计算机科学
COMPUTER SCIENCE

一种基于层次聚类和模拟退火的选择性集成算法的风控模型研究

王茂光, 冀昊悦, 王天明

引用本文

王茂光, 冀昊悦, 王天明.一种基于层次聚类和模拟退火的选择性集成算法的风控模型研究[J]. 计算机科学, 2022, 49(11A): 210800105-7.

WANG Mao-guang, JI Hao-yue, WANG Tian-ming. Study on Risk Control Model of Selective Ensemble Algorithm Based on Hierarchical Clustering and Simulated Annealing [J]. Computer Science, 2022, 49(11A): 210800105-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800105

一种基于层次聚类和模拟退火的选择性集成算法的风控模型研究

王茂光　冀昊悦　王天明
中央财经大学信息学院　北京 100081
(wangmg@cufe.edu.cn)

摘　要　集成学习模型可有效解决单一模型出现的模型结构单一、稳定性和预测能力弱的问题。但是由于结构复杂,其常出现运行效率低下、存储代价过大等问题,一般使用选择性集成算法优化集成学习模型来解决这些问题。目前提出的选择性集成算法仍存在运行效果和效率提升不够明显的现象。为解决这些问题,提出一种基于 Stacking 集成框架的选择性集成算法,算法主要使用了凝聚型层次聚类(AHC)算法和模拟退火的 Metropolis 准则对基学习器的种类和个数进行筛选。在实证分析方面,分别使用了国内外网贷对模型进行搭建。实验结果证明,AHC-Metropolis 选择性集成模型可有效提升计算效率、预测能力、稳定性和泛化能力,有助于规范互联网金融行业秩序,协助开展金融监管任务,为建立我国金融风控管理体系和保障国家金融安全提供有效依据。

关键词:层次聚类;模拟退火;选择性集成;金融风控
中图法分类号　TP181

Study on Risk Control Model of Selective Ensemble Algorithm Based on Hierarchical Clustering and Simulated Annealing

WANG Mao-guang,JI Hao-yue and WANG Tian-ming
School of Information,Central University of Finance and Economics,Beijing 100081,China

Abstract　Eensemble learning model can effectively solve the problems of single model structure,stability and weak predictive ability. However,due to the complexity of its structure,problems such as low operating efficiency and excessive storage cost often occur. Selective ensemble algorithms are often used to optimize ensemble learning models to solve these problems. The currently proposed selective ensemble algorithm still has the phenomenon of insufficient operating effect and efficiency improvement. In order to make up for these shortcomings,a selective ensemble algorithm based on the stacking ensemble framework is proposed. It mainly uses the agglomerated hierarchical clustering(AHC) algorithm and the metropolis criterion of simulated annealing to select the type and number of base learners. In terms of empirical analysis,domestic and foreign online loan data are used separately to build the model. Experimental results prove that the selective ensemble model of AHC-Metropolis can effectively improve the computational efficiency,predictive ability,stability and generalization ability. It is helpful for regulating the order of the Internet financial industry,assist in financial supervision tasks,and provide an effective basis for establishing our country's financial risk control management system and guaranteeing national financial security.

Keywords　Hierarchical clustering,Simulated annealing,Selective ensemble,Financial risk control

1　引言

互联网金融是金融行业在科技时代的产物,主要内容为通过各种科技手段为金融行业提供新的业务和服务。科技的融入使金融业的发展模式产生了变革,互联网金融将持续成为金融业的发展重点。国内外互联网金融的迅速兴起,直接导致了风险问题的集中爆发,大量互联网金融平台出现运营困难、提现困难、破产退市等问题。为保障国家金融安全和维护经济市场秩序,我国政府高度重视此类问题的发生,自2016年起,针对互联网金融业务开展专项整治工作。

2019年9月发布的《金融科技(FinTech)发展规划(2019—2021年)》为我国互联网金融业务的发展方向进行规划,旨在强调科技需对金融发展提供正向作用,并着重说明了互联网金融监管的重要性,提出要专业化、统一化地进行风险控制管理。2020年9月发布的《关于加强小额贷款公司监督管理的通知》中,提出要明确小额信贷公司的业务内容和贷款明细,并对企业所在地的监管责任进行明确。2021年2月发布的《关于进一步规范商业银行互联网贷款业务的通知》中规范了商业银行贷款的流程秩序,对贷款业务中的监管任务进行了进一步的细化和明确。互联网金融作为市场经济的重要组成部分,其

基金项目:国家自然科学基金(62072487);中央财经大学科技项目(020676116004,020676114004)
This work was supported by the National Natural Science Foundation of China(62072487) and Research Projects of Central University of Finance and Technology(020676116004,020676114004).
通信作者:冀昊悦(jihaoyuew@163.com)

