论文题目：基于改进的自适应遗传算法与多模型融合的信贷风险预测模型研究

学位类别：工程硕士

学科专业：软件工程

年　　级：2018

研 究 生：王　祥

指导教师：楼新远

二零二一年五月

国内图书分类号：TP311 密级：公开
国际图书分类号：004

西 南 交 通 大 学
研 究 生 学 位 论 文

基于改进的自适应遗传算法与多模型融合的信贷风险预测模型研究

年　　级　2018
姓　　名　王祥
申请学位级别　硕士
专　　业　软件工程
指 导 教 师　楼新远

二零二一年五月二十三日

Classified Index: TP311
U.D.C: 004

Southwest Jiaotong University

Master Degree Thesis

RESEARCH ON CREDIT RISK FORECASTING MODEL BASED ON IMPROVED ADAPTIVE GENETIC ALGORITHM AND MULTI-MODEL FUSION

Grade: 2018
Candidate: Wang Xiang
Academic Degree Applied for : Master Degree
Speciality: Software Engineering
Supervisor: Lou Xinyuan

May.23,2021

西南交通大学

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权西南交通大学可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复印手段保存和汇编本学位论文。

本学位论文属于

1．保密□，在　　年解密后适用本授权书；

2．不保密□，使用本授权书。

（请在以上方框内打“√”）

学位论文作者签名：　　指导老师签名：

日期：　　日期：

西南交通大学硕士学位论文主要工作（贡献）声明

本人在学位论文中所做的主要工作或贡献如下：

1、本文根据脱敏的用户行为数据，挖掘信贷风险关联因素，并基于建立的序列浮动双向搜索算法选择了 198 维显著性特征，构建了基于深度神经网络 DNN、集成学习算法的用户信贷风险预测模型。

2、引入多模型融合技术，确定各基学习器预测结果的最佳权重。通过单模型和融合后模型的结果对比分析，多模型融合比单模型的预测结果有显著提升。

3、本文建立了基于交叉概率和变异概率的改进线性自适应遗传算法，并使用改进后的算法对四种集成学习方法进行优化，寻找最优的模型超参数组合。同时，从不同模型及模型的不同参数两方面构建融合模型，在测试集上取得了较好的预测效果。

本人郑重声明：所呈交的学位论文，是在导师指导下独立进行研究工作所得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本论文的研究做出贡献的个人和集体，均已在文中作了明确说明。本人完全了解违反上述声明所引起的一切法律责任将由本人承担。

学位论文作者签名：

日期：

摘要

互联网技术的飞速发展以及互联网技术与传统金融领域的结合，使得互联网金融相比于传统金融领域商业模式的优势也逐步显现出来，以 P2P 借贷、网络贷款等各种形式的互联网金融服务在我国呈现出生机勃勃的发展景象。在实际贷款过程中，平台过于依赖央行的征信系统，然而征信系统在数据时效性、全面性等方面存在明显的短板。如何快速、准确地评估个人信用状况，建立精准的信贷风险预测模型，是各金融机构风险控制的核心环节。

本文就用户信贷风险预测的问题，围绕用户的基本信息、消费记录、还款记录以及浏览行为，挖掘与信贷风险关联特征，引入 PolynomialFeatures 做特征构造，增强特征的表达能力，并基于建立的序列浮动双向搜索算法得到 198 维显著性特征。最终通过深度神经网络 DNN、集成学习算法和多模型融合技术建立了信贷风险预测模型。

针对标准遗传算法容易陷入局部最优和“早熟”的问题，通过对遗传算子中的交叉率和变异率做进一步的优化，建立了一种基于交叉率和变异率的改进线性自适应遗传算法（ILAGA）。本文通过对 GBDT 等集成学习算法的关键超参数进行研究，使用改进后的遗传算法对信贷风险预测模型进行优化，寻找其最优超参数组合，建立了 ILAGA-GBDT 等风险预测模型。通过仿真实验对比分析，优化后的单模型各项评价指标有了一定程度的提升。同时，为了增强个体学习器之间的差异性，引入深度神经网络 DNN，使用 Stacking 算法分别从多种模型和模型的不同参数两个维度进行差异性多模型融合。仿真实验对比显示，利用遗传算法改进后的集成学习做模型融合能够显著提高模型预测的准确性。相较于遗传算法改进前的单模型，最终融合模型的准确率、AUC 等指标提升超过 5%。最终模型在测试集上 AUC 值达到 0.97 以上，准确率超过 97%，预测高风险用户的查准率达 96%以上，对用户信贷风险有较好的预测效果，更适应于解决信贷风险预测的问题。

实验结果表明，本文研究的信贷风险预测模型具有良好的预测效果和泛化能力，对信贷风险评估具有一定的实际价值。

关键词：信贷风险预测；多模型融合；自适应遗传算法；集成学习

Abstract

With the rapid development of Internet technology and the combination of Internet technology and traditional financial field, the advantages of Internet finance compared with the business model of traditional financial field have gradually emerged. Various forms of Internet financial services, such as P2P lending and online loan, have shown a dynamic development scene in China. In the actual loan process, the platform relies too much on the credit investigation system of the Central Bank. However, the credit investigation system has obvious shortcomings in terms of data timeliness and comprehensiveness. How to evaluate individual credit status quickly and accurately and establish accurate credit risk prediction model is the core link of risk control of financial institutions.

Based on the user's basic information, consumption record, repayment record and browsing behavior, this thesis excavates the features associated with credit risk. The features are constructed in PolynomialFeatures to enhance the expression ability of the features, and 198-dimensional significance features are obtained based on the established sequential floating bidirectional search algorithm. Finally, a credit risk prediction model is established through deep neural network DNN, integrated learning algorithm and multi-model fusion technology.

Aiming at the problem that the standard genetic algorithm is easy to fall into local optimum and "premature", an improved linear adaptive genetic algorithm (ILAGA) based on the crossover rate and mutation rate is established by further optimizing the crossover rate and mutation rate in the genetic operator. In this thesis, the key hyperparameters of GBDT and other integrated learning algorithms are studied, and the improved genetic algorithm is used to optimize the credit risk prediction model to find the optimal combination of the hyperparameters, ILAGA-GBDT and other risk prediction models are established. Through the comparative analysis of simulation experiments, evaluation index of the optimized single model has been improved to a certain extent. At the same time, in order to enhance the difference between individual learners, the deep neural network DNN is introduced, and the Stacking algorithm is used to carry out the difference multi-model fusion from two dimensions of a variety of algorithms and different parameters of algorithms. The comparison of simulation experiments shows that model fusion using the improved ensemble learning of genetic algorithm can significantly improve the accuracy of model prediction. Compared with the single model before the improvement of genetic algorithm, the accuracy, AUC and other

indexes of the final fusion model were improved by more than 5%. The AUC value of the final model on the test set reached more than 0.97, the accuracy was more than 97%, and the precision of predicting high-risk users was more than 96%, which had a good prediction effect on users' credit risk, and was more suitable for solving the problem of credit risk prediction.

The experimental results show that the credit risk prediction model studied in this thesis has good forecasting effect and generalization ability, and has certain practical value for credit risk assessment.

Key words: Credit Risk Forecasting; Multi-model Fusion; Adaptive Genetic Algorithm; Integrated Learning

目 录

第 1 章 绪论

1.1 课题研究背景及意义

1.1.1 背景及意义

近年来，随着数据挖掘、机器学习等技术的热潮，国内互联网金融发展态势迅猛，形式越来越多元化。传统金融和机器学习等互联网技术的融合发展，相比于传统金融领域的模式，其优势也逐步显现出来[1]。我国 2015 年的政府工作报告明确指出，需加快推进传统金融机构与互联网元素的融合，并将其确定为我国未来发展的一项重大国家战略计划[2]。以 P2P（Peer to Peer，即个人对个人）在线网络贷款为例，2017 年，我国全年 P2P 在线借贷行业交易额总规模达 28049 亿元人民币，累积在线贷款机构数目逾六千家。随着国家对不良贷款平台整顿力度的加强，2018 年清退取缔众多网贷平台，在线贷款交易额急剧下降，但成交额规模仍达 17948 亿元，2020 年迅速增长至 22485 亿元[3]。

对于以 P2P 借贷、在线信贷为代表的互联网金融服务商，最大的运营风险源自信贷客户的违约信用风险。在对信贷客户的评估环节中，可利用平台自身收集的历史信贷数据作为依据外，中国人民银行的信用调查系统数据也可作为一个机构平台辅助的决策依据。目前，央行的信用调查系统已基本创建了包括企业和个人在内的电子信用档案数据库，截至到 2019 年 9 月，央行已收集自然人信息量超 9.9 亿人，覆盖范围广且基数已经非常庞大[4]。但其中仅有约 3 亿人存在历史往期的信贷记录，存在大量的数据记录缺失，其严重限制了信贷客户评估的应用范围，较难反映客户的真实信息，使得平台对用户的信用状况评估及风险管理存在较大的困难。如何快速、准确地评估个人信用状况，建立精准的信贷风险预测模型，是各金融机构风险控制的核心环节，更是其核心竞争力之一[5]。

本文对信贷客户的风险评估，具有重要的理论和现实意义。

（1）有助于提高 P2P、在线信贷等平台的风险控制能力

如上文所述，由于缺少对客户的征信数据，因此，仅凭征信来评估客户的信用风险往往存在较大的偏差。本文的研究基于用户的基本信息、银行卡消费记录、信用卡还款记录以及用户的浏览行为记录挖掘客户的信息，建立模型全面评估用户的信贷风险，从而提高金融平台的风险控制水平，降低平台的坏账率，减少损失，达到系统性防范的目的。

（2）对信贷风险模型的开发具有一定的借鉴意义

当前，我国政府正在逐步放开征信管制，各个互联网金融平台和民营征信机构创新个人信用风险评估模型，背靠大数据充分挖掘用户的相关特征，并与其业务紧密结合。本文研究中包括的关键特征、选取的模型算法等方面对该类模型的开发具备一定的借鉴意义。

（3）对于完善国内金融服务的模式有积极意义

目前国内现有的商业模式中，发展较好的有京东白条、花呗借呗等小贷服务，因为其有平台的交易、支付等众多数据作为支撑。在本文的研究中，可以通过客户的基本信息以及消费行为等重新评估贷款客户的信用风险状况，拓宽个人的融资渠道，拓展和完善国内金融的服务模式。

1.1.2 拟解决的问题

信贷风险管理一直是金融服务行业重要的组成部分，尤其对于互联网金融服务尤为重要。在当前传统的个人信用风险评估形式下，主要依赖于中国人民银行的信用调查系统电子信用档案[6]。如前所述，央行的征信系统存在大量的数据缺失，会大大限制评估客户的范围及评估的准确性。对于互联网金融企业，信贷业务由于用户的宏观因素变化等问题，导致用户风险模型预测准确率低，客户评级不准确，进而导致不良贷款余额和不良贷款率的增加。本文通过分析国内外的研究动态，从用户的基本信息、银行卡消费记录、信用卡还款记录以及用户的浏览行为出发，通过数据的探索、预处理，挖掘相关特征，将遗传算法与机器学习技术相结合，建立个人信贷风险评估模型，解决模型预测准确率低、对高风险用户预测查准率低的问题。同时为了解决遗传算法容易陷入局部最优和早熟的缺陷，建立了基于交叉率和变异率的改进线性自适应遗传算法，并使用该算法对集成学习方法等单模型进行优化，寻找其最优模型超参数。此外，运用多模型融合技术，从多种模型和模型的不同参数两方面增强个体学习器之间的差异性，以最优权重参数确定最终融合模型，取得较佳的预测效果，以降低信贷平台的坏账率损失，提高优质客户的收益，也在一定程度上完善央行的征信系统。

1.2 相关文献综述

1.2.1 信贷风险研究现状

世界范围内小额贷款最早可追溯到 20 世纪 70 年代的孟加拉农业银行，有效缓解了部分个人和企业的资金压力，此后该模式不断发展并趋于成熟，成为个人、企业甚至发展中国家减轻贫困的有效途径[7]。21 世纪以来，互联网的普及和互联网技术的发展，一种依靠网络的在线信贷开始逐步发展起来。

关于信贷风险和互联网金融，闫真宇、王汉君[8]等人提出在互联网金融运营中需要

解决传统金融以及互联网技术可能会带来的各项信贷风险。王裕粟[9]认为目前由贷款用户信用风险评估与控制导致的较高不良贷款余额与坏账率已经严重影响到我国金融行业体系的正常发展。建立精准的信贷风险评估模式不仅是各金融平台的迫切需求，更是其机构的核心竞争力。洪娟、曹彬、李鑫[10]等认为互联网技术与传统金融的结合，由于其非标准化、信息技术、安全体系以及法律和监管等问题，互联网信贷的风险评估、控制及管理等方面要比传统金融复杂的多。

对于借贷客户的信用风险评估，需借助统计学、运筹学和计算机科学等理论，挖掘该用户数据中所蕴涵的关于信用风险的特征和行为特点，建立相应的预测模型。周贤[11]在新形势下信贷风险管理问题研究中分析了当前阶段我国在信贷风险管理方面存在的一些问题，指出问题主要反映在信用对象以及行业问题两个维度。同时，针对这两个问题提出了新形势下提高信贷风险管理的有效措施。刘佳蒙[12]在商业信贷风险管理存在的问题与对策研究中分析了强化商业信贷风险管理的意义，并从社会环境和认知因素这两大方面的因素下分析了其中存在的问题。徐溪蔓[13]在基于时间序列模型的商行信贷规模与风险管理分析中提出了信贷/GDP 比率这一指标，这个比率能够反映出未来信贷规模的长期走向，并根据我国商业银行近十年的数据探索了未来信贷风险的管理和监管方向。李帅鹏[14]在基于贝叶斯决策规则的商业银行信贷风险研究中，利用大数据样本分析，建立贝叶斯信贷风险评估模型，通过直观性的预测结果制定各类客户风险评分值，并根据风险评分值确定是否发放贷款。

