的应用前景。

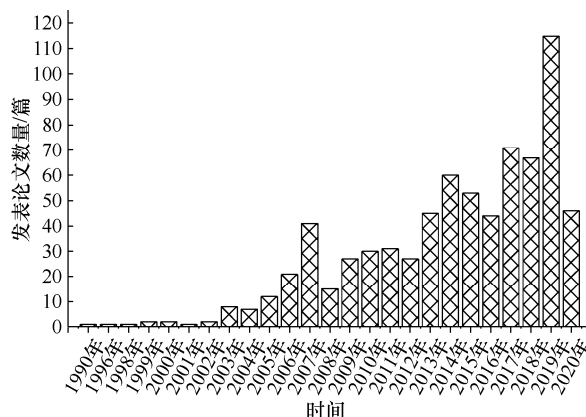


图2 1990—2020年 Web of Science 数据库相关文章数目变化趋势

在当前的研究方法中, 一些研究^[39-41]分别对基于分类、回归和聚类问题的集成学习方法进行了综

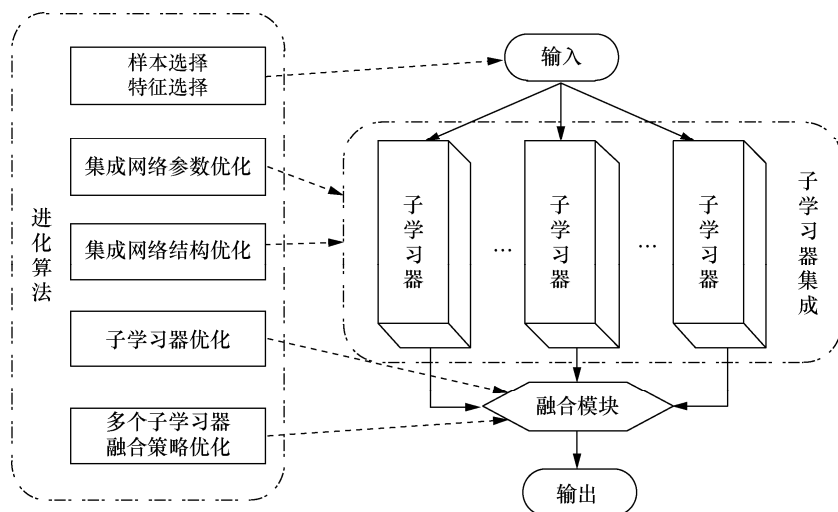


图1 进化集成学习算法示意

述。尽管这些研究对进化集成学习算法进行了简单的介绍,但并未对一些代表性的进化集成学习算法进行详细综述。并且在笔者目前查阅到的文献中,一些关于进化集成学习算法的综述只是针对分类、回归和聚类问题中的一种或两种,尚未发现有学者对包含这3种问题的进化集成学习算法进行全面的综述。分类和回归是两种重要的有监督学习方法,而聚类是一种重要的无监督学习方法,这3种学习方法在机器学习、数据挖掘和模式识别中有非常广泛的应用。虽然这3种学习方法具有不同的特点,然而它们组建进化集成学习模型的方法却类似。近年来,不同的学者基于这3类学习方法分别提出许多相应的进化集成学习算法,提高了解决问题的能力。因此,将分类、回归和聚类问题整合起来,更全面地分析进化集成学习算法的发展、应用和未来方向具有十分重要的意义。

本文第2节对进化集成学习的理论基础与分类问题进行探讨;第3节根据进化算法的优化任务对分类、回归和聚类问题中的进化集成学习算法进行详细的综述;第4节对本文内容进行总结,并展望进化集成学习算法在未来的研究方向。

2 进化集成学习的理论基础与分类

2.1 集成学习的定义与理论依据

集成学习既适用于有监督学习(分类、回归),又适用于无监督学习(聚类),其是通过将一组模型(学习器)以特定方式集成以获得最终预测结果的一个学习过程^[42]。

集成学习使用不同的学习器解决给定的问题,然后结合群体学习器的结果来补偿误差、提高整个学习系统的性能。常用的集成方法有 Boosting^[43]、Bagging^[44]、Dagging^[45]、随机子空间法(random subspace method, RSM)^[46]、Stacking^[47]和基于投票的集成方法^[48]等。Dietterich T G^[49]从3个方面解释了集成学习的有效性:统计学(以平均的方法降低错误假设的风险,从而提高整个学习系统的预测性能)、计算数学(多个子学习器集成输出可以帮助算法逃离局部极值点)与表征方法(当最优假设在单个学习器的搜索空间之外时,通过集成不同子学习器可以扩展搜索空间,从而使最终输出结果更好地匹配数据空间)。

在证实集成学习有效性的各种理论基础中,偏差-方差-协方差分解(bias-variance-covariance

decomposition)被认为是集成学习主要的理论依据之一。以基于回归问题的集成学习为例,其偏差-方差-协方差分解理论的数学表达式如下:

$$E[\bar{y} - z]^2 = \text{bias}^2 + \frac{1}{N} \times \text{var} + \left(1 - \frac{1}{N}\right) \times \text{covar} \quad (1)$$

$$\text{bias} = \frac{1}{N} \sum_i (E[y_i] - z) \quad (2)$$

$$\text{var} = \frac{1}{N} \sum_i E[y_i - E[y_i]]^2 \quad (3)$$

$$\text{covar} = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} E[y_i - E[y_i]](y_j - E[y_j]) \quad (4)$$

其中, bias、var 和 covar 分别表示偏差、方差和协方差; E 表示期望; y_i 为第 i 个子学习器的输出结果, $i=1, 2, \dots, N$, N 为集成学习系统中子学习器的个数; z 表示期望输出值。式(1)为生成误差表达式,右边第一项用于衡量子学习器实际预测值与期望值的平均差,第二项用于衡量子学习器的平均变化,第三项用于衡量不同子学习器的相互成对差异。式(2)~(4)分别为偏差、方差和协方差的分解式。

由式(1)可知,当 covar 为负实数时,整个集成学习系统就可以在保持 bias 和 var 不变的情况下降低生成误差。此外,随着 N 的增加, var 在生成误差中所占的比例会不断降低,而 covar 的重要性会逐渐增加。因此,提高整个集成学习系统的关键在于设计相关性低的子学习器,即提高子学习器的多样性。

然而,上述生成误差的分解表达式并不能被直接应用于基于分类和聚类的集成学习系统中。因此,针对离散类标签问题,一些学者提出不同的偏差-方差分解方法^[41,50],这些方法从一定程度上补充了集成学习的理论依据。此外,强相关^[44]、随机判别^[51]和边界理论^[52]等也是集成学习方法奏效的理论基础,这些理论的解析与偏差-方差-协方差分解类似。

2.2 集成学习的组成

集成学习系统主要包括集成学习的生成和子学习器的融合两部分。集成学习的生成主要通过一定的技术生成一系列“好而不同”的子学习器。其中,集成学习模型的多样性对于学习系统的预测性能至关重要,这种多样性主要包括如下几种。

(1) 数据多样性。将源数据划分为多个数据集,并分别用于测试不同的子学习器(划分后的数据集

必须具有差异度），从而使子学习器的决策具有多样性。常用的集成方法有 Bagging、AdaBoost、RSM、随机森林（random forest, RF）等。

（2）参数多样性。当参数被设置得不同时，即使是相同的学习器，其输出结果的差别也很大。如支持向量机（support vector machine, SVM）、多核学习（multiple kernel learning, MKL）方法和 K 均值聚类算法（ K -means clustering algorithm, K -means）等。因此，调整子学习器的内核参数与组合参数也可以达到增强学习系统的目的。

（3）结构多样性。给予学习器增加内核或在其损失函数中增加去相关惩罚因子，以增加自身结构的多样性。学习器间的结构组合有同质学习器和异质学习器组合结构、样本特征组合和学习器组合结构、串行组合和并行组合结构、在线集成学习结构等。相同子学习器采用不同的组合结构也会对整个集成学习系统的输出产生影响。

（4）其他。基于不同的特征子集或不同问题构建学习模型以及在不同子空间将学习目标进行投影等方法，在一定程度上也会影响集成学习系统的性能^[41]。

为了提高集成学习系统的多样性，可以采用基于数据处理的方法和基于建模过程处理的方法^[42]。其中，基于数据处理的方法包括训练集子采样（如 Bagging、Boosting 和 AdaBoost 等，而这种方法比较适合一些如决策树（decision tree, DT）、人工神经网络（artificial neural network, ANN）等输出不太稳定的学习器）、输入特征的处理（特征选择和特征属性转换等，如随机子空间、输入调制法^[53]、旋转森林和主成分分析等方法）和输出变量处理（如输出调制法^[54]等，这种方法主要针对分类问题）；基于建模过程处理的方法包括学习器参数调整（主要针对如 SVM 和 K -means 等对参数比较敏感的算法）、学习器组合结构调整（主要的组合方式有串行和并行两种）和学习器自身结构调整，如调整学习器遍历假设空间的方式，引导学习器的收敛路径（给 SVM 增加一个内核，使其能够遍历假设空间的所有可能性^[55]），或者结合偏差-方差-协方差分解为学习器增加负相关项^[12]。