监管的严密性对我国金融业的发展起到了重要作用。

近几年来，互联网金融行业经历了金融市场去杠杆化、市场经济结构变革、网贷平台暴雷和清零等一系列事件。2020年，疫情的爆发导致企业和个人对贷款的需求量大规模增加。监管指标不明确、不统一的问题仍然存在，使得贷款风险类事件持续发生，企业和个人通过平台贷款实施欺诈和使用银行信用卡进行套现的恶性事件屡禁不止。在这样的背景下，政府逐步清退网络贷款平台，重整银行信用体系。为保障国家的金融安全，加强金融信用风险的管理能力，我国亟需建立健全互联网金融风险控制管理体系，完善风险控制评估模型。为助力完善金融风控的管理体系，本文致力于建立一个精准有效的金融风控模型，从而支持和保障我国互联网金融市场可以健康、平稳和持续的发展。

目前，机器学习中的多种单一学习器算法已经应用于金融风控场景中，但是单一学习器存在不稳定和监督性弱的情况。使用集成学习算法可以将较弱的学习器进行组合，得到稳定性高的强监督性模型，并且异质集成的效果往往优于同质集成。针对金融领域出现的风险问题，本文提出一种基于凝聚型层次聚类和模拟退火 Metropolis 准则的异质集成学习模型退，模型名称为 AHC-Metropolis 选择性集成模型。该模型的总体框架为 Stacking 集成模型架构，模型有两个主要内容。

(1)使用凝聚型层次聚类(AHC)算法，对预测效果较好的单一学习器通过相似簇合并的思想进行“自下而上”结合。根据学习器之间的距离进行差异性选择，有助于降低学习器之间的相似性。差异性保证了基学习器之间的独立性，可以有效提升模型的泛化能力和稳定性。

(2)运用模拟退火 Metropolis 准则中能量值逐步变化至稳定的思想进行学习器筛选，设立特定度量指标测量集成基学习器组合的能量值，根据能量值的变化达到对各类基学习个数筛选的结果。在迭代过程中使用该准则可以得到分类器集合中的全局最优子集，完成针对分类器的二次筛选，保证在后续训练的过程中，模型仅使用基学习器子集就可以得到最优效果。

通过这两部分搭建的选择性集成学习模型，不仅可以有效减少模型在训练和预测时的计算成本，降低存储预测结果开销，还能提升模型的效率、准确性和泛化能力。

本文第 2 节主要对金融风控领域的相关文献和发展现状进行阐述；第 3 节对 AHC-Metropolis 集成模型中的核心算法进行描述；第 4 节进行实证实验和对比实验；最后总结全文。

2 相关工作

随着大数据时代的来临，信贷数据的体量逐渐增大，数据中特征复杂和噪音多的问题也越来越明显，仅使用专家经验对风险进行判断，会出现评估结果不准确的现象，给金融行业造成损失。为保证信用风险评估的准确性，常使用机器学习算法对金融领域数据进行信用评估和建模。

2.1 单一学习器模型

目前，应用于金融风控领域的机器学习算法主要包括 Logistic 回归、支持向量机、神经网络、朴素贝叶斯和 K-近邻算法等，这些算法都可以很好地对金融数据进行建模和评估。为进一步提升模型的准确率，学者们对这些单一机器学习算法进行改进，在基础的算法中融入 Lasso 方法、社交媒体信息、FCM 算法等[1-5]。为了验证学习器的优劣性，学者将不同类型的学习器进行比较，例如 Li 等人使用多种贝叶斯分类器与 David West 提出的 5 种神经网络模型相比较[6-7]。Li 毅等人通过不同平衡化方法处理后的数据建立树状模型和支持向量机模型进行对比[8]。树状结构的模型常用于金融数据中，极度随机树和随机森林的基模型均为决策树，随机森林是将决策树模型以 Bagging 集成策略进行组合，极度随机树使用数据的随机特征和所有样本进行训练，实验结果证明极度随机树可以得到更好的分类效果[9]。另外，AdaBoost 算法也常使用决策树算法作为基模型，通过设置决策树的权重大小提升分类效果，可用于金融风控的模型建立[10]。

2.2 集成学习策略

虽然单一学习器建立的模型效果表现已经较为不错，但模型的稳定性和泛化能力仍有欠缺，因此对多学习器组合形成的集成学习模型进行研究和分析。集成学习作为机器学习中一个重要的部分，受到了较多的关注。许多学者将研究重点放在集成学习方法中。集成学习将多个单一学习器通过特定的方法进行组合，有助于提升模型的性能[11]。

集成学习根据学习器类型，分为同质集成和异质集成，同质集成使用单一类型学习器进行组合[12-13]，异质集成使用不同种类学习器进行组合[14-15]。Liu 等[16]将单一学习器、同质集成和异质集成的效果进行比较，实验证明异质集成效果表现更优。通过集成学习中基学习器的组合模式进行划分，主要分为 Bagging，Boosting 和 Stacking。Bagging 是一种并行的集成学习策略，Qi 等[17]使用不同的特征选择方法搭建 SVM 模型，并使用 Bagging 策略进行建模。Boosting 是一种基于迭代方法和权重计算的集成学习策略，Li 等人使用 K-近邻作为 Boosting 框架的基学习器，与单纯的 K-近邻相比，模型的精准度有明显的提升[18]。Stacking 是一种两阶段模式的集成学习策略，第一阶段为多个同质或异质学习器，第二阶段常使用一个强学习器。其中，第一阶段的输出作为第二阶段的输入。Yu 等[19]将极限学习机作为初级学习器，DBN 作为次级学习器。Wang 等[20]使用 LightGBM 和 FM 的融合作为基学习器，再使用多分类算法进行训练。Cao 等[21]使用 Stacking 集成策略搭建的模型分别与单一分类器模型和投票法相比较，证明具有两层分类器特性的 Stacking 集成策略有更高的精度和稳定性。