1.2.2 互联网金融风险研究现状

二十一世纪以来，随着互联网的发展，互联网+金融成为金融领域的一个最新发展趋势。基于智能移动设备的普及，用户可直接完成金融服务的注册并使用其金融产品，如京东白条、360 借条等。当前发展的互联网金融提供的服务和产品总共包括两种形式，一种是同时提供线上和线下金融服务或产品的销售运营等；另一种是新兴的互联网金融业务类型，只针对互联网用户在线上进行销售及服务[15]。互联网金融的出现拓宽了个人和企业的融资渠道，拓展和完善了国内金融的服务模式。互联网金融给社会生产生活带来极大便利的同时，表现出了比传统金融服务更大的风险，且这种风险往往是不可逆的，更容易引起金融恐慌。

（1）互联网背景下的金融传统风险

互联网背景下的金融传统风险，首当其冲的是市场风险。金融市场除了本身具备较大的市场风险，还包括许多外部风险。我国的金融市场虽处于国家的监管之下，但主要依赖于市场经济的管控。其次是流动性风险。当金融机构由于发展方向等原因导致资金供给不足无力偿还用户，此时金融机构便会陷入危机，甚至倒闭，并且这种风险往往是不可逆的[16]。最后是贷款客户的信用风险。对于互联网金融的信贷业务，用户只

需要注册并提供身份信息证明，即可从平台借贷到一定数量的资金，而忽视了对借贷用户的可用风险的严格评估。若平台没有充足的资金链，一旦资金链断裂，部分公司无法承担后果只能宣布其破产处理。

（2）互联网背景下的金融新风险

在互联网发展的新形势下，新的金融风险也应运而生，这些新的风险正在制约着互联网金融的发展。新形势下的金融新风险，第一个是技术风险。线上的互联网金融依托于智能移动设备上的 APP 软件来提供相应的服务，尤其对于高并发量的 APP 来说，技术可能会成为一个制约发展的瓶颈[17]。此外，对于提供信贷服务的平台来说，如何建立精准的评估贷款客户违约风险模型是一个核心的技术难题。第二个是数据风险。互联网金融交易均是通过对数据信息的处理来实现对资金的管理。在交易过程中，互联网内部会收集大量的交易数据保存在数据库中，若在交易过程中或事后出现数据库数据的丢失或者出错，则相应的资金往来交易状况便会出现混乱，造成无法想象的严重后果。

综上所述，由于我国一些法律法规和规章监管制度尚未完善，互联网金融仍存在较多的风险和危机诱导因素，需要政府和互联网金融机构的共同努力，进一步完善我国监管制度，稳定我国互联网金融市场的秩序[18]。

1.2.3 信用风险评估模型研究

对于信贷客户的信用风险评估，需借助统计学、运筹学和计算机科学等理论，挖掘该用户的数据中所蕴涵的关于信用风险的特征和行为特点，建立相应的预测模型[19]。

目前国内外关于信用风险评估模型主要有如下几种：

（1）专家类模型

专家类模型是一种传统的信贷风险评估方法，在互联网时代未到来前，信贷用户的数据和历史资料很难全面获取到，对客户的信用风险等级评估依赖于信贷专家的主观判断[20]。如“5C”法，挑选信贷风险评估的专家，对借款人的品格、偿还能力等五个方面分别进行评估，并赋予其合理的指标权重，最终做出判断。此外，常用的专家类模式还包括 LAPP 法、CAMEL 体系等。专家类模型在指标处理上具有突出优势，且灵活性较好，但此模型过多依赖信贷风险评估专家的主观判断，且评估效率低下。

（2）Logistic 模型

Logistic 模型是目前应用较多的信贷风险评估模型，它的模型简单、计算速度快、鲁棒性强、可解释性好，并且可直接预测出借款人发生违约的概率。陈晓兰、任萍[21]在对企业信用风险评价研究中，通过将层次分析法和 Logistic 这两种模型结合构建组合模型，并对国内部分采样的企业信用风险进行了测评评估。实验结果证实该组合模型对企业的信用风险评估具备重要的参考意义。喻光丽[22]在对网贷平台借款人的信用

进行评估中，通过数据处理、特征提取、显著特征筛选出最终特征指标，再运用 Logistic 算法进行训练测试建模，得到最终的在线信用贷款风险评估模型，为网贷平台提供了一定的参考意义。呼振凯[23]在网贷中借款人的信用风险评估研究中构建了与用户在线信用风险关联的指标作为训练的特征，并运用 K-Means 算法将样本做聚类分析，并根据聚类结果从样本中选择其中最合适的正负样本，最后使用逻辑斯蒂回归算法进行训练预测。

（3）神经网络模型

神经网络模型是一个复杂的非线性学习函数，在多个领域可以达到较好的预测效果[24 25]。神经网络最初应用于金融方面是 Dutta 等在债券信用评级引入的[26]。1994 年 Altman[27]在公司的财务危机中引入了神经网络作为预测模型，取得了比较好的效果。方先明、熊鹏、张谊浩[28]在信用风险评价模型研究中通过 Hopfield 神经网络建立了对用户信用风险等级的评估模型。实验证明，该模型相较于其他模型能够更显著地反映出样本的特征，更适用于信用风险的评估。张佳维[29]在我国商业信贷现有的用户风险评估特征的基础上，使用了模糊神经网络算法建立模型，为商业信贷的个人及企业的信用风险评估提供了更多的依据和参考价值。

1.3 创新点

本文研究针对当前评估模型可能导致信贷风险评估出现偏差，高风险用户识别查准率低的现状，基于深度神经网络 DNN 和集成学习算法建立了个人信用风险评估模型。为了提高模型预测的准确性，运用 Stacking 方法构建融合模型，以最优权重确定最终模型。同时，为了进一步优化模型效果，使用改进后的基于交叉率和变异率的遗传算法对集成学习算法的超参数进行优化，完成了基于用户基本信息、银行卡消费记录、信用卡账单记录和用户浏览行为的个人信贷风险评估模型的设计与实现。本文研究的创新点主要包括如下三个方面：

（1）在信贷风险预测模型中，为了提升模型的预测精度和风险评估的准确性，基于 Stacking 算法对深度神经网络 DNN、集成学习算法构建融合模型，从多种模型和模型的不同参数两方面进行差异性模型融合，以最优权重确定最终模型，模型的效果有了显著提升。

（2）本文建立了一种序列浮动双向搜索算法 SFBSA。在该算法搜索特征子集时采用两种不同的重要性度量方式，避免了被单一重要性度量所约束，在一定程度上减小了陷入局部最优的缺陷。

（3）针对标准遗传算法容易陷入局部最优的缺陷，建立了基于交叉率和变异率的改进线性自适应遗传算法，并使用改进后的遗传算法对四种集成学习方法进行优化，寻找最优模型超参数组合，降低陷入局部最优和“早熟”的缺陷。结果表明使用遗传算

法改进后的集成学习相较于改进前有了一定程度的提升。同时，使用多模型融合技术构建融合模型，在测试集上取得了较好的预测效果。

1.4 论文组织结构及章节安排

本文围绕国内互联网金融平台信贷风险模型评估中存在的问题，结合用户信贷业务流程，利用深度神经网络、集成学习、多模型融合、遗传算法等机器学习技术建立行之有效的信贷风险评估预测模型，用以预测借贷客户的违约风险，减少不良贷款率，降低平台的风险。此外，本文通过对 GBDT 等四种集成学习算法的超参数进行研究，利用改进后的基于交叉率和变异率的线性自适应遗传算法对其进行优化，寻找最优训练超参数并进行多模型融合，对客户的信贷风险进行了精确的评估。具体地，论文的技术路线及结构框架安排如图 1-1。

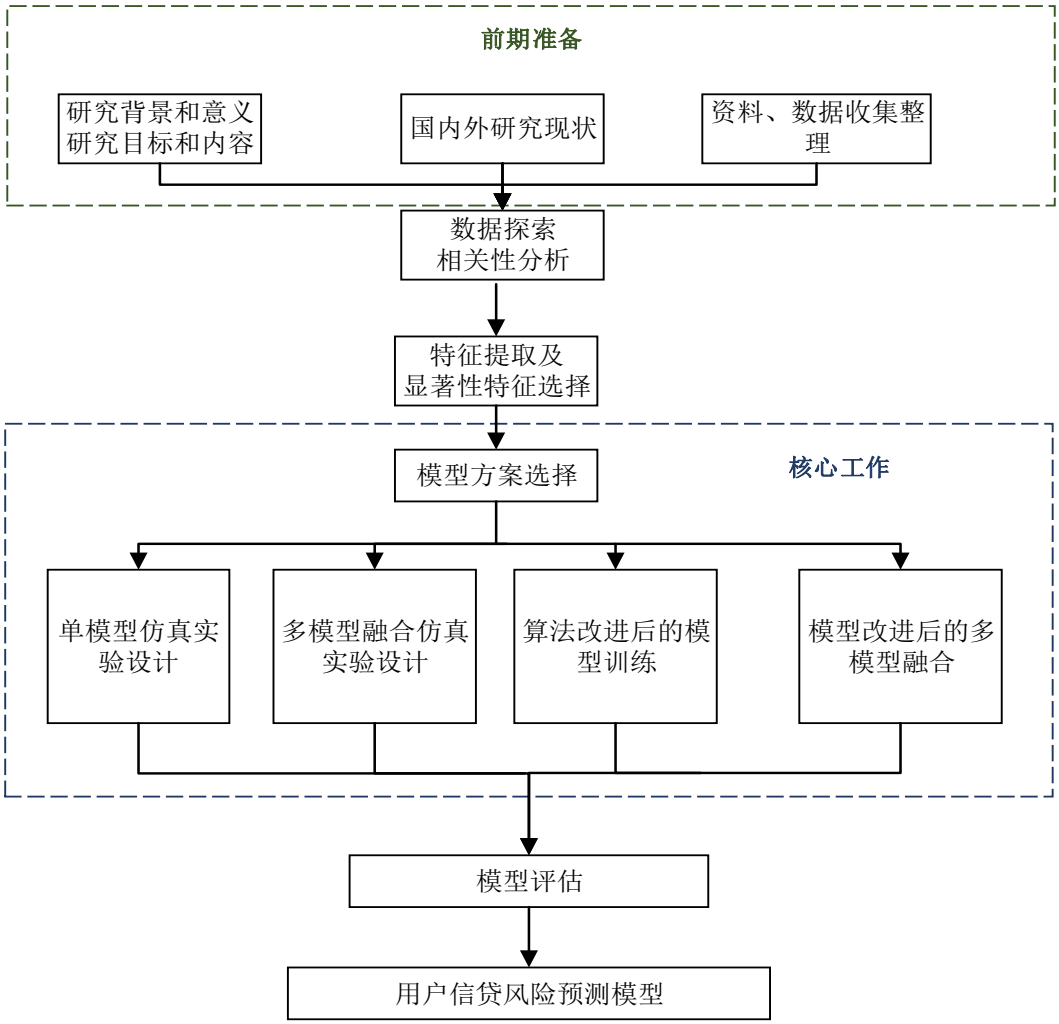


图 1-1 论文技术路线及结构框架图

本论文共五章，主要内容如下：

第一章首先介绍了我国信贷风险评估模型研究的背景和意义。其次，在国内外研

究综述中，简要介绍了信贷风险模型在国内外研究的现状。最后介绍了论文的三个主要创新点和组织结构框架。

第二章主要是本文研究的相关知识介绍。首先是对集成学习思想的概述，主要包含了序列集成学习算法和并行集成学习算法，分析对比了它们的不同点。其次是机器学习算法的基础理论研究介绍，主要包含了 GBDT 等集成学习方法。再次是对深度学习算法中的深度神经网络 DNN 进行简单介绍。最后是模型融合的理论概述、数据集和预测模型的评价指标。

第三章主要为信贷风险预测模型研究的实验部分，包含数据的探索和预处理、特征工程构建、深度神经网络 DNN 和集成学习算法单模型的仿真实验对比分析以及差异性多模型融合的仿真实验，多模型融合主要从多种模型和模型的不同参数两个维度加大个体学习器之间的差异性，提升模型融合的整体效果。

第四章重点介绍了改进的线性自适应遗传算法的过程以及使用该算法优化 GBDT、等四种集成学习算法超参数的过程。针对四个模型确定不同的域空间，优化不同的超参数组合，有效提升了单模型的泛化能力。同时，使用 Stacking 从多种模型和模型的不同参数提升模型融合的效果，并确定最终的信贷风险预测模型。

第 2 章 相关理论基础及数据集介绍

本章节是关于机器学习中集成学习算法预测模型、深度神经网络 DNN、多模型融合理论、数据集以及预测模型相关评价指标的内容介绍。

2.1 集成学习算法概述

在有监督学习任务中，由于模型在训练时存在偏好，所以实际情况有时并不理想，这时可通过多个不同学习器之间的合作、取长补短来完成分类或回归的学习任务，集成学习的思想正是基于这一点[30]。集成学习就是通过算法和训练集产生若干个基学习器，再使用某种结合策略将它们集成在一起作为整体的学习器，该学习器往往具有更好的泛化能力，同时也可降低陷入局部极小点的风险[31]。