多样性和预测性能是集成学习方法的两个重要的指标，但是，对差异度大的子学习器进行组合并不一定能提高学习系统的预测性能^[56]。整个集成学习系统的预测性能与不同子学习器生成误差的

不相关程度呈正相关关系。如何较好地平衡集成学习系统的多样性与预测性能是一个非常复杂的问题，采用经验法在大多数情况下并不能达到最佳效果，而进化优化算法的全局搜索特性为解决两种指标相互冲突的问题提供了契机，因此进化集成学习算法具有十分重要的理论意义和实际应用价值。

集成学习系统的另一个重要组成部分就是子学习器的融合，即将多个子学习器的输出整合为最终的结果。常见的融合策略有多数投票、简单平均、基于权值的投票方法和基于 Median Partition 共识函数的融合策略等。其中，多数投票方法的表达式如下：

$$Y_1 = \operatorname{argmax} \sum_{i=1}^N f_i(x) \quad (5)$$

其中， x 为子学习器的输入量； $f_i(x)$ 为第 i 个子学习器的输出量； Y_1 为集成学习模型的最终输出量。

基于权值的投票方法可以表示为：

$$Y_2 = \sum_{i=1}^N [w_i(x) \times f_i(x)] \quad (6)$$

其中， $w_i(x)$ 为权值方程。当权值的更新采用静态权值组合方法时，融合策略一般为线性组合方法；当 $w_i(x)$ 为 $1/N$ 时，则称之为简单平均；当权值更新采取动态权值组合方法时，集成学习系统会根据不同子学习器的性能（如均方误差、绝对误差等）分配不同的权值。

基于 Median Partition 共识函数的融合策略可以表示为：

$$Y_3 = \operatorname{argmax} \sum_{i=1}^N \Phi[f(x), f_i(x)] \quad (7)$$

其中， Φ 为相似度度量函数，适用性较广的度量指标有样本之间的欧氏距离等，还有一些研究使用基于信息论的测量指标（如类别熵、归一化互信息等指标）作为相似度度量。

针对不同的问题，融合策略也不尽相同。分类问题输出的是类标签、秩序列或概率分布^[57]，集成分类学习的融合方法常使用多数投票或者基于权值的投票方法；回归问题的输出是连续数值，集成回归学习的融合策略常使用基于权值的投票方法；集成聚类学习常使用基于 Median Partition 的共识函数作为融合策略，此外，基于 Objects Co-occurrence 的共识函数也是集成聚类学习问题中常用的融合策略^[41]。

在集成学习方法中，一般的融合策略是将不同

子学习器学习的结果整合输出。然而,一些研究将集成学习系统中子学习器的输出作为特征代入新的学习器中,从而组建新的学习系统,这类融合策略又被称为可训练的集成学习融合策略,其比较适用于对不同样本数据比较敏感的子学习器组合^[58],有关这类融合策略的论述见参考文献[59-60]。

2.3 基于进化算法的集成学习方法

在集成学习方法中,如何平衡集成学习模型的多样性与预测性能是整个学习过程中必须面对的难题。传统的集成学习通过组合不同种类的子学习器来增加集成模型的多样性,但是这种方法过于依赖经验分析,往往难以判断适用于不同问题的学习器的种类。使用传统的数学优化方法调整集成学习模型中的结构参数,从而增加集成模型多样性的方法,具有容易陷入局部最优、计算过程较复杂且受问题性质的限制等弊端。而进化算法具有鲁棒性强、适用性广泛以及不受问题性质限制的特点,尤其是在全局搜索方面优于传统的数学优化方法。基于进化算法在解决复杂优化问题中的优点,许多学者将集成学习与进化算法结合,并组成进化集成学习算法。

在进化集成学习算法中,首先将多个子学习器以 Bagging、AdaBoost 等方法组建成集成学习模型,以样本子集、特征子集、网络参数和融合权值等为编码对象,对种群进行初始化。然后以最小化集成学习模型的预测精度或网络的复杂度,以及最大化网络结构的多样性为优化目标,对种群个体进行适应度评价。如果此时进化算法满足终止条件,则输出算法获得的解,否则采用交叉和变异策略、小生境策略或档案集管理方法等机制,对种群中的个体进行更新,直至达到最大迭代次数或满足其他终止条件。当进化算法包含一个目标函数时,选取最优解代入集成模型,得到最终结果;当进化算法包含多个目标函数时,则需要通过拐点分析、最小曼哈顿距离分析等方法获得期望解,或者保留所有 Pareto 前沿解,并将获得的解代入集成模型,以得到最终结果。针对不同的集成学习模型,可以通过选择合适的进化算法来提高模型的整体学习能力。如果仅考虑提高集成学习模型的预测准确性,则可以采用单目标优化算法。如果同时考虑预测准确性与所建立模型的复杂性,则需选择多目标优化算法。

随着问题难度的不断提高,面临大数据、高维度、分布式计算和云计算等问题时,传统集成学习方法的适应性大打折扣。问题的复杂性增加了集成

学习与进化算法相结合的必要性,并促进产生一系列解决问题的新技术、新理论。许多研究表明,将进化算法与集成学习结合可以大大提高集成学习模型解决问题的能力。因此,进化算法与集成学习相结合是机器学习方法发展的必然趋势。两者相结合可以丰富集成学习的理论,并且在实际应用方面 also 具有重要的意义。

2.4 进化集成学习算法的分类

目前关于集成学习的文章有很多。鉴于集成学习技术的多样性,为了帮助相关学者找到学习的方向以及开发新的集成学习技术和算法,不同的学者提出了各种各样的分类法。Rooney N 等人^[61]根据子学习器是否属于同一种类型将集成学习分为同质和异质两类;针对需要解决的问题,集成学习可以被分为集成分类学习、集成回归学习和集成聚类学习;根据在学习的过程中是否产生新的子学习器以适应数据结构或提高学习系统的整体性能,集成学习可以被分为生成性集成学习和非生成性集成学习^[62];根据是否使用进化算法来优化集成学习系统,集成学习可以被分为进化集成学习和非进化集成学习;根据子学习器的选择和组合,集成学习可以被分为选择性集成学习与非选择性集成学习^[63]等。

本文主要对进化集成学习算法进行综述。下文根据进化算法的优化任务(如数据处理、集成模型参数组合优化、集成模型结构组合优化以及集成模型融合策略优化)总结了一些比较有代表性的进化集成学习算法。本文还在每一类别里分别总结了采用的进化算法的名称、优化问题的分类、目标函数及最终解的选择方法和优化任务信息以及采用的子学习器的名称、属性、集成方法、融合策略及解决的问题等内容。其中,有些作者并未概括所提方法和提出简称,为了更好地对不同进化集成学习算法的特点进行总结,本文对这些无简称的方法按照对应英文名称的首字母或根据算法特点进行命名。需要注意的是,本文所述的集成学习特指由多个子学习器组合集成一个整体学习系统的学习方法,一些包含多种集成策略的进化算法^[64-65]以及由集成了多种数据分析策略的单学习器构成的集成方法^[66-67],虽然也被命名为集成算法,但并不在本文所述的集成学习的范畴内。

3 分类、回归和聚类问题中的进化集成学习算法

3.1 基于数据处理的方法

一些学者使用不同的进化算法对集成学习的

样本进行数据处理。这方面的研究主要集中在样本选择、特征选择和特征构建上。样本选择主要被用于减少样本规模，最大限度地检测并消除冗余和错误的样本信息。特征选择通过一定的方法筛选并剔除数据的冗余特征，以达到降低数据特征空间维度、提高解决问题效率的目的。特征构建则对源数据进行推断，并构建有利于数据分析的新的特征。样本选择、特征选择和特征构建在数据挖掘和机器学习中有重要的意义和广泛的应用。在集成学习系统中，选取最小规模的样本子集或有效特征子集可以最大限度地减轻学习系统的计算负担，提升学习系统的综合学习能力。

针对分类问题中的基于数据处理的进化集成学习研究，参考文献[4]和参考文献[5]都采用多目标优化算法来提高集成模型的预测能力，然而二者的优化任务并不相同。前者主要针对样本选择与特征选择问题，以最大化预测性能以及最小化样本规模为优化目标，且在最终解的选择方案中以曲线下面积（area under the curve, AUC）值为参考。这样的优化方案可以较好地处理不平衡数据分类问题，却忽视了多样性的分析。而后者将 SVM 的结构优化作为其优化任务之一，以提高集成模型的多样性，并且将 SVM 结构优化、参数优化与样本选择同时作为优化任务，在进化的过程中同时兼顾集成模型的特性及数据特征的属性。参考文献[6]则通过集成不同属性的子学习器（SVM、DT）达到增强集成模型多样性的目的，且其集成方法相对于参考文献[4]和参考文献[5]也较为复杂。然而，在进化算法的选择上，参考文献[6]却选择了基础的 PSO 算法，作者并未增加其他进化策略以增强算法的局部搜索能力和解决高维度数据等问题的能力，且以最小化分类错误率为目标函数的单目标优化算法忽视了集成模型的结构属性对特征选择问题的影响。