随着集成学习中基学习器数量的增多，模型训练较慢，预测效果无明显增长甚至负增长的问题逐渐显现。因此 Zhou 等人提出选择性集成学习[22]。选择性集成学习算法既保证模型可以通过综合多种基学习器算法提升分类效果，又优化了模型结构，减少了不必要的运算和存储开销。选择集成方法主要分为聚类、排序、选择和优化共 4 种[23]。每种方法都可以使用多种不同算法进行实现。学者们在研究过程中可单独使是用这些算法，也可将不同种类的选择性集成算法组合使用[24]。例如 Wu 等人使用分层筛选和动态更新方法选取出最优的基学习器集合，其可有效解决分类器选择效率低的问题[25]。Du 和 Zhang 利用最大相关性和最小冗余准则对基分类器对 AdaBoost 生成的基分类器进行选择，并使用加权投票的方法进行集成[26]。Yu 提出基于熵算法的异质集成学习

模型，通过计算学习器组合的熵进行选择和删除，有效减少了基学习器个数，降低了存储空间[27]。Yang 在熵算法的基础上引入 AP 聚类算法，在减小了基学习器规模的同时保证了种类的差异性[28]。

基于异质集成的选择性集成策略可以在保证集成模型准确度、稳定性和泛化能力的同时，提升模型效率，降低计算和存储的开销成本。但将此策略运用于金融风控领域的研究还较少。因此，本文将该策略应用于国内外金融领域数据中，建立准确可靠的金融风控模型。

3 基于 AHC-Metropolis 选择性集成模型

集成学习模型将多个单一分类器进行组合，这些单一分类器被称为基学习器。随着集成学习模型中使用到的基学习器种类和数量的逐步增多，模型的准确度和稳定性也有一定的提升，但是经过实验发现，当模型中的基学习器数量达到一定量级时，预测效果开始出现增长停滞甚至负增长的情况。由此可以得出，基学习器种类和数量的增加虽对模型性能有着一定的正向作用，但也并非越多越好。

在迭代过程中，基学习器的数量出现成倍数增长的情况，可能会导致下面几种问题：1）模型中每个基学习器都需要对数据进行训练和预测，当基学习器的量级过大时，计算和存储的成本较大；2）当各个基学习器之间相似性较高时，会存在部分基学习器的使用效率较低的情况；3）可能会出现组合内部分基学习器预测错误率较高的问题，当错误率高于一定的阈值时，集成后模型会出现准确度提升不显著甚至下降的现象。

目前，学者们开始针对现存问题进行实证分析和研究，但仍没有统一的标准对基学习器进行筛选，并且筛选后得到的模型依旧存在计算时间过长和模型性能不够优秀的情况。在领域场景适用性方面，针对金融风控领域的选择性集成模型的研究仍有欠缺。为解决这些问题，本文建立了适用于金融风控场景的 AHC-Metropolis 选择性集成模型，模型整体框架如图 1 所示。