集成学习方法根据基学习器间的关系，可分为两类，一种是，在该集成学习方法中，第 N 个学习器依赖于前 N-1 个学习器的学习效果串行生成，每次通过赋予上一轮分类错误样本更高权重来提高模型对样本的拟合能力[32]。另一种是个体学习器之间不存在强依赖关系，可独立地并行生成。如 Bagging 通过 Boostrap 方法对样本做有放回地抽样并使用算法做相应的训练以获得不同的个体学习器。表 2-1 为两类集成学习算法 Boosting 和 Bagging 从样例权重等方面的细节对比。

表 2-1 Boosting 和 Bagging 算法的细节对比

集成学习算法	样例权重	预测函数	并行计算
Boosting	每一轮根据拟合情况调整训练集的权重	每个个体学习器根据训练情况拥有不同的权重	不支持
Bagging	训练过程中每个样本的权重相等	所有个体学习器权重相等	个体学习器之间可并行计算

2.2 机器学习算法理论

2.2.1 序列集成算法介绍

集成学习中的序列集成算法属于迭代算法，在学习过程中，通过不断地使用一个弱学习器弥补之前学习器的不足来串行地构造一个较强的学习器，代表是 Boosting 算法。Boosting 算法的基本思想是先赋予每个训练样本相同的权重，之后进行 N 次迭代。在每次迭代中，对分类错误的样本加大权重，使得学习器在下一次的迭代中更加关注这些样本[33]。Boosting 系列算法中最著名的有 AdaBoost 和提升树系列算法，包括 GBDT、XGBoost 等。Boosting 算法的过程示意图如图 2-1 所示，算法每次迭代生成一个基学习

器 f_i 及其权值 w_i，并根据基学习器的预测效果重新更新样本的权值。当所有基学习器训练完成后，将其以对应的权重集成在一起形成融合模型。

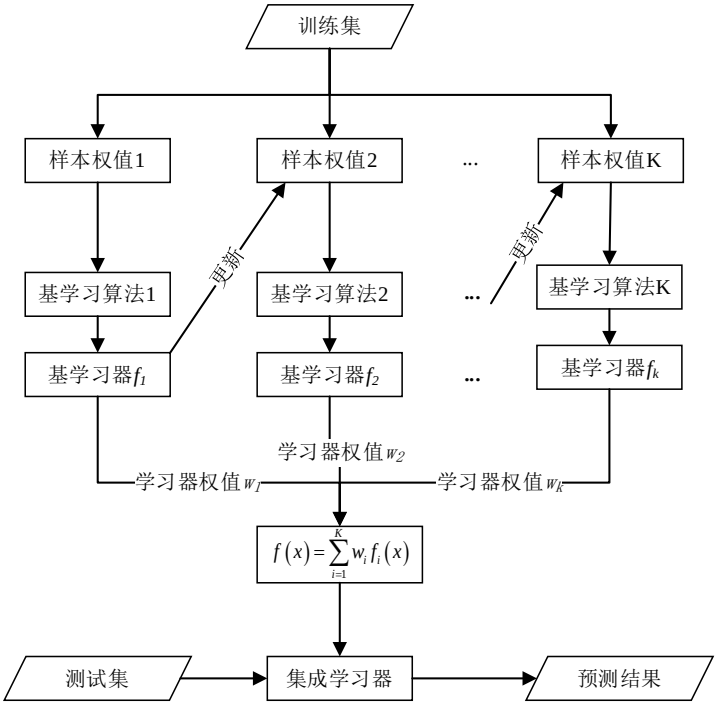


图 2-1 Boosting 算法过程示意图

（一）GBDT 算法

GBDT 是一种序列集成学习的决策树模型，在每一轮模型训练中，梯度提升每一次的计算是为了减小之前决策树所产生的残差，以使得总体模型具有更好的数据拟合能力，提高模型的训练效果[34]。GBDT 做分类或回归任务均可。

（1）梯度提升算法

梯度提升算法在模型训练迭代的过程中，随着个体学习器的数量增多，损失函数的值显著降低[34]。假设已训练到第 N 轮，前 N-1 轮个体学习器融合得到的强学习器为 $f_{t-1}(x)$，在训练时的损失函数为 $L(y, f_{t-1}(x))$，本轮的基学习器为 $h_t(x)$，则本次的损失函数 $L(y, f_t(x))$ 为：

$$L(y, f_t(x)) = L(y, f_{t-1}(x) + h_t(x)) \tag{2-1}$$

强学习器的公式如式 2-2，其中 $h_i(X)$ 为各基学习器。

$$f_t(X) = \sum_{i=1}^{t} h_i(X) \tag{2-2}$$

假设损失函数为 $L(y, f_t(x))$，使用贪婪的方式保证每一次新加入子模型 $h_t(x)$ 后损失函数值减少，如公式 2-3 和 2-4 所示，并利用梯度下降的方式进行残差拟合。

$$f_t(X) = f_{t-1}(X) + h_t(X) \tag{2-3}$$

$$L(y, f_t(X)) < L(y, f_{t-1}(X)) \tag{2-4}$$

（2）决策树算法

决策树模型本质为划分数据集，树的根节点包含全部样本，每个叶子节点包含的是同种类别的样本[35]。决策树的学习过程关键点在于如何确定最优划分属性和最优属性值[36]。以 ID3（Iterative Dichotomiser 3）为例，选取信息增益最大的特征作为决策树的划分属性将样本分到不同的结点中。

信息熵 $Ent(D)$定义如式 2-5，其中 p_k 表示样例集合 D 中第 k 类样本所占比例。

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{2-5}$$

属性 A 在数据集 D 上产生的信息增益 $Gain(D, A)$ 公式如 2-6 所示。

$$Gain(D, A) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2-6}$$

其中，D^v 表示在属性 A 的 V 个分支节点中第 v 个属性值包含的样本。

GBDT 算法中决策树的分裂方法可分为两种，深度优先和广度优先。深度优先是采用递归思想按最大收益叶子生长的方式，广度优先是按照层次来构建树。

按深度优先的方式可以花费较小的代价构建最终的决策树。这种方式的优点是精准度高，拟合数据的能力强，且可以在较短的时间内完成树的生长[37]。另外，树节点构造的过程是顺序的，无法直接进行并行计算。图 2-2 是决策树按叶子生长的过程。

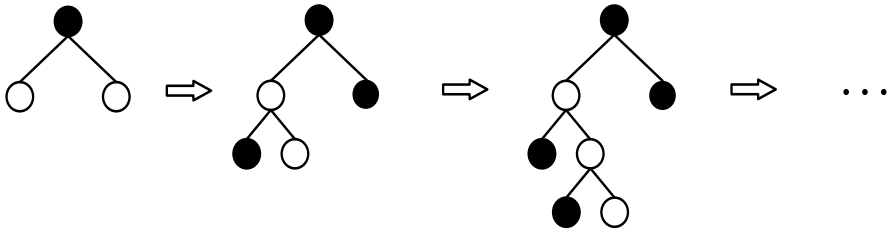


图 2-2 按叶子生长分裂的决策树

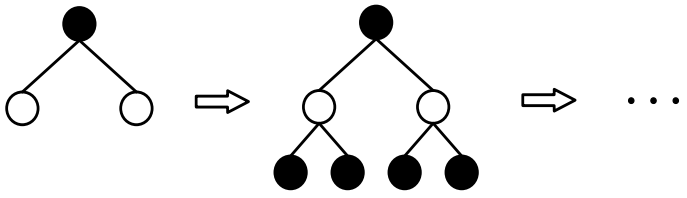


图 2-3 按层分裂的决策树

按广度优先的方式（按层生长的方式）指每一层的每一个结点都进行分裂，因此以这种构造方式可以并行加速计算，但在训练时会产生多余的分裂节点[38]。此外，模型

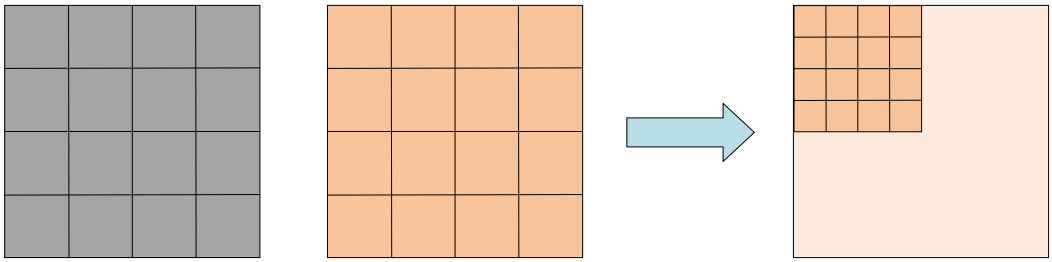
在训练过程中每次迭代都需要遍历整个数据集，对运算设备的运行内存要求较高[39 40]。图 2-3 是按层分裂的决策树生长过程。

（二）LightGBM 算法

LightGBM 是由微软开发、基于梯度提升树的优化算法，其使用的是 histogram 算法，模型训练时消耗的内存更低，计算速度更快[41 42]。为了降低算法在训练时内存的占用以及计算的代价，LightGBM 在 XGBoost 的基础上使用了直方图优化算法、按叶子生长算法、差加速等[43]。相较于其他提升树算法，在内存占用、训练速度和准确性方面均有较大幅度的提升。

（1）直方图算法优化

相比于其他提升树算法遍历每一个分割点计算分裂增益，直方图算法将连续的浮点数据转换为 bin 数据，实现特征浮点数据离散化，并最终用直方图的形式表示，整个过程将大规模的数据映射到了直方图上[44]。训练时只需保存特征离散后的值，桶 bin 值的存储比原本训练样本所占内存大大减少，内存消耗可以降低为原来的八分之一[45]。



左侧为int_32存储索引和float_32 存储值
右侧为uint8_t存储#*bin*值

图 2-4 LightGBM 直方图算法内存优化图

（2）LightGBM 模型决策树的查找最优分裂点原理

LightGBM 查找最优分裂点的过程如表 2-3 所示。

（三）三种 Boosting 算法细节对比

表 2-2 为 GBDT、XGBoost 以及 LightGBM 算法的细节对比，包括树生长模式等。

表 2-2 GBDT、XGBoost 和 LightGBM 算法的细节对比

模型细节	GBDT	XGBoost	LightGBM
树生长模式	按层生长	按层生长	受深度限制的按叶子生长
分裂点搜索方式	直接遍历	特征预排序	直方图算法
内存开销	大	大	小
类别特征	需做特殊处理	One-hot 编码	直方图数据处理
并行	不支持	特征并行	特征、数据并行
是否支持缺失值处理	不支持	支持	支持

表 2-3 LightGBM 查找最优分裂点过程

算法：**FindBestSplitByHistogram**
输入：训练集X,模型$T_{c-1}(X)$
1： **for all** 叶子p in $T_{c-1}(X)$:
// 为每个特征构造直方图
2： **for all** f in $X.features$:
3： H = *new Histogram*()
4： **for** i in (0, num_of_row)
5： $H[f.bins[i]].g$ += g_i // 获取每个*bin*的梯度之和
6： $H[f.bins[i]].n$ += 1 // 获取每个*bin*中样本数量
7： **end for**
// 从直方图寻找最佳分裂点
8： **for** i in (0, $len(H)$):
// G_L和n_L代表当前*bin*左侧所有*bin*的梯度和样本数量
9： G_L += $H[i].g$; n_L += $H[i].n$
// 与父节点数据进行差计算获得右侧*bin*梯度和G_R和数量n_R
10： $G_R = G_P - G_L$; $n_R = n_P - n_L$
11： $\Delta loss = \frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} - \frac{G_P^2}{n_P}$
12： **if** $\Delta loss > \Delta loss(p_m, f_m, v_m)$:
13： $(p_m, f_m, v_m) = (p, f, H[i].value)$
14： **end if**
15： **end for**
16： **end for**
17： **end for**

从 LightGBM 的伪代码可看出，LightGBM 在寻找最优分裂点时，先为每个特征构造直方图，再遍历所有样本，计算每个 bin 中的梯度之和和样本数量，并使用差加速[46]计算右边结点的梯度和及样本数量，最后在遍历过程中取最大的信息增益，以此时的特征和 bin 的值作为分裂结点的特征和分裂特征取值[47]。

2.2.2 并行集成算法介绍

并行集成算法中，其个体学习器间不存在强依赖关系，在训练时个体学习器可独立并行生成。Bagging 属于并行式集成学习算法，Bagging 的基本思想是对整体数据集做自助采样法形成 T 个训练数据集，然后针对上述数据集利用相应的算法训练出 T 个个体学习器，再使用某种结合策略将它们集成在一起作为整体的学习器[48]。算法的过程示意图如图 2-5 所示。以随机森林为例，Random Forest 样本采样方式同样基于自助采样法生成训练集，再利用决策树相关算法生成基学习器，基于结合策略融合成为随机森林，因此随机森林在训练时更注重降低模型的方差[49]。