Liu Y 等人^[7]以 SVM 为子分类器，通过设计一种改进的多目标蚁群优化（MOACO）算法对高维度的实体解析（ER）问题进行处理。参考文献[7]中提出的 EC-ER 算法的目标函数表达式如下：

$$\begin{cases} \max \alpha_i \\ \max 1 - \max_{j=1}^{N-1} \{\beta(\alpha_j, \alpha_i)\} \\ \min \gamma_i \end{cases} \quad (8)$$

其中， α_i 表示第 i 个子分类器的准确率； $\beta(\alpha_j, \alpha_i)$ 表示第 i 个分类器与第 j 个分类器的差异度，差异度

的值通过计算不同分类器预测结果的皮尔逊相关系数得到； γ_i 表示第 i 个子分类器的特征数量。参考文献[7]中设计的多目标优化算法旨在使集成学习系统保持更高的分类准确率和子分类器多样性的同时，获得更少的样本特征数量。在获得最优 Pareto 前沿解后，作者按照优先度梯度的方法获得最终的 Pareto 最优解。优先顺序从高到低分别是准确率最高、差异度最大和特征数量最少，在满足前两个目标的前提下选择最少的特征数量。然而在特征选择中，相同的特征数量对应着多种不同的特征组合，这显然是一个多模态的优化问题^[68]，在实际情况中可以采用多模态优化技术^[69-70]获得同一最佳特征子集下的多种特征组合。这样在保证集成学习模型预测性能和多样性的前提下，赋予了决策者更多选择最佳特征子集的机会。

参考文献[71]和参考文献[72]是进化集成学习的应用性研究，它们分别使用进化集成学习算法解决图像分类问题和激光诱导集成光谱元素定量分析问题。在优化算法上，二者使用了简单的遗传规划（genetic programming, GP）算法或 GA，并且都以最大化预测准确性为适应度函数。不同的是，参考文献[71]针对图像分类问题，使用 GP 算法对原有图像的数据进行特征构建，并生成有利于图像分析的特征。而参考文献[72]的优化任务为使用 GA 从包含噪声的数据中进行特征选择。为了保持集成学习模型的多样性，二者都选择集成多种不同的子学习器。虽然二者建立的进化集成学习模型取得了较好的预测效果，但是 GP 算法和 GA 作为基础的进化算法，在解决复杂的实际问题时，搜索最优解的能力有限。如果加入局部搜索策略或者小生境策略，则可以进一步增强算法的局部搜索能力。

李毓等人^[8]对回归模型的响应变量与协变量进行分析，并指出变量选择旨在去除冗余协变量，找到与响应变量相关度最高的协变量，这显然是一个特征选择问题。基于此，作者提出一种结合了带权值的单路径遗传算法和 Boosting 回归模型的进化集成学习算法（BoostGA），并将此方法用于回归模型的变量选择问题中。在变量选择过程中，首先每个训练个体被赋予不同的权值，然后根据赤池信息量准则（akaike information criterion, AIC）设计目标函数，计算出不同的带权值样本个体对应的适应度值，再通过遗传算法获得部分拥有较大适应度值的样本，最后计算不同变量在所获得的样本中的比

重,进而确定不同变量的重要性,并选择出期望的最佳变量集。然而,文献中所提模型主要采用基于误差分析的单目标遗传优化算法,虽然在变量选择过程中也通过信息熵对集成学习系统的多样性进行分析,但由于分析中带有经验参数设计,限制了该算法在实际应用中的推广。

Jiménez F 等人^[9]使用非支配径向基演化算法(ENORA)对销售预测问题的样本数据进行特征选择。文中包含两个优化目标,一个是最小化选择的特征数量,另一个是最小化随机森林的预测结果与期望值的均方根误差。最终从得到的非支配解中,选择具备最小均方误差的特征子集用于回归预测任务。然而,由式(1)可知,提高不同子学习器的差异度可以提高集成网络的多样性,在集成学习中可以在保持偏差和方差不变的情况下降低生成误差。因此在样本选择和特征选择问题中,加入多样性保持策略或以多样性为指标设计目标函数更有助于找到最佳样本子集或特征子集。

由于缺乏准确的标签,在无监督学习的过程中很难找到合适的准则用于选择样本的最佳特征子集,并且随着特征数量的增加,选择特征子集的难度不断加大,这些都限制了特征选择在无监督学习中的实际应用。在参考文献[10]中,作者以 *K*-means 为子学习器,通过随机子空间方法构建集成聚类学习模型 CEFS。其中,种群增量学习(population based incremental learning, PBIL)算法被应用于集成聚类学习的特征选择问题中。该方法对所有特征进行二进制编码,以最大化聚类算法的产生解与集成解的相似程度为目标进行优化。这种相似程度的表达式如下:

$$S(y, Y) = \frac{2 \times \sum_{i=1}^{\xi-1} \sum_{j=i+1}^{\xi} (1 - |y_{ij} - Y_{ij}|)}{\xi \times (\xi - 1)} \quad (9)$$

其中, ξ 表示要解决问题的类别数目; y 和 Y 分别表示子学习器和集成学习器的输出结果。由于单个聚类算法进行特征选择时难以选择合适的评价准则,常用的 scatter separability 和 DB-index 方法分别容易造成高维和低维数据偏差。Hong Y 等人^[10]通过集成多个聚类算法来提高集成聚类学习的鲁棒性,并利用集成学习模型确定特征选择评价准则,提出基于集成聚类的无偏差特征选择方法。然而所提方法并没有包含对整个集成网络的多样性分析,可以

通过集成多个差异度大的子学习器或者设计多样性指标,并使用进化计算的方法,寻找同时满足相似度和多样性指标的最佳特征子集。

3.2 基于集成学习模型参数优化的方法

集成学习网络中的参数对整个模型的性能有很大影响。需要注意的是,这里的参数主要指集成学习生成阶段子学习器包含的参数或集成模型中包含的其他参数。在机器学习中,即使是相同的学习器,当设置的参数不同时,预测效果也大相径庭。因此,许多研究采用不同的进化优化方法对集成学习中学习器的参数进行优化,以增加集成学习模型参数的多样性,从而提高整个集成学习系统的多样性。常见的集成学习模型中的参数包括神经网络的权值和阈值、SVM 的正则化参数和核参数、径向基等核函数中包括的参数以及集成模型学习过程中人为设置的一些系数(如负相关系数、误差系数)等。

在基于集成学习模型参数优化的研究方法中,按照进化算法优化问题的类别,可以将分类问题中的优化方法分为单目标优化方法和多目标优化方法两类。

基于单目标优化算法的研究有参考文献[15]和参考文献[73]。在参考文献[15]中, Padilha C A 等人以最小二乘支持向量机(LSSVM)为子学习器构建集成分类学习系统(MLGALSSVM),并且以最小化误差为目标函数,采用 GA 同时进行参数优化、融合权值优化、特征选择和最佳子学习器的选择。该研究同时兼顾集成模型的结构特性与数据特征属性,在进化过程中不断地优化整个集成模型。然而,基础的 GA 不能很好地同时处理多个复杂的优化任务。针对不同的优化任务应设计相应的目标函数,因为单目标优化的方法无法很好地处理不同优化任务之间的关联性与冲突性。

Huang T 等人^[73]提出了一种自适应小生境差分(ANDE)算法。该算法通过在 DE 算法中加入启发式聚类算法来构造小生境,并以此来增强算法解决多模态问题的能力。ANDE 算法中还使用自适应的方法对 DE 算法中的参数进行调节。作者还使用 ANDE 算法对神经网络的权值进行优化。基于 ANDE 算法在解决多模态问题中的优势,文献中建立的 ANDE-NNE 可以更有效地发掘更多最优神经网络参数组合。然而文献中建立的集成模型结构相对简单,且该文献缺乏对所建集成模型的多样性分析。

基于多目标优化算法的研究有参考文献[12-14]。

Chen H 等人^[12]以径向基神经 (RBF) 网络为子学习器设计集成学习算法, 从而解决分类问题。提出的 MRNCL 算法通过带有适应度共享和分配的非支配排序方法设计一种多目标进化优化算法, 并将该算法用于优化 RBF 网络的中心、宽度和隐含层到输出层的权值。优化算法的适应度函数的表达式如下:

$$\begin{cases} \min \sum_{j=1}^R (f_i(x_j) - z_j)^2 \\ \max \sum_{j=1}^R (f_i(x_j) - Y)^2 \\ \min \sum_{j=1}^R \omega_j^2 \end{cases} \quad (10)$$