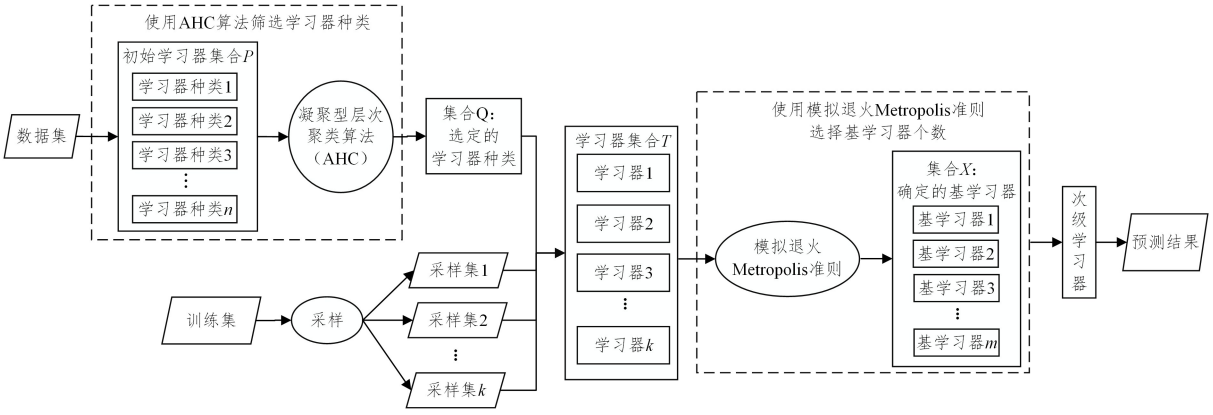

图 1 模型框架图

Fig. 1 Framework of the proposed model

AHC-Metropolis 选择性集成模型使用 Stacking 集成学习策略进行搭建，采用两层堆叠的模式，第一层由异质基学习器组成，第二层使用单一强学习器。在第一层中，使用凝聚型层次聚类算法可以保证各个基学习器之间的差异性，然后将训练集的采样子集放入已经确定的学习器种类中进行训练。同时在迭代过程中，使用模拟退火的 Metropolis 准则对基学习器集合进行确定，达到对种类和数目进一步筛选的效果。通过这两步的筛选，模型中会保留差异性大且对集成模型贡献作用较大的基学习器。在选择的过程中删除一定数量的基学习器，可以在保证模型效果的同时，有效降低存储所花费的成本。

3.1 基于凝聚型层次聚类的学习器种类选择算法

在阅读文献时，我们发现使用在金融风控领域的单一学习器种类较多，若将它们全部用于集成学习模型的搭建，可能会有比使用单一学习器更好的预测效果，但是将会带来很大的计算代价。使用数据对单一学习器进行训练，通过其预测结果进行聚类选择，可以直观地得出学习器之间的相似性，有助于在建模过程中选择差异性强的学习器，从而提升模型性能。因此，本文使用聚类算法对学习器的种类进行选择，主要过程是使用聚类算法将数据分类到不同簇，保证每个簇中的元素具有较高的相似性。使用已经训练好的单一学习器的预测结果进行聚类分析，根据结果保留差异性大且准率高的学习器种类，从而达到减小内存开销的效果。

通过大量的实证分析，本文选取在信贷风控领域表现较好的单一学习器，包括：决策树、Logistic 回归、BP 神经网络、K-近邻、支持向量机、朴素贝叶斯、随机森林、极度随机树和 AdaBoost，共计 9 种。本文引入聚类中的凝聚型层次聚类算法，这种聚类算法不需要提前设定初始聚类簇中心和聚类数量。AHC 算法的中心思想为“自下而上”，首先将每一个对象都看作一个簇，然后根据一定的距离计算方式将簇进行合并，终止条件为所有对象都在一个簇中，或是某个特定条件被满足。

本文选取的簇合并方式为单链合并，将簇的邻近度定义为：不同簇的两个最近点之间的邻近度。公式如下：

$$dist(C_u,C_v)=\min(dist(x_{ui},x_{vj}));i=1,2,3,\cdots,m;\quad j=1,2,3,\cdots,n \tag{1}$$

其中，x_{ui} 和 x_{vi} 分别簇 C_u 和簇 C_v 中的数据点（$x_{ui}\in C_u$，$x_{vi}\in C_v$），m 和 n 分别为簇 C_u 和簇 C_v 中数据点的数量。

使用欧氏距离法作为数据点之间距离的计算方法，公式如下：

$$d_{ij}=\sqrt{\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2} \tag{2}$$

其中，i 和 j 分别代表数据对象 i 和 j，n 为数据对象的个数。

本文使用 AHC 算法筛选学习器种类的过程如下：

Step 1 将数据集输入到上述 9 个单一学习器的集合 P 中，依次对每个学习器进行训练，得到 F1 值、准确率、召回率等评估指标。

Step 2 使用 $F1$ 值对单一学习器进行有效性判断，将阈值设定为 80%，当学习器 $F1$ 值小于 80%时，在集合 P 中删除此学习器，将 P 中剩余学习器预测的结果组合存入矩阵 G 中。

Step 3 将矩阵 G 中的数据使用 AHC 算法进行计算，得到层次聚类结果。聚类结果中包含每个簇的索引值以及簇之间的距离。

Step 4 当出现两个簇邻近程度较大时，使用 Step1 中得到的各学习器评估指标对簇进行筛选，得到最终的筛选结果，存入集合 Q 中，集合 Q 中的学习器种类为后续用于模型搭建的基学习器种类。

3.2 基于模拟退火 Metropolis 准则的学习器选择算法

Metropolis 准则是一种接受准则，通过对固体状态变动前后能量值的计算，决定是否接受或是以一定概率接受变动后的新状态。模拟退火算法是将 Metropolis 准则通过迭代进行组合优化，从而解决固体退火问题的算法。Metropolis 准则应用于本文场景中的公式如下：

$$p_t(X_{\text{old}} \Rightarrow X_{\text{new}}) = \begin{cases} 1, & F \leqslant F' \\ \exp\left(\dfrac{\Delta F}{t}\right), & F > F' \end{cases} \tag{3}$$

其中，F 为 Metropolis 中的能量值，$p_t(X_{\text{old}} \Rightarrow X_{\text{new}})$ 表示接受新解 X_{new} 的概率。当 $F \leqslant F'$，组合接受新解；当 $F > F'$ 时，对概率值进行判定，若概率值 $\exp(\Delta F/t)$ 大于一定的阈值，组合接受新解；否则，不接受新解。