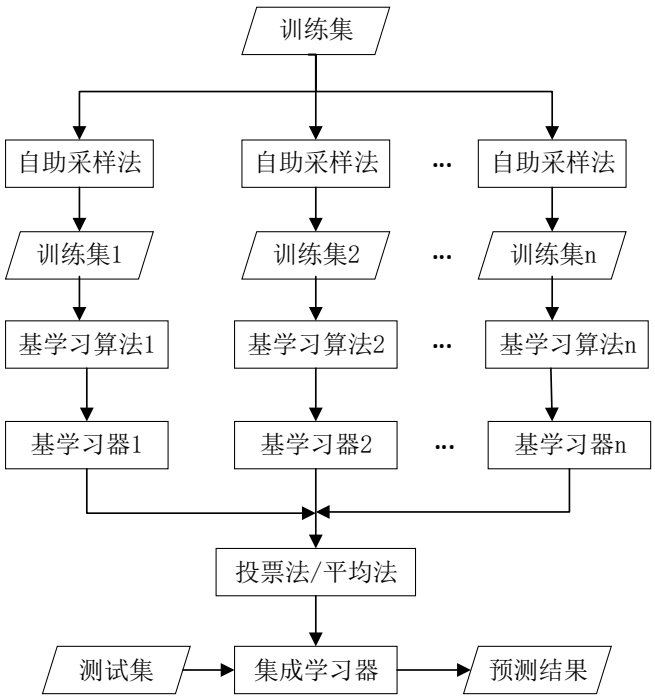


图 2-5 Bagging 算法过程示意图

2.3 深度学习算法理论

深度学习是机器学习的一个分支，其来源于人工神经网络的研究，是基于人工神经网络的机器学习方法的一部分。学习可以是有监督、半监督或是无监督。深度学习的网络架构使得深度学习特别适合处理含有较多变量的问题。现已证明，深度学习在图像识别和自然语言处理等领域非常有效。深度学习算法包括卷积神经网络（CNN）、循环神经网络（RNN）、深度神经网络（DNN）等。以深度神经网络 DNN 为例。

深度神经网络是基于感知机的扩展，有较多的隐藏层网络。DNN 的内部神经网络层可分为三类，分别为输入层、隐藏层和输出层。一般来说，第一层是输入层，最后一层是输出层，中间的层都是隐藏层。如图 2-6 所示。

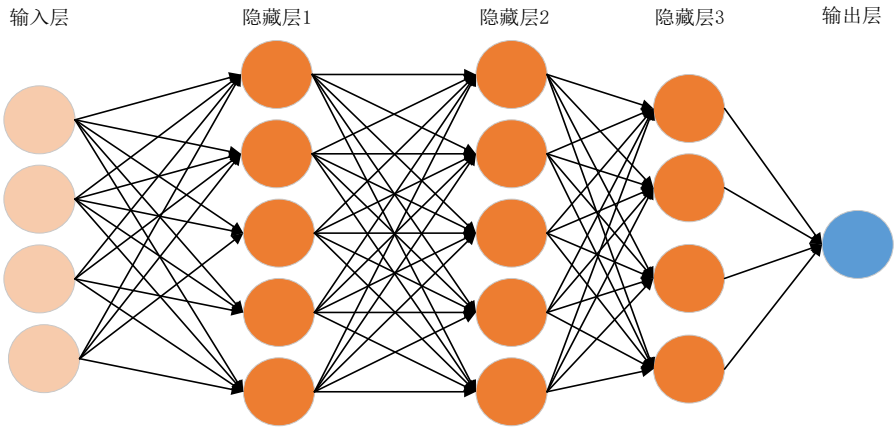


图 2-6 DNN 示意图

如图 2-6 所示，层与层之间是全连接的，第 i 层的任意神经元一定与第 i+1 层的任意神经元相连。从局部来看，其神经元为一个线性关系，即：

$$z = \sum w_i x_i + b \tag{2-7}$$

其中，w_i 为线性关系系数，b 为偏移。

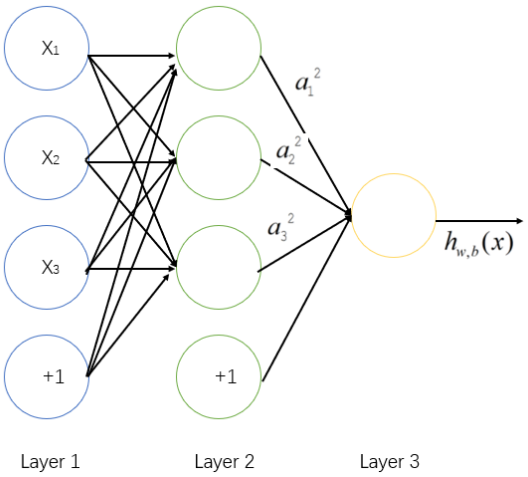


图 2-7 DNN 前向传播算法

以图 2-7 的 DNN 前向传播算法为例，$w^i{}_{mn}$，i 表示线性系数 w 所在的层数，而下标 m 和 n 对应的是输出的和输入的层的索引。我们可以利用上一层的输出计算下一层的输出。从图 2-8 可以看出，对于第二层的输出 a_1^2，a_2^2，a_3^2，有：

$$\begin{aligned} a_1^2 &= \sigma(z_1^2) = \sigma(w_{11}^2 x_1 + w_{12}^2 x_2 + w_{13}^2 x_3 + b_1^2) \\ a_2^2 &= \sigma(z_2^2) = \sigma(w_{21}^2 x_1 + w_{22}^2 x_2 + w_{23}^2 x_3 + b_2^2) \\ a_3^2 &= \sigma(z_3^2) = \sigma(w_{31}^2 x_1 + w_{32}^2 x_2 + w_{33}^2 x_3 + b_3^2) \end{aligned} \tag{2-8}$$

其中，σ 为激活函数。一般地，假设第 $l-1$ 层共有 m 个神经元，则对于第 l 层第 j 个神经元的输出 a_j^l 计算公式如下：

$$a_j^l = \sigma\left(\sum_{k=1}^{m} w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{2-9}$$

依据此公式，选择合适的激活函数，从输入层开始，一层一层向后计算，一直到输出层，得到输出结果值。

2.4 模型融合理论

机器学习理论中的 NFL 定理告诉我们，不存在某个算法一定优于另外一个算法，算法的优劣仅体现在针对某个领域或某个问题时其是否会产生较好的效果[50]。因此我们可以以适当的方式将多个模型融合，发挥各个模型的长处，扬长避短，提升模型结果

的准确度。模型融合就是训练多个模型，按照某种策略集成成为一个更复杂、泛化性能更好的模型。模型融合的策略有如下几种：

（1）投票法

这是一种最简单的模型融合方法，采取投票机制的方法，依据每个类别的投票票数确定最终的分类结果[51]。

（2）平均法

对于回归任务来说，最简单的融合就是平均法，最终的结果为所有模型结果的平均，如加权平均法等。

简单平均法：

$$H(x)=\frac{1}{T}\sum_{i=1}^{T}h_i(x) \tag{2-10}$$

加权平均法：

$$H(x)=\sum_{i=1}^{T}w_i h_i(x) \tag{2-11}$$

其中，w_i 是个体学习器 h_i 的权重，T 是模型的个数。

（3）Stacking 算法

除了投票法和平均法，还可以使用一个更强大的方法，再利用一个学习算法的训练过程作为结合方式将各个个体学习器进行组合，称之为学习法，如 Stacking[52]。

表 2-4 Stacking 算法流程

算法：**Stacking**
输入：训练集D={$(x_1,y_1),(x_2,y_2),...(x_m,y_m)$}； 初级学习算法$\Im_1,\Im_2,...\Im_T$； 次级学习算法$\Im$. 过程： 1：**for** t=1,2,...T do 2：$h_t=\Im_t(D)$ 3：**end for** 4：$D'=\varnothing$； 5：**for** i=1,2,...m do 6：**for** t=1,2,...T do 7：$z_{it}=h_t(x_i)$； 8：**end for** 9：$D'=D'\cup((z_{i1},z_{i2},...z_{iT}),y_i)$； 10：**end for** 11：$h'=\Im(D')$； 输出：$H(x)=h'(h_1(x),h_2(x),...,h_T(x))$

Stacking 算法的步骤如下：

Step1:根据已获得的训练数据集使用 N 个相同或不同的初级学习算法训练获得 N 个初级学习器。

Step2:对于测试数据集中的样本，用 N 个初级学习器分别预测得到对应的输出，类别标记为它原来的标记，并用次级学习器进行训练。

具体地，Stacking 算法的流程伪代码如表 2-4 所示。模型训练分为两层，首先利用训练集训练出 N 个个体学习器，称为初级学习器。再将初级学习器的输出与原来样本的标签拼接作为第二层次级学习器的输入进行训练，以寻求最优的权重系数。

Stacking 的思想如图 2-8，图中所采用的为五折交叉验证：

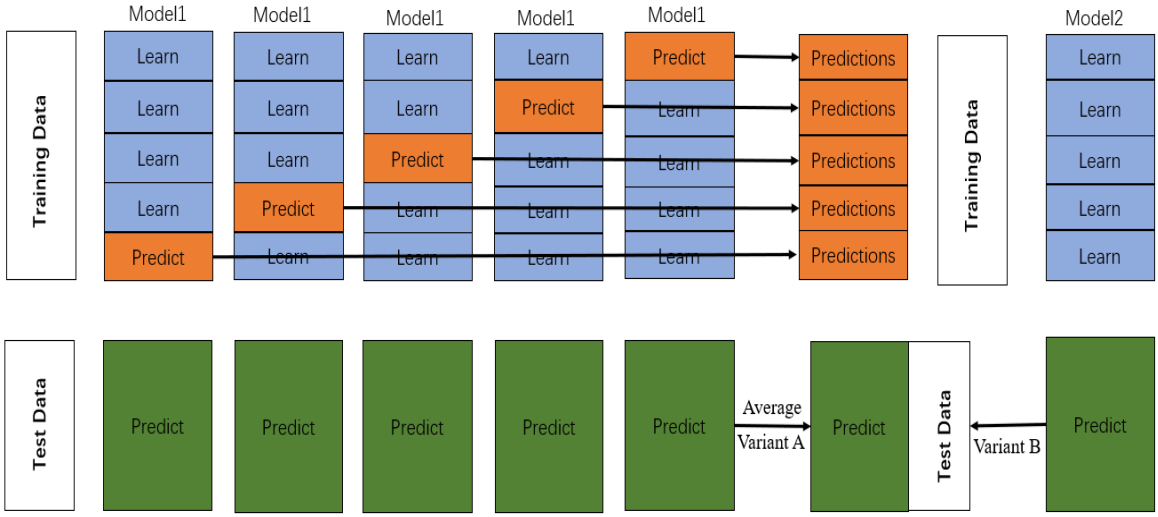


图 2-8 Stacking 五折训练图

图 2-8 中使用了五折交叉验证，融合模型的第一层中，利用 Training data 训练并输出模型的预测结果，并将预测结果作为输入进入到第二层。此时，使用次级学习器进行训练及预测。通常来说，第二层的次级学习器为了避免陷入过拟合的风险会采用 LR 等简单的学习器[53]。

2.5 数据集介绍

本文信贷风险预测模型研究的数据来源于融 360 数据挖掘竞赛。所用数据包括六张表，分别是基本信息表、历史违约记录信息表、浏览行为信息表、银行卡收入支出流水记录表、平台发放贷款时间信息表、信用卡历史账单记录信息表，总共约 2000 万条数据，其中包含共 55596 个用户，用户浏览行为数据约 1200 万条，信用卡账单约 200 万条，银行卡流水记录约 600 万条。其中存在逾期记录的用户数为 22230，占比约 40%。训练数据和测试数据都做了相应的脱敏处理。下面分别介绍所用表的详细字段及示例数据。放款时间与逾期记录合为一表。

（1）用户基本信息表

用户基本信息表包含用户的个人信息状况，包括用户性别、用户职业、用户教育程度、用户婚姻状态、用户户口类型等，表中均为脱敏数据，且这五个字段的值均为离散、无序的。用户基本信息表如表 2-5 所示。

表 2-5 用户基本信息表

用户标识	用户性别	用户职业	用户教育程度	用户婚姻状态	用户户口类型
3150	1	2	4	1	4
6965	1	2	4	3	2
1265	1	3	4	3	1
6360	1	2	4	3	2

（2）用户放款时间表及逾期记录表

用户放款时间表中记录了为每个用户放款的时间。用户逾期表记录了用户逾期记录。利用放款时间可将用户的行为分为放款前和放款后以及不区分放款前和放款后。有逾期记录的用户表明其可能具有较大的信贷风险，应予以重点关注。如表 2-6 所示。

表 2-6 用户放款时间及逾期记录表

用户标识	放款时间	是否逾期
1	68458	0
2	68458	0
3	68458	0
4	68458	1

（3）用户浏览行为记录表

用户的浏览行为与信贷风险的评估也存在直接联系。在本文的研究中，用户的浏览行为记录包含浏览时间、浏览行为编号、浏览子行为编号。用户的信贷风险往往也体现在用户的浏览行为中，且这部分的数据量占比达 60%以上，包含的信息量较大。表 2-7 为用户浏览行为记录表。

表 2-7 用户浏览行为记录表

用户标识	浏览时间	浏览行为数据	浏览子行为编号
34801	68588	173	1
34801	68588	164	4
34801	68588	38	7
34801	68588	45	1

（4）银行卡历史账单流水记录表

对于用户的银行卡历史账单流水记录表，主要包含流水时间、交易金额等。用户银行卡的账单及收入直接关系到用户的消费能力、还款能力等，与信贷风险直接相关。但该表中存在较大比例的用户缺失，80%以上的用户不存在银行卡流水记录。表 2-8 为用户的银行卡历史流水记录表。

表 2-8 银行卡历史流水记录表

用户标识	流水时间	交易类型	交易金额	工资收入标记
6965	68221	0	13.756664	0
6965	68221	1	13.756664	0
6965	68258	0	14.449810	0
6965	68258	1	10.527763	0

（5）信用卡历史账单记录表

用户的历史信用卡账单表包含上期信用卡账单以及本期信用卡账单等记录。包括时间、上期账单金额、上期还款金额、信用卡额度、本期账单金额、还款状态等。用户的信用卡历史账单记录与信贷风险存在直接的联系，且包含的用户数以及整体数据量都比较完整，是蕴含较多信贷风险相关因素的表。用户历史信用卡账单记录如表 2-9 所示。