其中, R 表示集成模型中训练样本的数量; f_i 表示第 i 个子学习器的输出值; x_j 为子学习器第 j 个训练样本的输入量; Y 为集成学习模型的最终输出量; z_j 表示第 j 个训练样本的期望输出值; ω_j 表示第 j 个训练样本对应的 RBF 网络输出层的权值。3 个计算式分别表示最小化单个神经网络中训练数据的均方误差、最大化不同神经网络间的负相关度和最小化集成网络的正则化值。负相关学习主要是在集成学习的生成阶段给每个子学习器 (主要是神经网络) 的误差函数增加一个惩罚函数。新增的惩罚函数可以在不同的子学习器间形成一个负相关的误差, 从而增加不同子学习器的差异化程度, 达到增强集成学习模型多样性的目的。Chen H 等人^[12]还在负相关集成神经网络的基础上加入正则化分析, 以降低算法的复杂度, 提高算法处理带噪声数据的能力。文献中分别针对集成学习模型的预测准确度、多样性和网络复杂度来设计优化目标。然而, 在使用多目标优化算法获得满足优化目标的非支配解后, 文献中选择集成所有非支配解对应的子学习器, 并采用平均权值的融合方法。研究表明, 有选择性的集成子分类器往往能获得更好的集成学习模型^[27,29]。

盛伟国等人^[13-14]利用加入小生境技术的遗传算法对集成神经网络的权值、阈值进行编码, 以最小化均方误差和负相关误差的合并值为优化目标对集成分类学习算法进行优化。虽然建立的集成模型采用负相关学习技术增加模型的多样性, 然而, 加权合并的方法同时增加了两个目标的合并系数, 其并未充分考虑两个目标的特性, 无法准确得到使两个优化目标均达到相对最优的解集, 其本质上仍

属于单目标优化。

在使用基于集成学习模型参数优化的方法来解决回归问题的研究中, 采用单目标优化算法的有参考文献[16]和参考文献[18]。在参考文献[16]中, Yao X 等人使用改进的进化规划 (EP) 算法对集成神经网络中的负相关惩罚系数进行优化, 并提出 NNE-TFP 算法。胡梦月等人^[18]使用加入局部搜索和莱维飞行策略的蝙蝠优化算法 (BA) 对 KELM 算法中的回归模型参数和核参数进行优化。二者均采用进化集成回归学习方法解决实际问题 (电信网络流量预测及风电预测)。针对所解决问题的特性, 二者均设计了对应的集成学习模型, 前者采用负相关学习的方法增加模型的多样性, 而后者采用改进的 AdaBoost.RT 算法构建集成预测学习算法 Ada-KELM。并且二者分别使用改进的 EP 算法与 BA 提高进化算法适应特定问题的能力。

Jin Y 等人^[17]分别使用动态加权法 (DWA) 和 NSGAII 对 MLP 组成的集成学习方法的神经元连接结构和权值进行多目标优化。不同的是, DWA 是将最小化均方误差和神经网络的正则化信息的加权合并值作为一个优化目标, 而 NSGAII 采用非支配排序的方法分别对两个目标函数进行优化。最后, 文献中分别对比了两种优化算法对集成神经网络的结构和参数的优化效果。其中, Ackley 函数被用于验证所提出的集成学习算法 NNRE-MOEA 在解决回归问题中的性能。文献中虽然对集成网络的复杂度进行了分析, 却未以最大化多样性为优化目标进行参数寻优。同样对集成分类学习模型中的神经网络权值进行多目标优化的还有 DIVACE^[11]等方法。

Bui L T 等人^[19]分别采用 NSGAII 和基于非支配排序的 DE 算法对以 BP 算法为子学习器的集成回归学习系统进行多目标优化。其中, 优化算法的编码对象是神经网络权值, 优化目标是最小化均方误差和最大化集成系统多样性。经过多目标寻优后得到非支配解集, 作者分别通过选择全部解集和只选择最优解集两种方式将对应的 BP 算法代入 Adaboost 模型中, 组建集成学习模型 Adaboost-FET, 并将其用于解决货币汇率问题。虽然建立的 Adaboost-FET 模型在测试问题中取得了较好的预测效果, 然而 BP 算法需要经过不断迭代来训练网络参数, 以达到较好的预测效果, 这个过程会耗费大量时间, 在集成模型中这个弊端更加明显。在子学习器的选择中,

可考虑 ELM 等单层神经网络,在不影响预测效果的前提下提高集成学习算法的效率。

聚类中心个数直接影响整个聚类学习器的预测性能。因此,通过进化算法来确定集成聚类学习模型中的聚类中心个数具有十分重要的意义。

参考文献[74]中使用模糊 C 均值聚类 (FCM) 算法作为子学习器,将链、超图结构和组织 P 系统 (tissue-like P system) 组合成集成聚类学习模型,融合策略采用概率轨迹积累 (PTA) 和基于概率轨迹的图划分 (PTGP) 两种共识函数。为解决聚类中心个数设置问题,作者采用 NSGAI、非支配邻域免疫算法 (non-dominated neighbor immune algorithm, NNIA) 和 Pareto 包络选择算法 (Pareto envelope-based selection algorithm, PESA) 3 种多目标优化方法,分别对集成聚类学习模型中的子系统进行优化。虽然所建模型在 UCI 测试集中取得了较好的效果,然而采用的 3 种优化算法均为基础优化算法,随着数据维度的增加,算法的搜索能力会逐渐退化或者限于局部最优。并且随着优化目标的增加,算法找到最优 Pareto 前沿解的能力会逐渐减弱。在解决复杂的实际问题时,应该根据要解决的问题加入局部搜索策略,以提高算法的局部搜索能力,或者加入决策变量的分析策略,以提高算法解决大规模优化问题的能力。

在参考文献[20]中,Handl J 等人将多目标优化算法——PESA-II 与 K-means 和最小生成树 (MST) 结合,提出一种名为 MOCK 的进化集成聚类算法。文献中并没有对聚类中心个数或样本对应的标签进行编码,而是采用一种基于轨迹邻接的编码方案,直接对样本形成的图形结构进行编码,最终在进化的过程中确定集成模型中包含的重要参数——聚类中心个数。Mohammadi M 等人^[21]提出一种使用 GA 在进化过程中确定关联矩阵和集成聚类结果的算法。通过使用进化算法,提出的 GACEII 可以在进化的过程中确定聚类算法中的重要参数,即聚类中心个数。上述方法均使用进化算法间接地实现对聚类中心的寻优,同时对聚类中心个数与样本标签等进行编码寻优的方法也值得更深入的研究。集成聚类学习模型还包括聚类学习器本身包含的一些参数 (如基于密度的聚类算法中关于密度的阈值等),更广泛的研究还包括采用进化算法对这些相关参数进行优化,以提高集成模型的整体性能。

3.3 基于集成学习模型结构优化的方法

结构的多样化是集成学习模型多样性的重要体现。对集成学习模型的网络结构进行优化包括对子学习器的结构优化以及对整个集成网络子学习器的组合优化。前者主要指利用进化算法优化神经网络或决策树等子学习器的节点分布及其对应的激活函数和学习规则等。后者可以被理解为采用进化算法在进化过程中选择最优的子学习器组合。

在基于集成学习模型结构优化的进化集成学习算法中,一些研究^[22,25,75]以神经网络为子学习器组建集成学习模型,通过进化算法对神经网络的结构 (节点分布) 进行优化,并且这些研究均通过改善子学习器的结构来提高整个集成学习模型的多样性。参考文献[22]和参考文献[25]均以最小化误差和集成网络的复杂度 (网络节点连接次数) 为目标进行优化。参考文献[22]将两个目标加权合并为一个目标函数,然后采用单目标优化算法进行优化。这样的方法增加了加权系数,但不能像多目标优化算法那样得到多个可行解。参考文献[75]虽然采用了多目标优化的方法,但是其针对不同的误差进行优化目标的设计,优化过程中并没有考虑集成学习模型的多样性或复杂度等性能指标。

Nag K 等人^[2]设计了一种基于决策树的集成分类器,并将多分类问题转化为二分类问题进行求解。作者以最小化假正 (false positive, FP) 和假负 (false negative, FN) 的数量,以及最小化决策树的树叶节点为优化目标对决策树结构进行优化。由于文献中以最小化 FP 和 FN 的数量为优化目标,设计的进化集成学习算法 ASMiGP 更加适合不平衡数据的分类问题,这点在实验分析阶段得到了验证。与其他集成方法不同,所提方法并未将集成学习系统的多样性作为一个优化指标参与优化过程,而是通过保障被划分为二分类问题的数据特征的差异性来保证整个集成系统的多样性。实验结果显示,该方法可以增强学习系统的多样性,然而,该方法在处理拥有较多样本的问题时会造成巨大的计算消耗。

Ojha V K 等人^[23]使用多目标遗传规划 (MOGP) 算法对子学习器的节点分布和激活函数类别进行优化。文献中选择神经网络和决策树两种不同的学习器组建一种名为异质柔性神经树 (HFNT) 的集成学习系统。Ojha V K 等人还使用 HFNT 解决分类、

回归和时间序列预测问题。在 HFNT 中，多层反馈神经网络的节点可以设定不同的激活函数，并且不同的节点还可以跨层连接。Ojha V K 等人以最小化学习器学习结果的错误率、最小化树的规模和最大化多样性指标为优化目标，对 HFNT 的网络节点和不同节点对应的激活函数种类进行寻优。3 个优化目标的数学表达式如下：

$$\begin{cases} \min \delta = \frac{1}{N} \sum_{i=1}^N (y_i - z_i)^2 \\ \min \text{tr} \\ \max \varepsilon = \frac{\text{num}_d}{\text{num}_a} \end{cases} \quad (11)$$