在模型使用 Metropolis 准则筛选基学习器的过程中，可能出现某一种或几种基学习器表现过于优秀的情况，若只是通过某种单一评估指标来确定组合的能量值，则可能出现集成模型在选择的过程中总是选择同一种类型的基学习器的现象。为避免这种问题的出现，能量值需要保证既可以考虑到模型的预测精准性，又可以保证基学习器的差异性，因此使用 F 来度量组合的能量值。F 的计算公式如下：

$$F(X) = Acc(X) \times \gamma + Ka(X) \times (1-\gamma) \tag{4}$$

其中，X 为每次迭代过程中筛选出的基学习器集合；Acc 为准确率(Accuracy)，常用于评估模型的性能，在这里用于计算基学习器集合的准确率；Ka 为 Kappa 系数，由 Cohen 于 1960 年提出[29]，被广泛应用于评估者之间的评分一致性检验。γ 用于调节评价指标 Acc 和 Ka 在度量方式 F 中的权重，当 γ 越大时，准确率 Acc 的作用越大，当 γ 越小时，Kappa 系数 Ka 的作用越大。

Kappa 系数的取值范围为(−1,1)，当 Kappa 系数越接近 −1 时，表示学习器的预测差异性越大；越接近 1 时，则结果相反。该系数计算公式为：

$$Ka = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

其中，p_0 为两个学习器分类的一致性，即使用分类相同的数量除以总样本数量；p_e 为两个学习器分类的机遇一致性，p_e 的计算公式如下：

$$p_e = \frac{\sum_{i=1}^{c} a_i \times b_i}{n^2} \tag{6}$$

其中，a_i 表示学习器 a 的第 i 类预测结果的个数，b_i 表示学习器 b 的第 i 类预测结果的个数，n 表示单一学习器预测结果的个数。因此，度量指标 F 既可以考虑到学习器的准确性，也可以保证在筛选过程中进一步维持基学习器的差异性。

本文使用基于模拟退火 Metropolis 准则的基学习器选择过程如下，其中 $F1$ 值和保留概率 p 的阈值取值均是通过大量实验分析得到的。

Step 1 向 3.1 节中选择的学习器种类输入数据进行训练，通过迭代选择 $F1$ 值高于 85%的基学习器，并放入集合 T 中。

Step 2 从集合 T 中选取第一个学习器放入集合 X 中，并计算集合 X 的度量值 F。

Step 3 在集合 T 中依次选取下一个学习器放入集合 X 中，计算集合 X 的新度量值 F'。

Step 4 当 $F \leqslant F'$ 时，当前学习器被保留在集合 X 中，且将度量值 F 更新。

Step 5 当 $F > F'$ 时，则需要对保留概率 p 进行计算，使用式(3)中的概率计算方式进行计算，若 $p < 0.9$，则删除该学习器；反之，则保留，且更新度量值 F。

Step 6 重复 Step 3−Step 5，直至集合 T 中所有的学习器均已参与迭代。最终生成的集合 X 为后续选择性集成学习模型的基学习器组合。

4 实验结果和分析

4.1 特征选择

本文使用的实验数据集是 2019 年网贷之家平台上的信贷相关信息，数据集中包含样本 5570 条，特征 27 个。首先对数据进行异常值处理、缺失值处理、特征初步筛选等数据预处理操作，得到的特征如表 1 所列。

表 1 数据预处理后数据特征

Table 1 Data characteristics after data preprocessing

特征	编号	特征	编号	特征	编号
参考收益	x_1	注册资金/万元	x_{10}	待收投资人数/人	x_{19}
投资期限/月	x_2	银行存管	x_{11}	借款人数/人	x_{20}
综合评分	x_3	自动投标	x_{12}	人均借款金额/万元	x_{21}
点评人数	x_4	债券转让	x_{13}	借款标数/个	x_{22}
注册年限	x_5	保障模式	x_{14}	待还借款人数/人	x_{23}
运营时间	x_6	资金净流入/万元	x_{15}	ICP 认证	x_{24}
待还余额/万元	x_7	平均借款期限	x_{16}	平台背景	x_{25}
昨日成交量	x_8	投资人数/个	x_{17}	加入协会	x_{26}
关注投友数	x_9	人均投资金额/万元	x_{18}	目标变量(Target)	Y

然后，使用特征工程对特征进行选择。本文选取的特征工程方法为 RFE 递归消除法。递归消除法是一种寻求最佳特征子集的优化算法，其在迭代过程中反复创建模型，剔除表现差的特征，根据保留和剔除的顺序对特征的重要程度进行排名，最终得到一个特征最优子集。数据通过 RFE 特征选择方法获取的特征子集如表 2 所列。

表 2　通过 RFE 选择得到的特征子集

Table 2　Feature subsets obtained through RFE selection

特征	编号	特征	编号	特征	编号
注册年限	x_5	银行存管	x_{11}	借款标数/个	x_{22}
运营时间	x_6	自动投标	x_{12}	待还借款人数/人	x_{23}
昨日成交量	x_8	待收投资人数/人	x_{19}	ICP 认证	x_{24}
关注投友数	x_9	借款人数/人	x_{20}	平台背景	x_{25}
注册资金/万元	x_{10}	人均借款金额/万元	x_{21}	加入协会	x_{26}
目标变量(Target)			Y		