表 2-9 信用卡历史账单记录表

用户标识	时间	上期账单金额	上期还款金额	…	信用卡额度	还款状态
3150	68365	18.626118	18.661937	…	20.664418	0
3150	68365	18.905766	18.909954	…	20.664418	0
3150	68365	19.113305	19.150290	…	20.664418	0
3150	68365	19.300194	19.300280	…	21.000890	0

2.6 预测模型评价标准

对预测模型的评价，是验证不同模型之间效果的对比标准[54]。在分类任务中，最常用的模型评价性能度量指标有准确率，AUC 等。

（1） 错误率和准确率

准确率与错误率是相对应的，准确率是指在模型预测结果中预测标签与样本真实标签一致的样本数量的比率，公式如下：

错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}II(f(x_i)\neq y_i) \tag{2-12}$$

准确率：

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}II(f(x_i)=y_i)=1-E(f;D) \tag{2-13}$$

其中，m是样本的数量。

（2）查准率和查全率

准确率和错误率并不能满足所有任务的需求，比如在判断是否为好瓜时，我们关心有多少比例的好瓜被挑选出来，挑出的好瓜当中有多少是真实的好瓜。本文的分类

任务为二分类，评价模型分类结果的混淆矩阵如表 2-10

表 2-10 分类评价结果的混淆矩阵

真实	预测	
	正例	反例
正例	TP	FN
反例	FP	TN

查准率定义：

$$P=\frac{TP}{TP+FP} \tag{2-14}$$

查全率定义：

$$R=\frac{TP}{TP+FN} \tag{2-15}$$

（3）P-R 曲线和 F1

P-R 曲线是描述查准率和查全率变化的曲线，P-R 曲线保存训练好的模型对每个样本的预测概率，并按照概率大小排序，逐一将样本作为正例处理，此时计算出查准率和查全率，以查准率为纵轴、查全率为横轴绘制 P-R 曲线。

查准率和查全率在一定程度上来说是互斥的，根据具体问题会侧重于一方，有时需在查准率和查全率间做一个平衡，即为 F-Measure，其表达式如下：

$$F_\beta=\frac{(1+\beta^2)*P*R}{(\beta^2*P)+R} \tag{2-16}$$

特别地，当β为 1 时，就是常用的$F1$度量，公式如下：

$$F1=\frac{2*P*R}{P+R}=\frac{2*TP}{样例总数+TP-TN} \tag{2-17}$$

（4）ROC 与 AUC

ROC 曲线与 P-R 曲线原理及绘制过程相似，唯一不同的是 ROC 曲线逐一计算的是“真正例率”和“假正例率”，ROC 主要侧重于研究样本判断概率排序的优劣[55]。

TPR 和 FPR 以及 AUC 的定义如下：

$$\begin{aligned} TPR&=\frac{TP}{TP+FN}\\ FPR&=\frac{FP}{TN+FP} \end{aligned} \tag{2-18}$$

$$AUC=\frac{1}{2}\sum_{I=1}^{m-1}(x_{i+1}-x_i)*(y_i+y_{i+1}) \tag{2-19}$$

其中，x_i 和y_i 是一组 FPR 和 TPR。AUC 值越大，预测概率排序的质量越好，模型的效果越好。

2.7 本章小结

本章主要是相关知识的介绍。首先是对集成学习思想的概述，分析对比了它们的不同点。其次是机器学习算法的一些理论研究介绍，主要包含了集成学习算法、深度神经网络以及模型融合的理论介绍。再次给出本文使用的数据集，最后给出了信贷风险预测模型的各项评价指标。

第 3 章 基于多模型融合的信贷风险预测模型

3.1 数据探索及数据预处理

3.1.1 数据探索

数据探索 EDA（Exploratory Data Analysis）是指对原始数据通过统计、作图、计算相关性等手段初步探索数据存在形式和内在关系的分析方法，是数据挖掘过程中的重要一环[56]。数据探索主要价值在于快速熟悉和把握数据集，了解变量之间的相关关系以及变量与预测值之间的关系等，为数据挖掘后续过程的数据结构和特征集做准备[56]。

（1）用户基本信息表

用户基本信息表中包括职业、教育程度等五个类型的数据，对于每个属性的每一个离散值逾期的比例如图 3-1 所示。

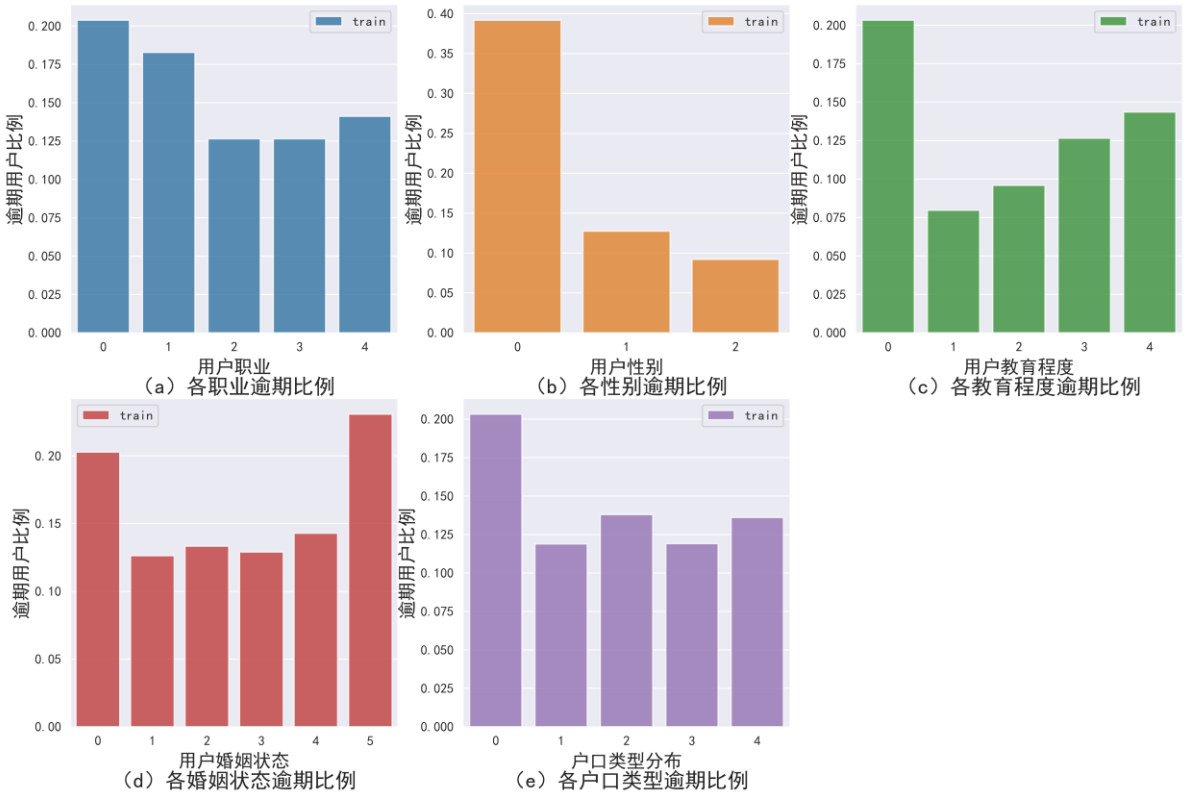


图 3-1 每个属性的离散值逾期比例

图 3-1 的柱状图描述的是用户基本属性的属性值与是否逾期的相关比例关系，以用户性别和教育程度为例，用户性别为 0 的逾期占比高达近 40%，远高于性别为 1 和 2 的用户，说明用户性别为 0 的逾期信贷风险可能较大，与预测标签的相关性较大。教育程度值为 0 的用户中逾期比例超过 20%，与其他类别相差较大，可能存在较大的信

贷风险。其他的属性如用户职业等，其各属性值逾期的比例相差不大。

（2） 银行卡流水记录表

银行卡流水记录数据表主要包括流水时间、交易金额等。对数据表做放款前后行为的对比，放款前/后用户收入和支出笔数、以及放款前/后用户的收入占比分布如图 3-2 所示。

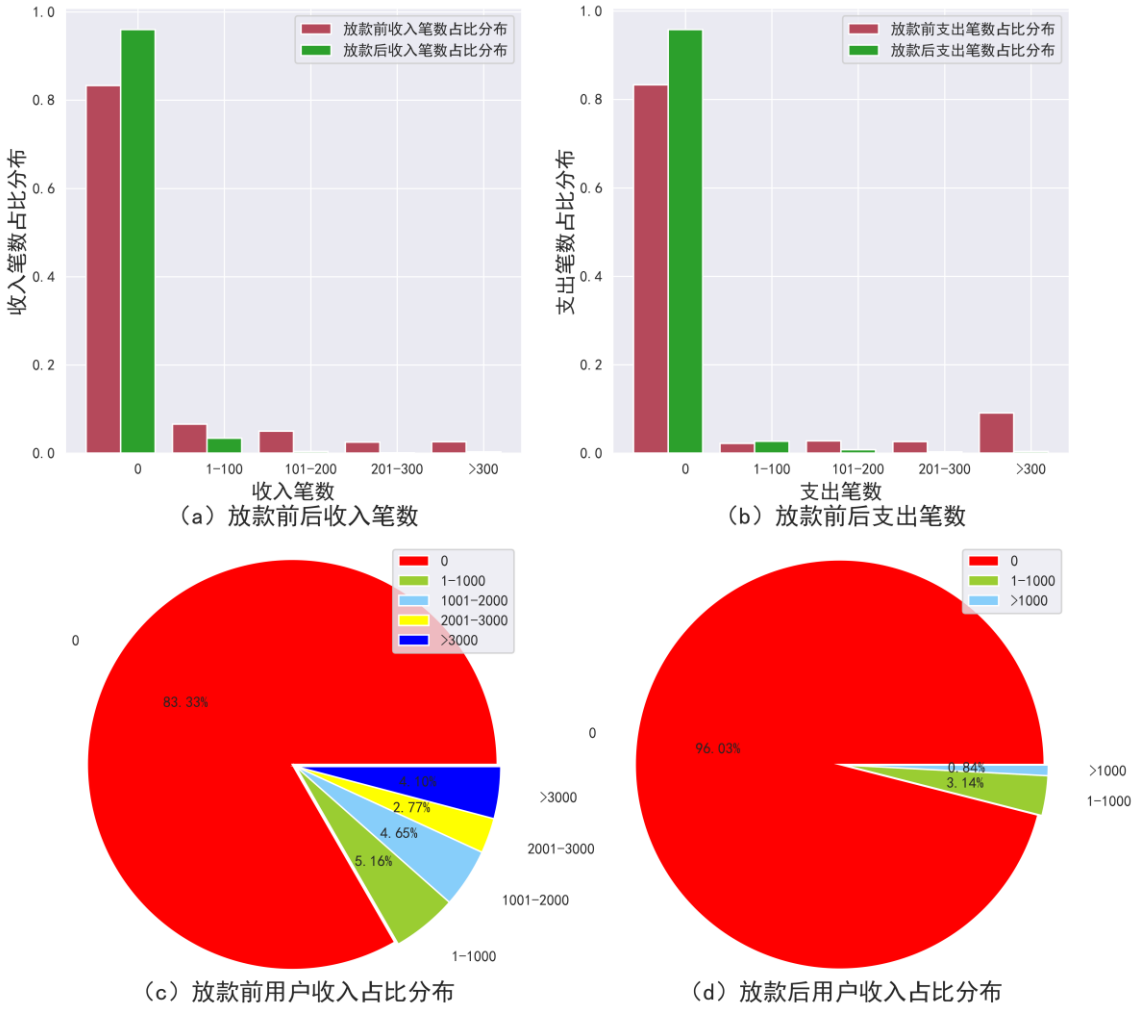


图 3-2 用户银行卡流水记录探索

图 3-2 描述的是放款前后用户收入支出笔数以及收入占比分布的对比情况。从图 3-2 中的柱状图可以明显看出放款前后收入和支出的笔数占比分布发生较为明显的变化。放款前后收入笔数为 0 的相差较大，放款后收入笔数为 0 的用户比例明显高于放款前。放款前用户的支出笔数为 0 的占比约 83%，放款后约为 96%，可能由于放款后需偿还其他账单，故整体用户的消费能力降低，其余支出笔数区间放款前后基本持平。从饼图可以看出，放款前后收入为 0 的占比达 83%以上，对于放款后达到了 96%以上，占据主体地位，说明大部分用户在放款后的收入降低。综上，用户的收入以及支出在放款前后差别较大，因此在做特征工程时可以根据放款时间分为放款前、放款后分别做特征的提取。

（3） 用户浏览行为表

用户浏览行为表包括浏览时间、浏览行为编号等。用户在还款期间若有较为频繁的浏览行为，其可能存在一定的信贷风险，用户的风险往往存在于行为之中，挖掘用户的浏览行为具有较大的价值。通过放款时间将用户的浏览行为分为放款前和放款后，观察放款前后用户行为的变化。图 3-3 描述的是放款前后用户的浏览行为变化图，两幅图中的横轴均为用户浏览的时间，前者属于放款前用户的浏览时间，后者属于放款后的用户浏览时间，纵轴为放款前和放款后用户浏览行为的总和。因浏览行为数据为脱敏数据，故统计每个用户放款前后浏览行为的总和，并结合浏览时间绘制成图 3-3 的散点图，由此可看出放款前后浏览行为的分布变化。从图 3-3 中的散点图对比可以看出，用户在放款前和放款后的浏览行为数据的分布发生较为明显的变化。从整体分布来看，放款前用户的浏览行为在各个时间段分布较均匀，放款后用户的浏览行为整体向时间后端聚集。