其中， y_i 表示第 i 个子学习器的实际输出值； z_i 表示第 i 个子学习器的期望输出值； δ 表示集成学习系统输出值与期望值的均方误差； tr 表示神经树中除根节点之外的节点个数； ε 表示 HFNT 的多样性指标，其值等于互异的树的个数 (num_d) 与所有树的个数 (num_a) 的比值。其中，如果两个树的节点数目不同、节点对应的激活函数不同或者输入的样本特征不同，则称两个树互异。确定了 HFNT 的结构后，文献中还采用 DE 算法对集成网络中的权值进行微调。在最后的组合阶段，融合策略采用多数投票的方法。文献中所提方法在针对测试函数的训练中取得了良好的效果，然而并没有在实际的问题中进行验证。

采用进化算法选择最佳子学习器组合的方法有参考文献[26]和参考文献[76]。Min S H^[26]以最大化集成学习模型分类准确度为目标函数，使用 GA 对集成学习算法中的子学习器进行二进制编码，以寻找最佳子学习器组合。文中将 K 近邻 (KNN)、逻辑回归 (LR)、DT 和 SVM 4 种方法作为子学习器，通过随机子空间的方法构建异质集成分类学习模型 GAHRS，并将之用于处理公司破产预测问题。其中，子学习器的输出采用多数投票法进行整合，以获得集成学习系统的输出。虽然文献中使用互异的子学习器组建集成学习模型，以增加网络的多样性，但是该方法过于依赖经验分析，而针对不同的问题，需要选取不同的子学习器，并且不同的集成模型对子学习器的要求也不同，如预测结果随机性较强的子学习器更适合以 Bagging 模型组建集成学习系统。

Liu W 等人^[76]以多个基于信念规则 (BRB) 的分类器构建集成分类学习模型。作者分别采用 DE 算法和 PAES 算法来优化 BRB 的参数并进行子学习

器的组合优化。虽然使用的进化算法在一定程度上提高了集成模型的预测能力，然而作者并未根据具体的问题对优化算法进行改进，以提高搜索能力。并且作者并未对构建的集成学习模型进行整体分析，而是分别使用两种优化算法对子学习器的内部参数以及子学习器组合进行优化。

在参考文献[77]中，Faceli K 等人根据多目标优化算法可以同时利用多个聚类评价准则，以获得多个最优解，并且集成多个聚类算法使获得的解具有更好的鲁棒性，将多目标优化算法与集成学习结合，提出一种多目标聚类集成 (MOCLE) 算法。在所提算法中，Faceli K 等人分别根据 NSGAI 和强度 Pareto 进化算法 (SPEA) 设计了两种适用于集成聚类的多目标进化算法。其中，Faceli K 等人将所有子学习器 (如 K-means、单连接算法 (SL)、平均连接算法 (AL) 和共享近邻 (SNN) 算法) 的结构参数以及样本对应的标签定义为初始种群，以最小化总体偏差 (所有样本与其对应的聚类中心的距离的总和) 和连接性指标 (反映相邻样本被放置到同一个聚类中心的频率) 为优化目标。与其他集成聚类不同的是，MOCLE 算法并不是将所有子学习器的学习结果整合到一起，然后通过共识函数进行输出，而是设置了具有共识函数作用的交叉算子。因此，MOCLE 算法可以在进化的过程中直接输出集成结果。然而，MOCLE 算法在设置初始种群时需要子学习器的一些设置 (如聚类数目等) 进行经验分析，以提高集成模型的多样性。所提优化算法旨在从初始设定的集成模型中寻找最佳集成效果，而不是对所有可能存在的子学习器组合进行寻优，因此，Faceli K 等人删除了变异的过程。这种操作虽然减轻了计算负担，却在一定程度上使算法失去了全局搜索的能力。Liu R 等人^[78]在 MOCLE 算法的基础上增加了基于不同聚类结果相似度的目标函数，以增强原有算法的预测能力。

Mahmood A 等人^[1]针对网络视频分类问题提出了基于 GA 的集成聚类学习算法，以提高预测结果的鲁棒性、稳定性和准确性。提出的 SS-EE 算法以子学习器 (图聚类 (GC) 算法、仿射传播聚类 (AP) 算法和谱聚类 (SC) 算法) 的种类和聚类集成方法 (基于簇的相似度划分 (CSPA)、元的种类聚类算法 (MCLA) 和超图划分算法 (HGPA)) 的种类等信息为编码对象，使用进化算法完成对整个集成学习模型结构的优化。Mahmood A 等人根据视频信

息设计了名为预配对百分比 (pre-paired percentage, PPP) 的指标, 并以最大化 PPP 指标为优化目标指导 GA 的进化过程。PPP 指标的计算主要依据网络视频信息的模型分析, 因此限制了所提集成聚类学习模型的推广应用。

3.4 基于融合策略优化的方法

进化集成学习方法的融合策略主要包括子学习器融合权值优化和共识函数优化。其中, 前者主要针对集成分类学习和集成回归学习, 而后者主要体现在集成聚类学习中。

在基于融合策略优化的进化集成学习方法中, 一些研究^[27,29-30,32,79]根据经验法预先确定子学习器, 通过选择差异较大的子学习器保证集成网络的多样性, 然后采用进化算法对子学习器的融合策略进行分析。将完全不同的几种子学习器的结果合并输出, 虽然保证了集成网络的多样性, 但需要很强的经验分析。因为不同子学习器的适用对象差异很大, 组建集成网络时, 并不能保证所使用的某个或某些子学习器对集成网络的输出起到积极的作用。

另一些研究采用多级优化策略, 首先对集成网络的结构或参数进行优化, 然后采用另一种优化算法在寻优得到的集成网络基础上对子学习器融合权值进行寻优。在参考文献[31]中, Chen Y 等人使用柔性神经树 (FNT) 集成技术对股票市场进行预测。文献中使用 GP 算法对神经树的结构进行优化, 当确定了集成回归学习模型 FNT 的结构后, 通过 PSO 算法实现对网络模型参数的寻优和对融合阶段的权值的优化。多级优化策略虽然采用进化算法调整子学习器的多样性和预测性能, 但是将子学习器融合权值优化与集成网络参数或结构优化分开进行的方法忽略了集成网络优化的整体性。

多级优化策略和根据经验选择子学习器组合优化策略的共同点在于两者在融合权值优化时并不会牵涉到子学习器结构或参数的改变。还有一些研究^[15,28]则将子学习器融合权值寻优与数据处理、网络系数优化或网络结构优化作为一个整体同时进行。

任克强等人^[28]利用人工鱼群算法 (AFSA) 的寻优特性来改进 PSO 算法的局部搜索能力, 然后使用提出的混合优化算法来选择 AdaBoost 模型中的最优子学习器组合及其融合权值。其中, 优化目标为最小化分类误差。最终的集成分类学习模型 AFSA-PSO-AdaBoost 被应用于人脸检测问题中。然

而, 所提算法以最小化分类误差为优化目标, 在进化过程中集成网络的多样性并没有被考虑在内, 集成网络多样性的保持主要依靠 AdaBoost 集成方法本身。

在集成聚类学习中, 子学习器的输出结果不仅像分类和回归问题那样可以通过权值分配的方法进行整合, 还可以通过代入共识函数中获得整个集成学习算法的输出。一般的共识函数主要通过共联矩阵法、超图法、信息论法和进化优化的方法等实现^[45]。因此, 共识函数的设计是不同集成聚类学习算法的主要区别点之一。

在基于进化算法的共识函数优化方法中, Dai H 等人^[80]以多种不同的聚类方法组建异质集成聚类学习模型, 并使用 NSGAII 在进化过程中完成类别的划分。与其他集成聚类方法不同的是, 作者以改动的二元相似聚类集成 (DSCE) 方法作为解的生成方法替代 NSGAII 中的交叉算子, 并以 DSCE 方法中得到的聚类中心数目为输入, 将其代入 K-means 算法中, 最终得到更多高质量的多样性解。参考文献[33]和参考文献[37]通过集成多个不同的子学习器来保持集成网络的多样性, 参考文献[34]将 K-means 算法作为子学习器建立集成学习模型。这些方法使用进化算法并以共识函数或基于共识函数的指标设计优化目标, 从而实现集成聚类学习模型融合策略的优化。还有一些方法 (如 MECEA^[35]、MOMVEC^[36]和 EMEP^[38]等) 以最小化分区总体偏差、连接性指标、同一聚类中心样本的致密性和隶属不同中心样本的相似度等为优化目标, 采用多目标优化算法提高集成聚类学习模型的总体性能, 并将建立的模型用于图形分割、多视图聚类、细胞类型识别等问题中。然而, 这些方法并没有利用进化算法的全局搜索能力对整个集成网络的结构或参数进行优化, 从而改善集成模型的多样性、复杂度或预测性能等。

3.5 混合进化集成学习算法

常见的集成学习方法通过集成多个相同属性的学习器 (如分类、回归或聚类学习器) 来完成有监督或无监督学习。然而, 有些研究却将分类学习器或回归学习器与聚类学习器结合, 共同组建混合集成学习系统, 并通过进化算法改善所建系统的学习能力。