4.2　信贷风控模型搭建

本文选用 Stacking 集成学习策略搭建信贷风控模型。Stacking 框架为两层堆叠模式，因此可以分为两个阶段来对模型进行构建。

第一阶段为对基学习器进行选取。本文首先使用了 AHC，用于保证基学习器之间的差异性。然后，引入了模拟退火的 Metropolis 准则，使用准确值和 Kappa 系数构建度量值，在迭代的过程中对基学习器进行筛选。第二阶段需选用一个强学习器作为次级学习器，由于 Stacking 集成学习模型的第一阶段的训练结果存在较强的多重共线性，因此选用树状结构的 XGBoost 算法作为次级学习器，可以最小化多重共线性对模型的影响。

将选取出的特征空间输入 3.1 节中提到的 9 个单一学习器中，根据学习器的 $F1$ 值是否大于 0.8 判定是否保留该学习器。各学习器的判定结果及对应索引如表 3 所列。

表 3　各学习器的判定结果

Table 3　Judgment results of each learner

学习器名称	索引值	$F1$ 是否大于 0.8	是否可以保留
决策树	0	是	是
Logistic 回归	1	是	是
BP 神经网络	2	是	是
K-近邻	3	是	是
朴素贝叶斯	4	是	是
随机森林	5	是	是
极度随机树	6	是	是
AdaBoost	7	是	是
支持向量机	8	是	是

根据表 3 的结果可以得出，以上 9 个学习器的表现均较优，$F1$ 值均大于 0.8。若仅通过模型性能进行判定，均可保留。因此需进一步计算模型的聚类结果。将选择出的最优特征子集输入到 3.1 节中提出的 AHC 中，得到的结果过如图 2 所示，图中横坐标对应表 3 中各个学习器的索引值。

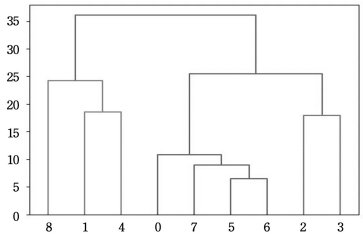

图 2　使用 AHC 算法的聚类结果

Fig. 2　Clustering results with AHC algorithm

根据图 2 中的 AHC 聚类结果，选取的学习器种类为：决策树(索引：0)、Logistic 回归(索引：1)、BP 神经网络(索引：2)、K-近邻(索引：3)、朴素贝叶斯(索引：4)和支持向量机(索引：8)，共计 6 种。

因此，将上述 6 种学习器作为模型的初级学习器，将 XGBoost 作为模型的次级学习器，通过穷举法设置训练的迭代次数，对不同迭代次数获得的结果进行分析和对比。通过大量的实验分析，我们得出迭代次数在 50 至 200 之间时，运行时间较为合理且模型表现较好，因此选取迭代次数的取值范围为[50，200]。

在 3.2 节中，度量值中用到了调节权重的指标 γ。为选取出表现较优的权重指标，我们对不同的指标值 γ 分别进行实验。由于 γ 的取值范围为[0，1]，我们将选取极端值($\gamma=0$ 和 $\gamma=1$)、中点值($\gamma=0.5$)和四分位值($\gamma=0.25$ 和 $\gamma=0.75$)分别进行实验。

首先需要对模型的迭代次数进行确定，选取指标 $\gamma=0.5$ 进行实验，结果如表 4 所列。

表 4　$\gamma=0.5$ 时不同迭代次数的实验结果

Table 4　Experimental results of different iteration times when $\gamma=0.5$

迭代次数	基学习器个数	$F1$	*Accuracy*	*AUC*	*Precision*	*Recall*	运行时间/min
50	20	0.9823	0.9768	0.9929	0.9843	0.9804	3
100	32	0.9816	0.9757	0.9927	0.9789	0.9844	7
150	47	0.9836	0.9783	0.9931	0.9806	0.9866	10
200	84	0.9833	0.9779	0.9931	0.9795	0.9871	25
250	90	0.9838	0.9786	0.9932	0.9811	0.9866	43
300	115	0.9830	0.9775	0.9938	0.9794	0.9865	72

从表 4 中可以看出，3.2 节中提出的结合 F1 值和 Metropolis 准则的学习器选择方法可以有效删除部分较差的学习器，为模型的训练节省了开销。当迭代次数为 150 时，选出 64 个基学习器，此时模型训练效果较好且运行时间较为合理。

因此选定迭代次数为 150，分别对权重指标 $\gamma=0$，$\gamma=0.25$，$\gamma=0.5$，$\gamma=0.75$ 和 $\gamma=1$ 进行实验，得到不同 γ 取值时模型的各评估指标值以及各种类基学习器的选择情况，如表 5 和表 6 所列。