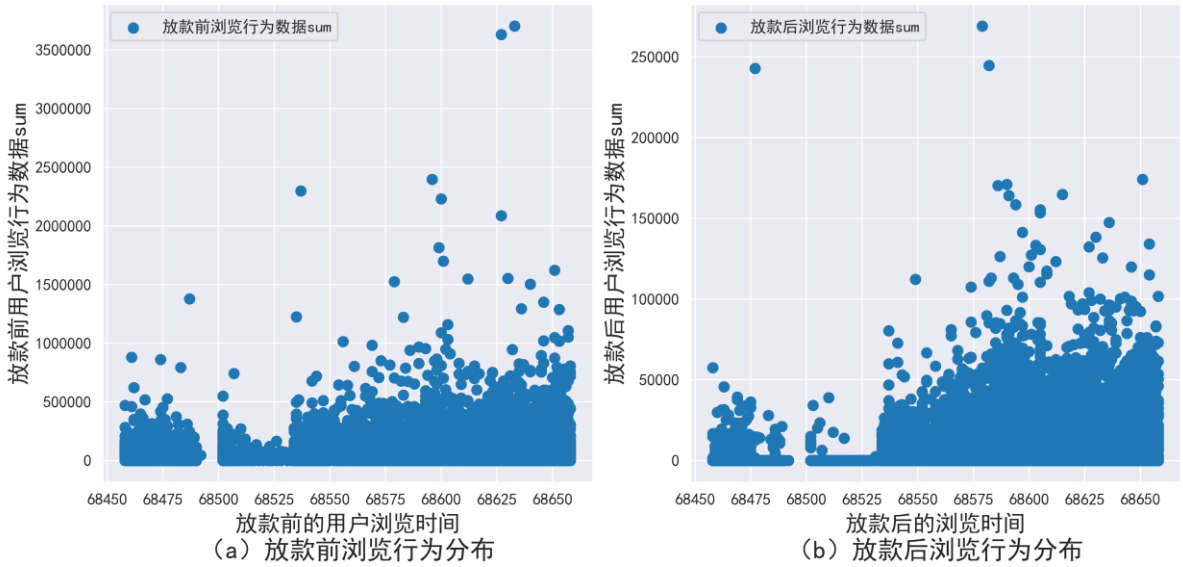


图 3-3 放款前后用户浏览行为变化图

（4） 信用卡账单记录表

用户的信用卡账单记录表包括时间、上期账单金额、上期还款金额、信用卡额度等。通过对放款前后用户的信用卡账单总笔数对比，观察用户的信用卡使用变化情况。图 3-4 是所有用户在放款前后信用卡账单总笔数的折线图。

图 3-4 描述的是放款前后用户信用卡账单总笔数的对比图。从图中可以看出，蓝色代表的放款前账单总笔数总体上要明显高于图中橙色所代表的放款后账单总笔数。当信用卡发放贷款后，用户可能使用此金额偿还其他账单，故总体上放款后的账单总笔数要小于放款前。图中存在少量放款后账单笔数远远大于放款前总笔数的用户，应注意的是此类用户可能存在较大的潜在信用风险，需重点关注。同时，放款前或放款后信用卡账单总笔数较大的用户同样可能存在较大的信贷风险。

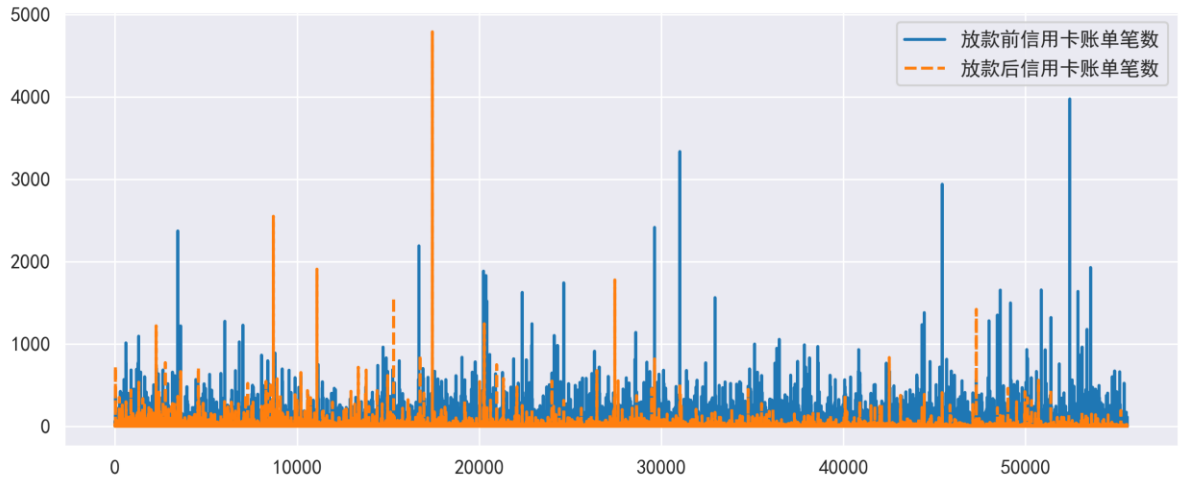


图 3-4 放款前后用户信用卡账单总笔数对比图

3.1.2 数据预处理

数据是整个建模过程中的基础，数据的质量很大程度上决定了模型最后的效果。在数据挖掘过程中，数据的缺失往往是真实存在、不可避免的。信息的遗漏或者是特征属性不存在均会导致数据的部分缺失[57]。

本文中，绝大部分用户都不会同时存在用户基本信息、银行卡流水、信用卡账单和浏览行为的记录。用户基本信息、信用卡账单、银行卡流水以及浏览行为数据中用户缺失的比例如图 3-5 所示。

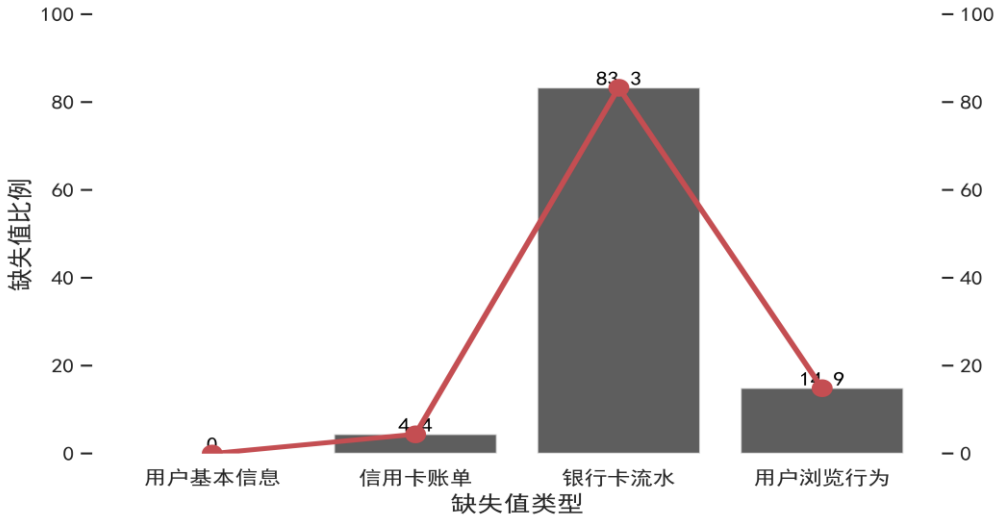


图 3-5 各原始表的用户缺失比例

从图 3-5 可以看出，用户基本信息表中无用户缺失，即对于所有的用户均存在相应的基本信息。用户缺失值比例最多的是银行卡流水记录，仅包含 16.7%比例的用户记录，大部分用户的银行卡流水记录缺失不存在。其次是用户浏览行为记录，存在 14.9%用户的缺失，信用卡账单较为完整，缺失的用户占比仅为 4.4%。

对于数据集合中缺失值的处理，主要有三种方式：分别是直接删除[58]、缺失值填充和不处理[59]。缺失值填充方式是指对缺失值按照一定规律或方法进行填充，主要的方法如表 3-1 所示。

表 3-1 常用的缺失值填补方法

插补方法	方法描述
平均值	使用该特征下所有存在值的平均值填充
固定值	使用某一固定值（如 0 或-1 等）填充
样本临近插补	使用缺失样本附近值填充
回归方法	使用回归模型预测
插值法	使用插值函数填充

在本文的研究中，由于缺失的用户样本数据较多，不可直接做删除处理。若采用直接删除的方式，则仅包含不到 10%的用户数据，会损失大量的数据信息。模型在学习过程中可用信息较少，可能存在过拟合的风险。

本文根据数据字段的具体定义和实际情况采用了零值、该列数据最小值对数据中缺失值进行填充，需要填充的表主要为银行流水记录、信用卡账单记录以及用户浏览行为表。在银行卡流水记录中使用零值对用户的交易金额进行填充；在信用卡账单中，使用零值对消费笔数、调整金额、循环利息填充，使用该列数据的最小值对信用卡额度、预借现金额度填充；在用户浏览行为中，使用零值对浏览行为数据和浏览子行为编号进行填充。具体填充方式如表 3-2 所示。

表 3-2 各特征缺失值填补方法

填充方法	特征
零值填充	上期账单金额、上期还款金额、本期账单余额、本期账单最低还款额、消费笔数、本期账单金额、循环利息、交易金额、浏览行为数据、浏览子行为编号、调整金额
该列数据最小值	信用卡额度、预借现金额度可用余额

3.2 信贷风险特征工程

特征工程是利用相关领域知识，从原始提供的数据集合中获取可用于模型训练的特征集合数据。这些特征数据需要较为完整地包含整个数据集合的信息，并且利用这些特征数据可以使得模型具有很好的泛化能力、性能达到最优[60]。在数据挖掘的整个流程中，特征工程的工作量可占据一半以上。最终模型的成功取决于数据和算法，而其中数据在算法中的表现就是特征，即特征工程结果的输出。因此，特征工程质量的优劣很大程度上决定了模型最终的预测效果和泛化能力。

3.2.1 基础特征构建

（1）类别特征--用户基本信息表

用户基本信息表中的数据类型均为类别型变量，是离散和无序的，考虑量纲等影响，对这些属性采取 One-Hot 独热编码。对属性做独热编码操作，一是为了消除离散、无序属性在模型计算距离时的影响，二是可以增加特征的维数，给予模型更多的信息。在独热编码后，对特征进行组合构造，进一步增加特征的数量。本文使用 PolynomialFeatures 进行特征的构造，其构造的方式就是特征与特征相乘，即多项式的方法。采取独热编码和特征构造可能产生特征稀疏的问题，可通过后续降维处理。

（2）非类别特征

信用卡账单表、银行卡流水记录表、用户浏览行为表中的数据均为脱敏数据，因此主要从统计特征出发做特征的提取。具体地，根据用户的行为时间和放款时间，将用户的所有行为记录分为放款前、放款后以及不区分放款前后这三种情况。对信用卡、银行卡账单、浏览行为的统计特征主要包括求和 sum，计数 count，方差 var 等。由类别特征及非类别特征组成初始基础特征共计 1636 维。

具体地，图 3-6 所展示的为部分类别特征和非类别特征信息。在非类别特征中，分为基本统计特征，构造差值、求和特征以及补充特征等。

图 3-6 类别及非类别特征示意图

（3）数据标准化

在特征工程中，数据的标准化是极为重要的环节。数据的标准化是将数据按比例

缩放，使之标准化到一个小的特定区间。标准化可以消除量纲的影响，优化算法的收敛速度和精度等。当算法中涉及距离计算时必须进行标准化处理，如 SVM、K 最近邻等。本文对特征数据使用标准化的方法为标准差标准化，μ 为样本均值，σ 为样本标准差。

$$x' = \frac{x - \mu}{\sigma} \tag{3-1}$$

3.2.2 基于序列浮动双向搜索算法的特征选择方法

在特征工程中，通过上述的特征构建会形成部分无用、冗余、重要性低的特征，这类特征对于模型训练无用甚至会影响模型的训练效果。特征选择就是从已提取的特征集合中选取最能代表数据集合信息、对模型训练有效的特征子集[61]。特征选择的要求，首先要求选取的特征子集对于学习器应具有很好的可分性，即该特征子集能够使得模型对于不同类别具有很好的区分度。其次，特征应具有可靠性，仅保留可靠真实的特征。再次，选取的特征之间应尽可能排除冗余保持独立，对于相关性较强的多个特征尽量只保留一个。最后要求所选择的特征数量应尽可能少，同时损失的信息量尽量小[62]。

机器学习中常用的特征选择方法有三种，分别为过滤法[63]、包裹法[64]以及嵌入法。Embedded 嵌入法主要包含两种，一种是利用在模型中加入惩罚项做特征选择，如在损失函数中加入 L1 正则化项等。另一种嵌入法是基于树模型的生成过程来做特征选择。如 XGBoost 根据特征分裂的次数 weight、特征平均增益值 gain 和特征平均覆盖率 cover 作为衡量特征的重要性。上述三种指标定义如下：

1）weight:权重形式，表示在所有树中，某个特征在树构建过程中被当作分裂节点的次数。

2）gain:平均增益形式，表示在所有树中，一个特征作为分裂节点存在时，带来的增益的平均值。

3）cover:平均覆盖度，表示在所有树中，一个特征作为分裂节点存在时，覆盖的样本数量的平均值。

当获取到特征的重要性之后，对全部特征根据重要性度量进行排序，可以得到原始的特征候选子集。从候选子集中搜寻最优的特征子集通常有三种策略：分别是前向搜索、后向搜索和双向搜索[65]，每次选择或舍弃一个可以使得模型性能提升的特征。然而上述的三种策略均是采取一种重要性度量方式且基于贪心策略，故最终选择的结果容易陷入局部最优，影响最终模型的效果。