Minku F L 等人^[81]提出一种基于聚类和协同进化算法的集成神经网络模型 CONE。CONE 首先采用进化聚类 (ECM) 算法将数据划分到不同的子空

间中,然后使用基于模糊逻辑的神经网络(EFuNN)对不同子空间中的数据进行分类。由于被划分到不同子空间中的数据的结构不同,导致对应的EFuNN网络结构也互不相同,Minku F L等人分别采用单目标协同进化遗传算法(CEGA)、多目标协同进化遗传算法(MO-CEGA)和多目标协同进化策略(MO-CES)对不同EFuNN的结构或参数进行优化,从而在保证分类精度的同时,优化网络规模、提高计算效率。通过聚类算法对问题进行知识提取的方法增加了集成算法的可解释性,使CONE更适用于解决在线预测问题。

Rahman A等人^[82]提出一种面向聚类的集成分类学习算法NULCOEC。该算法使用一种非均匀分层方法对数据进行划分,每一层包含不同数量的聚类学习器(层次聚类算法)。Rahman A等人分别使用单目标GA和多目标GA,以最大化准确率和多样性为目标对整个集成学习模型的分层层数和每层聚类学习器的数目进行优化。针对每一层的聚类学习器,文献中还使用分类学习器(如SVM)对聚类结果进行决策。最终Rahman A等人使用多数投票的方法对所有层的学习结果进行融合。

在参考文献[83]中,Coletta L F等人将NB和DT两种分类学习算法与经典的聚类学习算法K-means组合成一个集成学习系统C3E-SL。所提方法首先通过集成分类学习模型对训练数据进行学习,然后得到测试数据所属标签的概率分布,再通过集成聚类学习模型对上一步生成的概率分布结果进行重新提炼,进而得到最终结果。整个集成模型的学习过程可以概括为一个优化问题:

$$\min J = \sum_{i \in x} \mathcal{L}(y_i, \lambda_i) + \theta \sum_{(i,j) \in x} s_{ij} \mathcal{L}(y_i, y_j) \quad (12)$$

其中, y_i 表示聚类学习算法为每个样本数据赋予的标签概率; λ_i 表示分类学习算法对第 i 个样本数据学习后得到的标签概率分布; $\mathcal{L}(\cdot)$ 表示损失函数; s_{ij} 表示聚类算法共识函数联合矩阵中包含的元素; θ 表示一个正实数系数。该式子的两个组成部分分别表示集成分类学习部分和集成聚类学习部分的损失函数值。当损失函数 $\mathcal{L}(\cdot)$ 选择平方损失时,上述优化问题可以表示为:

$$\min J = \frac{1}{2} \sum_{i \in x} \|y_i - \lambda_i\|^2 + \theta \frac{1}{2} \sum_{(i,j) \in x} s_{ij} \|y_i - y_j\|^2 \quad (13)$$

在所提集成学习方法中,整个集成学习系统学习的过程即式(13)逐渐收敛的过程。其中,由式(13)

可知,系数 θ 在整个集成学习系统中起到连接集成分类学习模型和集成聚类学习模型的作用。基于系数 θ 对集成学习模型的影响,Coletta L F等人^[83]采用DE算法,以最小化整个集成系统的分类误差为优化目标,对 θ 和集成学习系统收敛过程中的迭代次数进行了优化。

参考文献[84]和参考文献[85]均使用单目标优化算法对混合集成学习模型进行优化。不同的是,前者的优化任务为获得聚类中心个数和子学习器的最优组合,而后者对集成模型中极限学习机(ELM)的权值参数进行优化。二者都使用聚类算法和分类算法构成混合集成学习模型,然而聚类算法在二者中的作用却不相同。前者使用基于PSO算法的进化聚类方法对数据进行聚类,并生成随机子空间,而后者则以多维缩放K均值聚类(MSK)算法为融合策略,对ELM集成网络的输出结果进行聚类。两种混合集成学习模型最终被分别用于解决分类和回归预测问题。虽然参考文献[84]和参考文献[85]均将聚类方法与分类或回归学习器结合起来提高混合集成模型的学习效果,然而,在优化算法的选择上二者均采用简单的进化算法,参考文献[85]虽然将PSO算法用于优化PIO算法,以提高算法的全局搜索能力,但这种方法增加了算法的复杂度,并且改进后的PIO算法的优化效果也没有得到验证。

在参考文献[86]中,Onan A等人使用集成聚类算法对整个集成模型中的集成分类学习器进行组合优化。提出的EPACMOEA首先使用多种分类学习器(5种贝叶斯分类器、14种基于核函数的分类器、10种不同的KNN分类器、3种基于规则的分类器和8种基于DT的分类器)组建集成分类学习模型,并对文本情感分析数据进行分类,然后使用自组织映射(SOM)、最大期望(EM)算法和K-means++对建立的集成分类模型进行聚类分析,同时结合ENORA对建立的集成分类模型进行子学习器组合结构优化,最终通过Q统计指标选择多样性更好的分类器。

有监督学习器与无监督学习器结合的方法主要分为以下3类:一是使用聚类学习器进行知识提取,将数据划分到不同的子空间,然后结合分类或回归算法进行有监督学习,如参考文献[81]和参考文献[84];二是使用聚类算法对有监督学习器的输出结果进行聚类分析,以提高整个学习系统的预测性能或计算效率,如参考文献[85]和参考文献[86];

三是使用有监督学习器对聚类的结果进行决策分析,通过有监督学习器来提高聚类算法的学习能力,如参考文献[82]和参考文献[83]。在混合进化集成学习中,进化算法分别被应用于集成模型结构优化、参数优化和子学习器组合优化等。然而,采用进化算法对集成模型的融合权值进行优化、多目标优化方法最终解的选择方法以及设计(包括集成模型稀疏度、复杂度在内的优化目标)等内容仍值得进行更广泛和深入的研究。并且这类混合进化集成学习方法大多被应用到解决分类问题中,后续的研究可以考虑使用该方法解决回归问题等。

4 结束语

本文将进化集成学习算法作为一个单独的问题进行详细的综述,并且对与分类、回归和聚类3种与机器学习方法相关的进化集成学习算法进行全面的综述。文中对进化集成学习的理论依据与组成结构以及分类情况进行了概述,然后分别根据进化算法在集成学习模型中的优化任务(即样本选择、特征选择、集成模型参数组合优化、集成模型结构优化以及集成模型融合策略优化)总结了一些相关的代表性论文,并分析了这几类方法的主要特点。此外,文中还对一些将无监督学习和有监督学习结合的混合进化集成学习算法进行了综述与分析。进化集成学习算法将集成学习与进化算法结合,不仅为提高集成学习的综合性能提供了新技术,同时也极大地丰富了集成学习的理论。未来关于进化集成学习算法的研究方向主要包括如下几个方面。

(1) 进化集成学习算法中目标函数的设计

选择合适的目标函数对提高进化集成学习算法的性能至关重要。较常用的目标函数包括预测结果的准确度、集成模型的多样性、网络结构的复杂性、网络的负相关度和样本的相似度等。如果通过改变集成模型中子学习器的组合方式来实现对上述指标的优化,在实际应用中往往难以实现,费时费力且效果差强人意。并且上述指标的设计往往是针对特定应用问题的,在设计的过程中并未有统一的标准,因此,进化集成学习中目标函数的设计是一个值得深入研究的问题。

(2) 进化集成学习算法中进化算法最终解的选择

单目标优化算法选择最优解作为输出,然而其无法协调多个优化目标之间的复杂关系;多目标优化算法通过非支配排序的方法获得 Pareto 前沿解,

其输出是一组解集,常用的最终解选择方法有选择全部解、按优先梯度选择解、通过 Pareto 曲线拐点或曼哈顿距离分析法选择解等。然而,分析现有的研究结果发现,最终解的选择方法与其需要解决的问题有很强的关联性。因此,在实际应用中,如何根据要解决的问题确定合适的最终解选择方案仍需要进行更广泛的研究。

(3) 进化集成学习算法中进化算法的选择

在进化集成学习中,需要根据集成模型的特点选择合适的进化算法。如在集成 ANN 的权值、阈值优化问题中,优化对象的维度可能达到几千维,应该选择大规模优化算法。而在集成学习的样本选择(或特征选择)问题中,相同的样本数量(或特征数量)对应的是一组样本子集(或特征子集),要解决该问题应选择多模态优化算法。但是,当前的进化集成学习算法往往缺乏对相关问题的详细分析。在面对大数据、云计算等复杂的机器学习任务时,相应的进化集成学习算法也需要根据实际问题重新设计针对性更强的进化算法。

(4) 进化集成学习方法应用范围的拓展

通过文献综述发现,当前的进化集成学习算法主要被用于解决分类问题,而面向回归和聚类问题的进化集成学习技术也值得更广泛的研究。此外,对进化集成方法进行改进以适用分布式计算、设计适用于在线学习的进化集成学习算法、将无监督学习与有监督学习方法结合以增强集成模型的学习性能等问题都是值得广泛和深入研究的课题。

参考文献:

- [1] MAHMOOD A, LI T, YANG Y, et al. Semi-supervised evolutionary ensembles for web video categorization[J]. Knowledge-Based Systems, 2015, 76: 53-66.
- [2] NAG K, PAL N R. A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification[J]. IEEE Transactions on Cybernetics, 2015, 46(2): 499-510.
- [3] DITERRICH T. Machine learning research: four current direction[J]. Artificial Intelligence Magazine, 1997, 4: 97-136.
- [4] FERNANDEZ A, CARMONA C J, JOSE DEL JESUS M, et al. A pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets[J]. International Journal of Neural Systems, 2017, 27(6): 1750028.
- [5] ROSALES-PEREZ A, GARCIA S, GONZALEZ J A, et al. An evolutionary multiobjective model and instance selection for support vector machines with pareto-based ensembles[J]. IEEE Transactions on Evolutionary Computation, 2017, 21(6): 863-877.
- [6] 姚旭, 王晓丹, 张玉玺, 等. 基于随机子空间和 AdaBoost 的自适应集成方法[J]. 电子学报, 2013, 41(4): 810-814.