表 5　不同 γ 取值的模型效果

Table 5　Model effects of different γ values

γ 取值	$F1$	*Accuracy*	*AUC*	*Precision*	*Recall*
0	0.9830	0.9775	0.9918	0.9805	0.9854
0.25	0.9831	0.9775	0.9928	0.9806	0.9855
0.5	0.9836	0.9783	0.9931	0.9806	0.9866
0.75	0.9830	0.9775	0.9917	0.9806	0.9855
1	0.9816	0.9756	0.9927	0.9789	0.9843

表 6　不同 γ 取值选取的基学习器个数

Table 6　Number of base learners selected for different γ values

学习器	$\gamma=0$	$\gamma=0.25$	$\gamma=0.5$	$\gamma=0.75$	$\gamma=1$
决策树	15	16	16	17	14
Logistic 回归	4	4	4	4	4
BP 神经网络	5	5	6	6	4
K-近邻	4	4	5	5	6
朴素贝叶斯	12	12	14	15	16
支持向量机	0	0	2	1	0
总个数	40	41	47	48	44

通过表 5 中所列出的结果可以得出，本模型在 γ 取不同值时相差不多，各项指标均在 0.95 以上，表现十分优秀。

因此，在保证模型性能的基础上，最大化基学习器差异性，有助于异质集成模型基学习器的多样性。从表 6 中可以看出，γ 取值为 0.5 时，模型性能最好，且基学习器种类的多样性最大。模型最终确定的初级基学习器种类及个数为决策树 16 个、K-近邻 5 个、朴素贝叶斯 14 个、Logistic 回归 4 个、BP 神经网络 6 个、支持向量机 2 个。

4.3 对比实验

为证明模型在提升了泛化能力和稳定性的前提下，提高了分类的准确性，将本文提出的 AHC-Metropolis 选择性集成模型与单一学习器、其他集成学习模型进行对比实验，将网贷之家平台数据集输入各个模型中进行训练和预测，实验结果如表 7 所列。

表 7 模型对比

Table 7 Model comparison

模型名称	*F*1	*Accuracy*	*AUC*	*Precision*	*Recall*
支持向量机	0.8084	0.6877	0.9140	0.6789	0.9788
朴素贝叶斯	0.8853	0.8321	0.9061	0.8055	0.9827
Logistic 回归	0.9073	0.8689	0.9409	0.8498	0.9732
基于熵的选择集成	0.9370	0.9153	0.9471	0.9189	0.9558
基于 Metropolis 的选择集成	0.9718	0.9550	0.9685	0.9468	0.9681
基于 AHC 的选择集成	0.9523	0.9381	0.9859	0.9682	0.9369
基于 AP 的选择集成	0.9380	0.9201	0.9710	0.9612	0.9151
AHC-Metropolis 选择性集成模型	**0.9836**	**0.9783**	**0.9931**	**0.9806**	**0.9866**

从表 7 中可以看出，本文提出的 AHC-Metropolis 集成学习模型的各项指标均表现较好，在预测准确性方面有着很大的提升。为验证模型在相关领域的其他数据集上仍然有效，我们使用了 Lending-Club 网贷数据集来测试此模型。Lending-Club 数据集是美国一家网贷公司的公开信贷数据，通过数据预处理和特征工程后，建模得到的结果如表 8 所列。

表 8 AHC-Metropolis 与单一学习器的对比(Lending-Club 数据)

Table 8 Comparison of AHC-Metropolis and single learner (Lending-Clubdata set)

模型名称	*F*1	*Accuracy*	*AUC*	*Precision*	*Recall*
支持向量机	0.9136	0.8535	0.8369	0.8423	0.9881
朴素贝叶斯	0.9478	0.9187	0.8817	0.9436	0.9522
Logistic 回归	0.9559	0.9288	0.8742	0.9196	0.9952
AHC-Metropolis 选择性集成模型	**0.9707**	**0.9532**	**0.9636**	**0.9452**	**0.9976**

从表 8 中可以看出，使用 Lending-Club 数据训练该模型得到的结果，可以证明该模型是可以运用于金融领域的其他典型场景的，在稳定性方面有着比较优秀的表现。

综上，本文提出的 AHC-Metropolis 集成学习模型预测效果有显著的提升，并且稳定性和泛化能力均较强，因此该模型可以运用于风控监管过程中，为衡量金融风控中的信用问题提供评价标准。

结束语 本文提出了 AHC-Metropolis 选择集成模型，模型主体是用 Stacking 集成框架，是一种典型的两阶段架构。使用凝聚型层次聚类(AHC)算法和模拟退火的 Metropolis 准则，可以在保证模型基学习器多样性的同时，有效筛选基学习器种类和个数，降低计算和存储的开销成本，提升模型预测和泛化的能力。综上，本文提出的 AHC-Metropolis 选择集成模型可以有效用于金融风险控制管理，为发展我国互联网金融行业和保障国家金融安全提供一定的应用价值。在今后的研究过程中，我们将对模型进行改进，注重多分类金融场景的数据研究，对基学习器的种类和个数筛选，以及次级学习器的选取方面进行更进一步的研究。

参 考 文 献

[1] HAND D J, HENLEY W E. Statistical lassification methods in consumer credit scoring [J]. Journal of the Royal Statistical Society, 1997, 160(3): 523-541.