（1）序列浮动双向搜索算法 SFBSA 算法的建立

本文基于上述问题建立了一种序列浮动双向搜索算法（Sequence Floating Bidirectional Search Algorithm，SFBSA）。在该算法搜索特征子集时采用两种不同的重要性度量方式 i_1 和 i_2，避免了被单一重要性度量所约束，在一定程度上减少了问题的局

限性。算法在每次搜索时，先从原始的候选特征集合中按照重要性度量方式i_1从大到小的顺序选取一个合适的特征加入目标集合，使得学习器性能提升，再从目标集合中按照重要性度量方式i_2从小到大的顺序删除能够使得学习器效果提升的特征，如此迭代进行。

具体地，SFBSA 算法的步骤如下：

Step1:使用 XGBoost 算法从两个重要性度量方式i_1和i_2计算出每个特征的重要性，并剔除特征重要性为 0 的特征。

Step2:前向特征添加。初始时先建立一个空的目标特征集合，每次从原始的候选子集中依据重要性度量i_1从大到小选择一个最重要的特征添加到目标集合中，使得该特征添加到目标集合中学习器的准确性提高。

Step3:后向特征删除。从第一步的目标集合中依据重要性度量i_2从小到大搜索并删除一个特征，使得去除该特征后学习器的准确性提高，一直删除直到不存在使学习器准确性提高的特征，返回至 Step2。

SFBSA 算法的伪代码如表 3-3 所示。

表 3-3 SFBSA 伪代码

算法：**SFBSA**
输入：包含两方面特征重要性度量方式i_1和i_2的特征集合I_1和I_2 输出：目标特征集合O 算法过程： 1: $O \leftarrow \varnothing$ 2: 两个候选特征子集I_1和I_2。I_1根据重要性度量i_1从大到小排序， I_2根据重要性度量i_2从小到大排序。 3: 从集合I_1中选取特征x_b，使得加入该特征后学习器评价指标上升。 4: ***while*** x_b存在 *do* 5: $O \leftarrow O \cup x_b$ 6: 从I_2中获取特征x_w，使得O删除x_w后学习器评价指标上升 7: ***while*** x_w存在 *do* 8: $O \leftarrow O - x_w$ 9: ***end while*** 10: 重复从I_1中获取特征x_b 11: ***end while*** 12: ***return O***

（2）SFBSA 搜索过程及结果

本文中，通过 XGBoost 模型的训练得到两种重要性度量方式，选取特征平均增益值 gain 和特征分裂的次数 weight 作为重要性度量i_1和i_2。经过 XGBoost 模型的训练，剔除重要性为 0 的特征。本文中特征平均增益（作为i_1）排名前十名和特征分裂次数（作为i_2）排名后十名的特征如表 3-4 和表 3-5 所示。

表 3-4 为特征按平均增益值排序前十名的特征，括号中数字即代表即特征重要性，可以看到放款后本期账单余额 max 的重要性最大，表中的数字特征是用户基本信息通过独热编码和特征构造形成。在特征重要性前十名的特征中用户信用卡账单特征居多，且其特征重要性均较大，后续将按照表 3-4 的顺序依次加入使得学习器性能提升的特征。

表 3-4 特征平均增益值排序前十名

排序	重要性度量 1(gain)
1	放款后本期账单余额 max(2618.84)
2	去重后放款后本期账单余额 max(2473.21)
3	去重后放款后信用卡额度 max(1473.17)
4	放款后本期账单最低还款额 max(937.77)
5	放款后信用卡额度 max(760.17)
6	去重后放款后本期账单最低还款额 max(339.20)
7	60(336.00)
8	3(325.46)
9	21(290.97)
10	13(283.79)

表 3-5 为按分裂次数排名后十名的特征（已删除特征重要性为 0 的特征），重要性即分裂次数均为 1，后续将按照此顺序依次删除可以使得学习器性能提升的特征。

表 3-5 特征分裂的次数排序后十名

排序	重要性度量 2(weight)
1	37(1)
2	放款前浏览子行为编号_11_sum(1)
3	时间未知上期账单金额 std(1)
4	放款后上期账单金额 std 与放款前上期账单金额 std 差值(1)
5	放款前预借现金额度 std(1)
6	去重后放款后可用余额 sum(1)
7	放款前信用卡额度 std(1)
8	19(1)
9	放款前上期还款金额 std 与放款前上期账单金额 std 差值(1)
10	71(1)
	去重后放款后循环利息 count(1)

SFBSA 按照特征平均增益值为前向特征添加顺序、特征分裂次数为后向特征删除顺序，得到共计 198 维显著性特征组成特征子集用于模型的后续训练，最终得到的特征子集中前 20 维特征的特征重要性（平均增益值 gain）如图 3-7 所示。

图 3-7 的纵轴为特征名称，横轴为特征平均增益，特征平均增益值前十名的特征全部都保留，特征重要性最大的仍为放款后本期账单余额 max。排名前五名的特征的重要性占据较大的比例，排名 6 到 20 名的特征重要性相差不大。最终筛选出的显著性特

征主要为信用卡账单的特征、用户基本信息特征以及浏览行为特征。

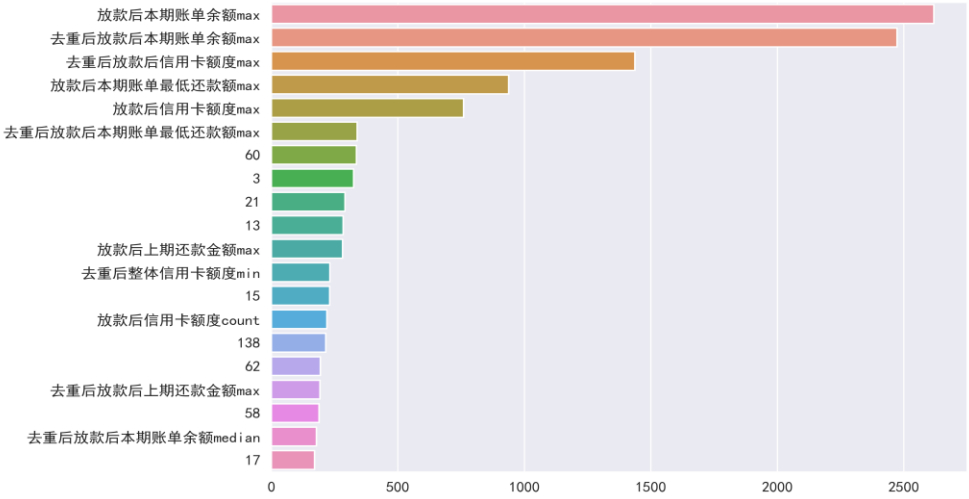


图 3-7 SFBSA 特征选择结果前 20 名特征

按照 SFBSA、前向搜索、后向搜索、双向搜索这四种方式选择出的特征子集在 GBDT、LightGBM 上的效果对比如表 3-6 所示。

表 3-6 不同特征搜索方式的准确率效果对比

搜索方式	GBDT	LightGBM
前向搜索	90.18%	90.20%
后向搜索	91.30%	92.08%
双向搜索	91.28%	91.35%
SFBSA	92.21%	92.30%

如表 3-6 所示，使用 SFBSA 做显著性特征选择在 GBDT 和 LightGBM 上的预测准确率要显著优于其他三种方法。总体来看，依据前向搜索的效果最差，平均比 SFBSA 低 2%左右，也证明了 SFBSA 做特征选择的有效性。综上，最终通过 SFBSA 算法做特征选择筛选出 198 维显著性特征用作后续模型的训练。

3.3 基于深度神经网络和集成学习算法的仿真实验及对比分析

3.3.1 单模型仿真实验环境

关于基于深度神经网络和集成学习算法的信贷风险预测研究仿真实验所使用的硬件和软件配置环境如表 3-7 和表 3-8 所示。

表 3-7 仿真实验硬件配置环境

硬件	详细信息
CPU	Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz
RAM	16G
硬盘	512GSSD

表 3-8 仿真实验软件配置环境

软件	详细信息
操作系统	Win10 64 位
实验平台	Python 3.6.1 和 Anaconda
Scikit-learn	0.23.2
Keras	2.2.4

3.3.2 单模型仿真实验数据及参数设置

（1）实验数据准备

在本文的实验中共包含 55596 个用户的行为数据。由上述特征工程得出的用户有效特征 198 维，构成 55596*198 维的特征向量矩阵。在实验中，通常采用测试误差来近似模型的泛化误差。在本文的实验过程中，使用留出法（hold-out）做模型的评估。留出法的思想是直接将数据集 D 划分为两个互斥的部分，其中一部分作为训练集 S，另一部分用作测试集 T。在划分训练集和测试集时应尽可能保持数据分布的一致性，避免因数据划分过程引入额外的偏差而对最终结果产生影响。在本文使用留出法时，为了避免单次使用留出法的不稳定性，采用了十次随机划分的方式。

（2）各模型参数设置

为了优化模型的泛化能力，需要对模型的参数进行调整。各模型参数不同，相同参数在不同算法中的作用也不尽相同。在本文的研究中，包括深度神经网络 DNN、集成学习算法的参数设置如表 3-9 所示。

表 3-9 仿真实验各单模型参数取值

模型名称	超参数	参数说明	参数值
DNN	—	隐层 1 和 2 神经元个数	550、860
	—	隐层激活函数	Relu
	—	输出层激活函数	Sigmoid
XGBoost	learning_rate	学习率	0.1
	n_estimators	个体学习器个数	120
	max_depth	最大深度	12
GBDT	learning_rate	学习率	0.2
	n_estimators	个体学习器个数	220
	max_depth	最大深度	15
	subsample	子样本比例	0.9
LightGBM	learning_rate	学习率	0.2
	n_estimators	个体学习器个数	150
	max_depth	最大深度	9
RF	n_estimators	个体学习器个数	100
	max_depth	最大深度	15

3.3.3 单模型仿真实验结果对比分析

在本文的研究中，单模型主要基于深度神经网络 DNN 和集成学习算法建立模型。DNN 通过搭建多层网络结构提升模型的拟合和泛化能力。集成学习算法比单个学习器具有更优的学习性能，主要包括两种，分别是序列化集成学习和并行化集成学习方法。本文采用了四种集成学习方法，包括 GBDT、XGBoost、LightGBM、Random Forest。本文模型评价的指标包括准确率、查准率、查全率、ROC 曲线、AUC 值以及 P-R 曲线。

（1）ROC 曲线以及 AUC 值对比

ROC 曲线是衡量模型在预测结果上排序质量好坏的标准，其体现了综合考虑学习器在不同任务下的“期望泛化性能”的好坏，或者是“一般情况下”泛化性能的好坏。若一个学习器的 ROC 曲线被另一个学习器的曲线完全覆盖，则可断言后者的性能优于前者。AUC 值为 ROC 曲线下的面积，AUC 值越大，模型的效果越好。五种模型的 ROC 曲线及 AUC 值如图 3-8 所示。

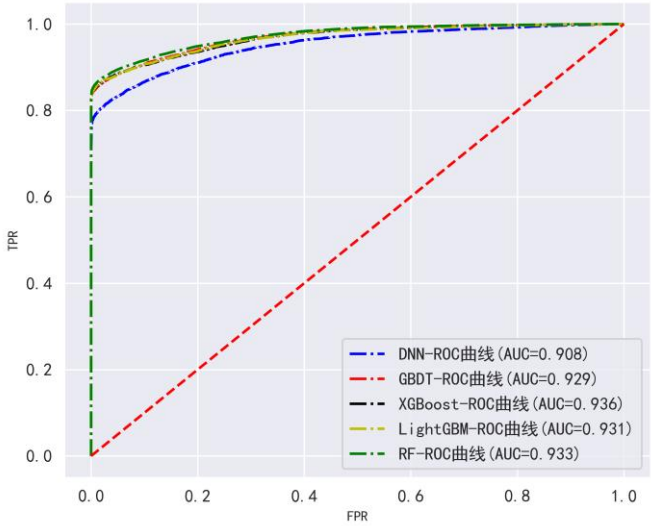


图 3-8 各单模型 ROC 曲线及 AUC 值

图 3-8 为各单模型在测试集上预测结果的 ROC 曲线及其对应的 AUC 值。从图中可以看出，五种模型的 AUC 值均超过 0.9，其中，XGBoost 的 AUC 值最大，达到 0.936，深度神经网络 DNN 的 AUC 最小，其值为 0.908。此外，五种模型中，四种集成学习算法的 AUC 值均相差不大，预测效果相近。综上，在 AUC 维度，集成学习算法的效果优于深度神经网络 DNN，其预测的概率结果排序更优。

（2）P-R 曲线

P-R 曲线直观地显示出学习器在样本总体上的查全率和查准率，同样地，若一个学习器的 P-R 曲线被另一个学习器的曲线完全“包住”，则可确定后者的性能优于前者。此外，也可比较 P-R 曲线下的面积大小，它在一定程度上表征了学习器在查准率和查全率上取得双高的比例。在本文中，各单模型的 P-R 曲线如图 3-9 所示。图 3-9 为五种

单模型在测试集上预测结果的 P-R 曲线。由图中可以看出，四种集成学习算法的 P-R 曲线将深度神经网络 DNN 的曲线完全“包住”，证明在该评价维度下，前者的性能优于后者。四种集成学习算法的 P-R 曲线比较接近，效果同 AUC 类似。此外，图中红色虚线为随机猜测的 ROC 曲线，其 AUC 值为 0.5，该红色虚线与 P-R 曲线的交点为平衡点，它是查准率和查全率相等时的取值，从图中可以看出，在平衡点评价维度下随机森林的效果要优于其他四种模型。