- YAO X, WANG X D, ZHANG Y X, et al. A self-adaption ensemble algorithm based on random subspace and AdaBoost[J]. *Acta Electronica Sinica*, 2013, 41(4): 810-814.
- [7] LIU Y, DIAO X C, CAO J J, et al. A method for entity resolution in high dimensional data using ensemble classifiers[J]. *Mathematical Problems in Engineering*, 2017(12): 1-11.
- [8] 李毓, 张春霞, 王冠伟. 线性回归模型的 Boosting 变量选择方法[J]. *工程数学学报*, 2015, 32(5): 677-689.
- LI Y, ZHANG C X, WANG G W. Boosting variable selection algorithm for linear regression models[J]. *Chinese Journal of Engineering Mathematics*, 2015, 32(5): 677-689.
- [9] JIMÉNEZ F, SANCHEZ G, GARCIA J M, et al. Multi-objective evolutionary feature selection for online sales forecasting[J]. *Neurocomputing*, 2017, 234: 75-92.
- [10] HONG Y, KWONG S, CHANG Y, et al. Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm[J]. *Pattern Recognition*, 2008, 41(9): 2742-2756.
- [11] CHANDRA A, YAO X. Ensemble learning using multi-objective evolutionary algorithms[J]. *Journal of Mathematical Modelling and Algorithms*, 2006, 5(4): 417-445.
- [12] CHEN H, YAO X. Multiobjective neural network ensembles based on regularized negative correlation learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(12): 1738-1751.
- [13] 盛伟国, 单鹏霄. 一种基于小生境的负相关神经网络集成算法[J]. *浙江工业大学学报*, 2016(5): 482-486.
- SHENG W G, SHAN P X. Negative correlation neural network ensemble based on a niching algorithm[J]. *Journal of Zhejiang University of Technology*, 2016(5): 482-486.
- [14] SHENG W G, SHAN P, CHEN S, et al. A niching evolutionary algorithm with adaptive negative correlation learning for neural network ensemble[J]. *Neurocomputing*, 2017, 247: 173-182.
- [15] PADILHA C A, BARONE D A C, NETO A D D. A multi-level approach using genetic algorithms in an ensemble of least squares support vector machines[J]. *Knowledge-Based Systems*, 2016, 106: 85-95.
- [16] YAO X, FISCHER M, BROWN G. Neural network ensembles and their application to traffic flow prediction in telecommunications networks[C]//*Proceedings of the 2001 International Joint Conference on Neural Networks*. Piscataway: IEEE Press, 2001: 693-698.
- [17] JIN Y, OKABE T, SENDHOFF B. Neural network regularization and ensembling using multi-objective evolutionary algorithms[C]//*Proceedings of the 2004 Congress on Evolutionary Computation*. Piscataway: IEEE Press, 2004: 1-8.
- [18] 胡梦月, 胡志坚, 仇梦林, 等. 基于改进 AdaBoost.RT 和 KELM 的风功率预测方法研究[J]. *电网技术*, 2017, 41(2): 536-542.
- HU M Y, HU Z J, ZHANG M L, et al. Research on wind power forecasting method based on improved AdaBoost.RT and KELM algorithm[J]. *Power System Technology*, 2017, 41(2): 536-542.
- [19] BUI L T, DINH T T H, et al. A novel evolutionary multi-objective ensemble learning approach for forecasting currency exchange rates[J]. *Data & Knowledge Engineering*, 2018, 114: 40-66.
- [20] HANDL J, KNOWLES J. An evolutionary approach to multiobjective clustering[J]. *IEEE Transactions on Evolutionary Computation*, 2007, 11(1): 56-76.
- [21] MOHAMMADI M, NIKANJAM A, RAHMANI A. An evolutionary approach to clustering ensemble[C]//*Proceedings of the 4th International Conference on Natural Computation*. Piscataway: IEEE Press, 2008: 77-82.
- [22] OBO T, KUBOTA N, LOO C K. Evolutionary ensemble learning of fuzzy randomized neural network for posture recognition[C]//*Proceedings of the 2016 World Automation Congress*. Piscataway: IEEE Press, 2016: 1-6.
- [23] OJHA V K, ABRAHAM A, SNASEL V. Ensemble of heterogeneous flexible neural trees using multiobjective genetic programming[J]. *Applied Soft Computing*, 2017, 52: 909-924.
- [24] MINKU L L, YAO X. Software effort estimation as a multiobjective learning problem[J]. *ACM Transactions on Software Engineering and Methodology*, 2013, 22(4): 1-32.
- [25] SMITH C, JIN Y. Evolutionary multi-objective generation of recurrent neural network ensembles for time series prediction[J]. *Neurocomputing*, 2014, 143: 302-311.
- [26] MIN S H. A genetic algorithm-based heterogeneous random subspace ensemble model for bankruptcy prediction[J]. *International Journal of Applied Engineering Research*, 2016, 11: 2927-2931.
- [27] EKBAL A, SAHA S. A multiobjective simulated annealing approach for classifier ensemble: named entity recognition in indian languages as case studies[J]. *Expert Systems with Applications*, 2011, 38(12): 14760-14772.
- [28] 任克强, 高晓林, 谢斌. 基于 AFSA 和 PSO 融合优化的 AdaBoost 人脸检测算法[J]. *小型微型计算机系统*, 2016, 37(4): 861-865.
- REN K Q, GAO X L, XIE B. AdaBoost face detection algorithm based on fusion optimization of AFSA and PSO[J]. *Journal of Chinese Computer Systems*, 2016, 37(4): 861-865.
- [29] ONAN A, KORUKOGLU S, BULUT H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification[J]. *Expert Systems with Applications*, 2016, 62: 1-16.
- [30] ABRAHAM A, GROSAN C, HAN S Y, et al. Evolutionary multiobjective optimization approach for evolving ensemble of intelligent paradigms for stock market modeling[C]//*Proceedings of the 2005 Mexican International Conference on Artificial Intelligence*. Heidelberg: Springer-Verlag, 2005: 673-681.
- [31] CHEN Y, YANG B, ABRAHAM A. Flexible neural trees ensemble for stock index modeling[J]. *Neurocomputing*, 2007, 70(4-6): 697-703.
- [32] ALAMANIOTIS M, IKONOMOPOULOS A, TSOUKALAS L H. Evolutionary multiobjective optimization of kernel-based very-short-term load forecasting[J]. *IEEE Transactions on Power Systems*, 2012, 27(3): 1477-1484.
- [33] YOON H S, AHN S Y, LEE S H, et al. Heterogeneous clustering ensemble method for combining different cluster results[C]//*Proceedings of the International Workshop on Data Mining for Biomedical Applications*. Heidelberg: Springer-Verlag, 2006: 82-92.
- [34] LUO H, JING F, XIE X. Combining multiple clusterings using information theory based genetic algorithm[C]//*Proceedings of the 2006 International Conference on Computational Intelligence and Security*. Piscataway: IEEE Press, 2006: 84-89.
- [35] QIAN X, ZHANG X, JIAO L, et al. Unsupervised texture image segmentation using multiobjective evolutionary clustering ensemble algorithm[C]//*Proceedings of the 2008 IEEE Congress on Evolutionary Computation*. Piscataway: IEEE Press, 2008: 3561-3567.