[2] SHENG J. Credit card cash out detection scoring model based on Logistic [J]. Computer Applications, 2009, 29(11): 3088-3091, 3095.

[3] FANG K N, ZHANG G J, ZHANG H Y. Personal credit risk early warning method based on Lasso-logistic model[J]. Quantitative Economics and Technical Economics, 2014, 31(2): 125-136.

[4] ZHANG Y J, JIA H Y, DIAO Y F, et al. Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending[J]. Procedia Computer Science, 2016, 91: 168-174.

[5] PANG S L, HOU X Y, XIA L H. Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine[J]. Technological Forecasting and Social Change, 2021: 120462.

[6] LI X S, GUO Y H. Personal credit evaluation model based on Naive Bayes classifier[J]. Computer Engineering and Applications, 2006(30): 197-201.

[7] WEST D. Neural network credit scoring models[J]. Computers & Operations Research, 2000, 27: 1131-1152.

[8] LI Y, JIANG T Y, LIU Y R. Research on Internet Personal Credit Evaluation Based on Unbalanced Samples[J]. Statistics and Information Forum, 2017, 32(2): 84-90.

[9] PIERRE G, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[10] ZHOU Q Y. Application Research of Improved AdaBoost Algorithm in Credit Imbalance Classification[D]. Huangzhou: Zhejiang Gongshang University, 2020.

[11] FINLAY S. Multiple cassifer achitectures and their apication to credit risk asessment[J]. European Journal of Operational Research, 2011, 210(2): 368-378.

[12] SUN J, LI H, CHANG P C, et al. Dynamic credit scoring using B & B with incremental-SVM-ensemble[J]. Kybernetes, 2015, 44(4): 518-535.

[13] DŽELIHODŽIĆ A, DONKO D, KEVRIĆ J. Improved Credit Scoring Model Based on Bagging Neural Network[J]. International Journal of Information Technology & Decision Making, 2018, 17(6): 17.

[14] NASCIMENTO D S C, COELHO A L V, CANUTO A M P. Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study[J]. Neurocomputing, 2014, 138: 347-357.

[15] LESSMANN S, BAESENS B, SCOW H V, et al. Benchmarking stat-of-the-art lassification algorithms for credit scoring: an update of research [J]. European Joumal of Operational Rescarch, 2015, 247(1): 124-136.

[16] LIU C Z, MA D L, XIA Y F. Application of Dynamic Heterogeneous Integrated Credit Scoring Model in P2P Network Lending [J]. Financial Development Research, 2018(9): 24-31.

[17] QI H,WANG W J,GUO H S. A SVM Bagging ensemble method based on feature selection[J]. Small Microcomputer System,2014,35(11):2533-2537.

[18] LI Y J,GUO H X,LI Y N,et al. Classification of an ensemble learning algorithm based on Boosting in imbalanced data[J]. System Engineering Theory and Practice,2016,36(1):189-199.

[19] YU L,YANG Z B,TANG L. A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment[J]. Flexible Services and Manufacturing Journal,2016,28(4):576-592.

[20] WANG M,CAO Q,SUN J Z,et al. A method of user basic attribute prediction based on ensemble learning [J]. Small Micro Computer System,2020,41(12):2509-2515.

[21] CAO Z H,YU D X,SHI J F,et al. Two-layer classifier model applied to personal credit evaluation [J]. Control Engineering,2019,26(12):2231-2234.

[22] ZHOU Z H,WU J X,TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence,2002,137(1/2):239-263.

[23] ZHANG C X,ZHANG J S. Overview of selective ensemble learning algorithms [J]. Chinese Journal of Computers,2011,34(8):1399-1410.

[24] XIA Y F. A novel heterogeneous ensemble credit scoring model based on bstacking approach[J]. Expert Systems with Applications,2018,93:182-199.

[25] WU M H,GUO J S,JU Y,et al. Parallel selective ensemble algorithm based on hierarchical filtering and dynamic update [J]. Computer Science,2017,44(1):48-52.

[26] DU H L,ZHANG Y. Network anomaly detection based on selective ensemble algorithm[J]. The Journal of Supercomputing,2020(prepublish):1-22.

[27] YU J Y. Research on corporate credit risk assessment based on heterogeneous learner integration strategy [D]. Beijing:Central University of Finance and Economics,2019.

[28] YANG H. Design and research of risk control model of micro-online loan platform based on migration learning [D]. Beijing: Central University of Finance and Economics,2021.

[29] COHEN J. A Coefficient of Agreement for Nominal Scales[J]. Educational and Psychological Measurement, 1960,20(1):37-46.

WANG Mao-guang, born in 1974, Ph.D, professor, is a member of China Computer Federation. His main research interests include intelligent risk control models and algorithms, big data and intelligent software engineering etc.

JI Hao-yue, born in 1998, postgraduate. Her main research interests include Internet financial risk control and credit investigation.