- [36] WAHID A, GAO X, ANDREA P. Multi-objective multi-view clustering ensemble based on evolutionary approach[C]//Proceedings of the 2015 IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2015: 1696-1703.
- [37] WANG Y, LIU X, XIANG L. Ga-based membrane evolutionary algorithm for ensemble clustering[J]. Computational Intelligence and Neuroscience, 2017.
- [38] LI X, ZHANG S, WONG K C. Single-cell RNA-seq interpretations using evolutionary multiobjective ensemble pruning[J]. Bioinformatics, 2019, 35(16): 2809-2817.
- [39] REN Y, ZHANG L, SUGANTHAN P N. Ensemble classification and regression-recent developments, applications and future directions[J]. IEEE Computational Intelligence Magazine, 2016, 11(1): 41-53.
- [40] MENDES-MOREIRA J, SOARES C, JORGE A M, et al. Ensemble approaches for regression: a survey[J]. ACM Computing Surveys, 2012, 45(1): 1-40.
- [41] VEGA-PONS S, RUIZ-SHULCLOPER J. A survey of clustering ensemble algorithms[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(3): 337-372.
- [42] MENDES-MOREIRA J, SOARES C, JORGE A M, et al. Ensemble approaches for regression: a survey[J]. ACM Computing Surveys, 2012, 45(1): 1-40.
- [43] FREUND Y, SCHAPIRE R E, et al. Experiments with a new boosting algorithm[C]//Proceedings of the 13th International Conference on Machine Learning. [S.l.:s.n.], 1996: 148-156.
- [44] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.
- [45] TING K M, WITTEN I H. Stacking bagged and dagged models[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1997: 367-375.
- [46] HO T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.
- [47] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.
- [48] MORENO-SECO F, INESTA J M, DE LEON P J P, et al. Comparison of classifier fusion methods for classification in pattern recognition tasks[C]//Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Heidelberg: Springer-Verlag, 2006: 705-713.
- [49] DIETTERICH T G. Ensemble methods in machine learning[C]//Proceedings of the 1st International Workshop on Multiple Classifier Systems. Heidelberg: Springer-Verlag, 2000: 1-15.
- [50] JAMES G M. Variance and bias for general loss functions[J]. Machine Learning, 2003, 51(2): 115-135.
- [51] KLEINBERG E. Stochastic discrimination[J]. Annals of Mathematics and Artificial Intelligence, 1990, 1(1-4): 207-239.
- [52] SCHAPIRE R E, FREUND Y, BARTLETT P, et al. Boosting the margin: a new explanation for the effectiveness of voting methods[J]. The Annals of Statistics, 1998, 26(5): 1651-1686.
- [53] FRANK E, PFAHRINGER B. Improving on bagging with input smearing[C]//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2006: 97-106.
- [54] BREIMAN L. Randomizing outputs to increase prediction accuracy[J]. Machine Learning, 2000, 40(3): 229-242.
- [55] LIN H T, LI L. Infinite ensemble learning with support vector machines[C]//Proceedings of the 16th European Conference on Machine Learning. Heidelberg: Springer-Verlag, 2005: 242-254.
- [56] BI Y. The impact of diversity on the accuracy of evidential classifier ensembles[J]. International Journal of Approximate Reasoning, 2012, 53(4): 584-607.
- [57] 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 C X, ZHANG J S. A survey of selective ensemble learning algorithms[J]. Chinese Journal of Computers, 2011, 34(8): 1399-1410.
- [58] SAGI O, ROKACH L. Ensemble learning: a survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1249.
- [59] PACL'IK P, LANDGREBE T C, TAX D M, et al. On deriving the second-stage training set for trainable combiners[C]//Proceedings of the 6th International Workshop on Multiple Classifier Systems. Heidelberg: Springer-Verlag, 2005: 136-146.
- [60] 张春霞. 集成学习中有关算法的研究[D]. 西安: 西安交通大学, 2010.
- ZHANG C X. Research on some algorithms in ensemble learning[D]. Xi'an: Xi'an Jiaotong University, 2010.
- [61] ROONEY N, PATTERSON D, ANAND S, et al. Dynamic integration of regression models[C]//Proceedings of the 5th International Workshop on Multiple Classifier Systems. Heidelberg: Springer-Verlag, 2004: 164-173.
- [62] VALENTINI G, MASULLI F. Ensembles of learning machines[C]//Proceedings of the 13th Italian Workshop on Neural Nets. Heidelberg: Springer-Verlag, 2002: 3-20.
- [63] ZHOU Z H, WU J, TANG W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1-2): 239-263.
- [64] WANG Y, LI B. Multi-strategy ensemble evolutionary algorithm for dynamic multi-objective optimization[J]. Memetic Computing, 2010, 2(1): 3-24.
- [65] YEN G G, HE Z. Performance metric ensemble for multiobjective evolutionary algorithms[J]. IEEE Transactions on Evolutionary Computation, 2013, 18(1): 131-144.
- [66] QIU X, SUGANTHAN P N, AMARATUNGA G A. Ensemble incremental learning random vector functional link network for short-term electric load forecasting[J]. Knowledge-Based Systems, 2018, 145: 182-196.
- [67] QIU X, REN Y, SUGANTHAN P N, et al. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting[J]. Applied Soft Computing, 2017, 54: 246-255.
- [69] YUE C, LIANG J, QU B, et al. Multimodal multiobjective optimization in feature selection[C]//Proceedings of the 2019 IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2019: 302-309.
- [69] YUE C, QU B, LIANG J. A multiobjective particle swarm optimizer using ring topology for solving multimodal multiobjective problems[J]. IEEE Transactions on Evolutionary Computation, 2017, 22(5): 805-817.
- [70] HU Y, WANG J, LIANG J, et al. A self-organizing multimodal multi-objective pigeon-inspired optimization algorithm[J]. Science China-information Sciences, 2019, 62(7): 1-17.

- [71] AIN Q U, SAHAF H A, XUE B, et al. A genetic programming approach to feature construction for ensemble learning in skin cancer detection[C]//Proceedings of the 2020 Genetic and Evolutionary Computation Conference. New York: ACM Press, 2020: 1186-1194.
- [72] ZENG J, XU H Y, GONG G L, et al. An heterogeneous stacking ensemble model for laser-induced breakdown spectroscopy elemental quantitative analysis[J]. IEEE Access, 2020, 8: 136141-136150.
- [73] HUANG T, DUAN D T, GONG Y J, et al. Concurrent optimization of multiple base learners in neural network ensembles: an adaptive niching differential evolution approach[J]. Neurocomputing, 2020, 396: 24-38.
- [74] YAN S, WANG Y, KONG D, et al. Hybrid chain-hypergraph p systems for multiobjective ensemble clustering[J]. IEEE Access, 2019, 9: 143511-143523.
- [75] MINKU L L, YAO X. An analysis of multi-objective evolutionary algorithms for training ensemble models based on different performance measures in software effort estimation[C]//Proceedings of the 9th International Conference on Predictive Models in Software Engineering. [S.l.:s.n.], 2013: 1-10.
- [76] LIU W, WU W, WANG Y, et al. Selective ensemble learning method for belief-rule-base classification system based on PAES[J]. Big Data Mining and Analytics, 2019, 2(4): 306-318.
- [77] FACELI K, DE CARVALHO A C, DE SOUTO M C. Multi-objective clustering ensemble[J]. International Journal of Hybrid Intelligent Systems, 2007, 4(3): 145-156.
- [78] LIU R, LIU Y, LI Y. An improved method for multi-objective clustering ensemble algorithm[C]//Proceedings of the 2012 IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2012: 1-8.
- [79] GOUDOS S, ATHANASIADOU G E. Application of an ensemble method to uav power modeling for cellular communications[J]. IEEE Antennas and Wireless Propagation Letters, 2019(99): 1.
- [80] DAI H, SHENG W. A multi-objective clustering ensemble algorithm with automatic k-determination[C]//Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis. Piscataway: IEEE Press, 2019: 333-337.
- [81] MINKU F L, LUDERMIR T B. Clustering and co-evolution to construct neural network ensembles: an experimental study[J]. Neural Networks, 2008, 21(9): 1363-1379.
- [82] RAHMAN A, VERMA B. Cluster oriented ensemble classifiers using multi-objective evolutionary algorithm[C]//Proceedings of the 2013 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2013: 1-6.
- [83] COLETTA L F, HRUSCHKA E R, ACHARYA A, et al. A differential evolution algorithm to optimise the combination of classifier and cluster ensembles[J]. International Journal of Bio-Inspired Computation, 2015, 7(2): 111-124.
- [84] JAN Z M, VERMA B K. Evolutionary classifier and cluster selection approach for ensemble classification[J]. ACM Transactions on Knowledge Discovery from Data, 2019.
- [85] JIANG F, HE J, TIAN T. A clustering-based ensemble approach with improved pigeon-inspired optimization and extreme learning machine for air quality prediction[J]. Applied Soft Computing, 2019, 85: 105827.
- [86] ONAN A, KORUKOGLU S, BULUT H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification[J]. Information Processing and Management, 2017, 53 (4): 814-833.

[作者简介]



胡毅(1988-), 男, 郑州大学电气工程学院博士生, 主要研究方向为进化计算、机器学习、智能控制及其应用。



霍博阳(1984-), 男, 博士, 中原工学院电子信息学院教授、院长, 主要研究方向为机器学习、遗传与进化算法、群集智能和多目标优化。



梁静(1981-), 女, 博士, 郑州大学电气工程学院教授、院长, 主要研究方向为计算智能、智能优化及机器学习。



王杰(1959-), 男, 博士, 郑州大学电气工程学院教授, 主要研究方向为模式识别与智能控制。



王艳丽(1982-), 女, 郑州大学电气工程学院博士生, 主要研究方向为智能计算和机器学习。