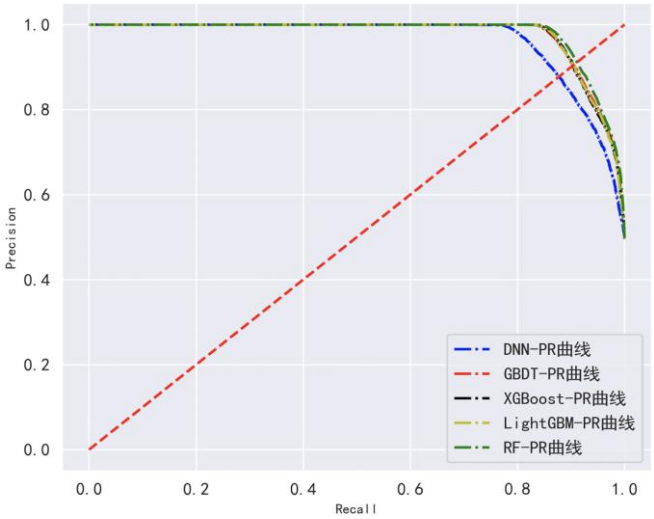


图 3-9 各单模型的 P-R 曲线

（3）准确率、查准率、查全率等指标

除了上述评价指标外，准确率、查准率、查全率也可作为重要的评价标准。本文模型为二分类模型，分为高风险用户和低风险用户。各单模型预测结果如表 3-10 所示。

表 3-10 单模型预测结果的各项评价指标

		Precision	Recall	F1-Score
DNN	低风险	0.92	0.84	0.88
	高风险	0.90	0.96	0.93
	Accuracy	—	—	0.91
XGBoost	低风险	0.99	0.87	0.93
	高风险	0.85	0.99	0.92
	Accuracy	—	—	0.92
GBDT	低风险	0.99	0.87	0.93
	高风险	0.86	0.98	0.92
	Accuracy	—	—	0.92
LightGBM	低风险	0.99	0.87	0.93
	高风险	0.85	0.99	0.92
	Accuracy	—	—	0.92
RF	低风险	0.99	0.88	0.93
	高风险	0.87	0.99	0.92
	Accuracy	—	—	0.93

表 3-10 为单模型预测结果的各项评价指标，包括整体的准确率、查准率、查全率、F1 值等。表中的 Accuracy、Precision、Recall、F1-Score 分别代表准确率、查准率、查全率、F1 值。从表中可看出，模型整体的预测准确率在 92%左右，其中，随机森林准确率最高，达到了 93%，深度神经网络 DNN 最低，为 91%。从模型对高风险用户预测的查准率来看，DNN 最高，达到了 90%，其余四种集成学习算法对高风险用户预测的查准率均较低，平均仅有 86%左右，此评价维度模型效果不理想。结合实际业务来看，信贷平台需准确识别高风险用户，降低平台的坏账率造成的损失。从模型对低风险用户识别的查准率来看，四种集成学习方法均达到了 99%，深度神经网络偏低，仅为 92%。

（4）预测结果概率分布

各模型在测试集上预测结果的概率分布如图 3-10 所示。

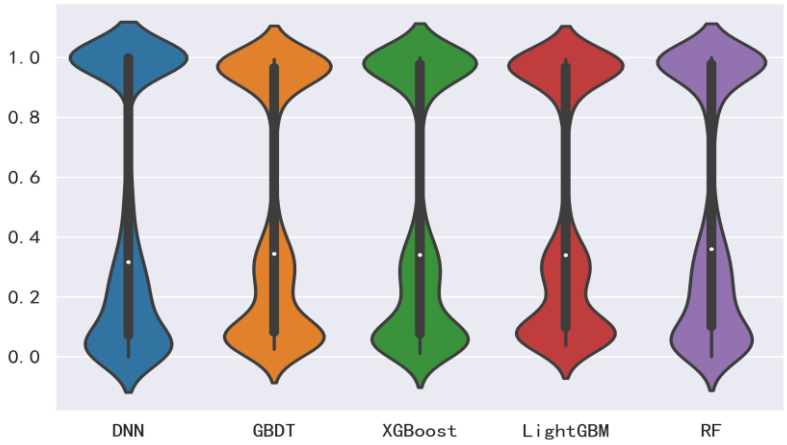


图 3-10 各单模型预测结果的概率分布

纵轴代表各个模型的概率数值分布。从图中可以看出，模型的预测概率在中间段分布较少，往 0 和 1 概率两极靠拢，说明模型可以以较高的确信度将正负样本分隔开。同时可以看出，深度神经网络对预测结果的概率分布与四种集成学习方法相差较大。对于三种 Boosting 算法（GBDT、XGBoost、LightGBM），其概率分布较为类似，与随机森林的概率分布有明显的差异。

从以上单模型结果可得，大多数单模型对低风险用户的预测准确率较好，相反，对高风险用户的预测准确率较低。而在信贷风险预测中，往往需要提高对高风险用户预测的准确率。因此，引入多模型融合技术，开展后续模型融合实验。

3.4 差异性多模型融合仿真实验及对比分析

模型融合是将多个同质或异质的学习器集成在一起做信息补充或互补，取长补短形成一个更复杂、泛化能力更强的学习器。在本文的研究中，模型融合主要采用 Stacking 算法。为了防止过拟合，次级学习器采用的是简单学习器 LR。如误差-分歧定理所述，集成学习中个体学习器之间的差异越大，模型融合的效果越好。因此为了增强学习器

之间的差异性，遵循个体学习器“好而不同”的原则，从多种模型及模型的不同参数两个方面入手，进行差异性多模型融合，建立信贷风险预测模型。

3.4.1 多种算法的模型融合

在机器学习中，不存在一个算法在所有领域均有较好的表现。因此可以融合不同算法发挥各个模型的优势达到理想的效果。本文使用 Stacking 做模型融合时，对不同的算法组合进行实验对比分析，建立泛化能力更强的信贷风险预测模型。

（1）多种算法融合实验步骤

本文对不同的算法组合做模型融合，提高模型的预测能力。多种算法的 Stacking 模型融合实验步骤如图 3-11 所示。

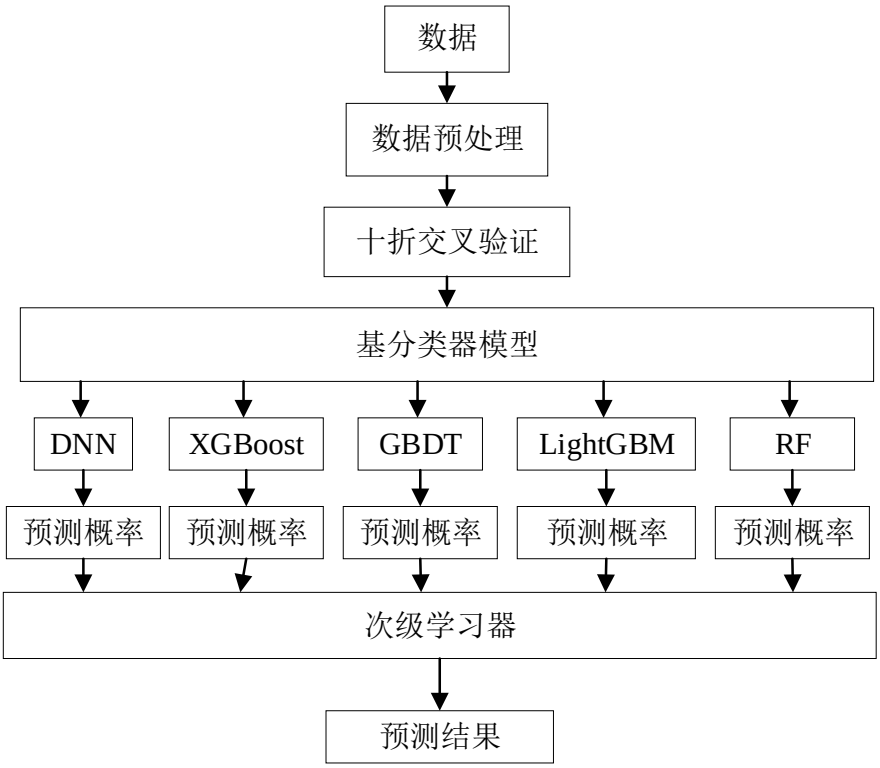


图 3-11 多种算法融合实验过程示意图

Step1:数据集划分：将整个数据集分为训练集 Training 和测试集 Testing，在 Stacking 融合过程中，使用十折交叉验证方法，将训练集 Training 再次划分为训练集 Training1 和测试集 Testing1。在每一次交叉验证过程中，将 Training 其中九折作为 Training1，剩下一折作为 Testing1;

Step2:每一折的交叉验证：在十折交叉验证过程中，每一次交叉验证包含两步：1) 基于 Training1 训练模型；2）基于 Training1 训练生成的模型对 Testing1 进行预测;

Step3:单个模型交叉验证：在第一折的交叉验证完成后，得到 Testing1 的预测值，记为 a1；以此类推，最终将得到 a2，a3……a10，将 a1 到 a10 拼接，即组成一个对

Training 的预测值矩阵，记为 A1。同时，在每一折训练完成后，对原始数据集的 Testing 进行预测，将得到当前 Testing 的预测值，记为 b1；以此类推，可得到 b2,b3……b10，对 b1 到 b10 取平均值则可得到 B1；

Step4:对融合模型组合中每个单模型重复 Step3 步骤。以融合五个模型为例，将得到 A2，A3，A4，A5 和 B2，B3，B4，B5 矩阵。将 A1-A5 矩阵并列拼接组成新的 Training Set，B1-B5 拼接成新的 Testing Set;

Step5：将新的 Training Set 和 Testing Set 输入到次级学习器进一步训练测试。在本文中，为了防止出现过拟合，选择的次级学习器为 LR。

（2）多种算法融合实验结果对比分析

通过对不同的模型组合进行融合加强差异性，提升模型的效果。本文对多种模型的组合进行试验，五种模型任意选取其中两种及以上共计 26 种组合分别进行 Stacking 模型融合仿真实验，以表 3-11 中的四种组合为例。

表 3-11 多种算法的组合示例

模型名称	模型组合
M1	DNN、GBDT、XGBoost、LightGBM、RF
M2	DNN、GBDT、XGBoost、LightGBM
M3	XGBoost、GBDT、LightGBM、RF
M4	DNN、XGBoost、RF

依据上述步骤，四种模型组合 M1、M2、M3、M4 融合的效果如图 3-12 所示。

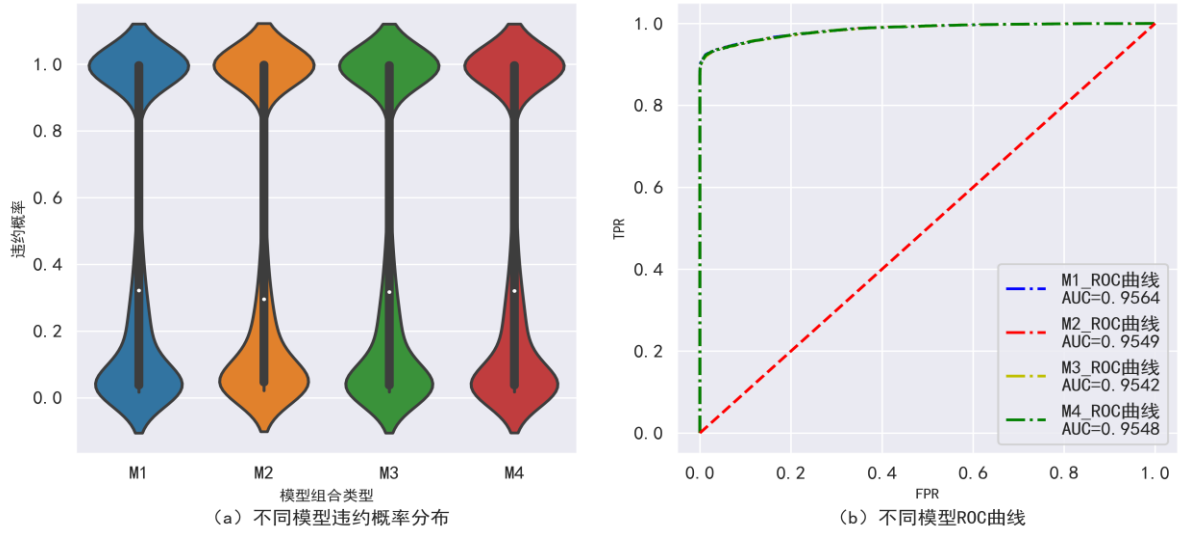


图 3-12 多种算法融合的模型评价结果

如图 3-12 所示，四种模型融合的效果相差不大，其 ROC 曲线几近重合。从四种融合模型测试集上预测结果的概率分布小提琴图来看，M1、M2、M3、M4 预测的概率分布基本相同，且基本都集中于两端概率的分布，中间段较少，分类结果较为可靠，ROC

曲线也证实了这一点。相比于未做模型融合前的单个模型，融合后的模型效果显著提升，相比于集成学习单模型，AUC 提升超过 0.02，均达到 0.95 以上。

M1、M2、M3、M4 各个融合模型的预测结果准确率等指标如表 3-12 所示。

表 3-12 多种算法的模型融合预测结果

		precision	recall	F1-score
M1	低风险	0.98	0.94	0.96
	高风险	0.93	0.97	0.95
	Accuracy	—	—	0.95
M2	低风险	0.97	0.94	0.95
	高风险	0.93	0.97	0.95
	Accuracy	—	—	0.95
M3	低风险	0.97	0.93	0.95
	高风险	0.93	0.97	0.95
	Accuracy	—	—	0.95
M4	低风险	0.97	0.93	0.95
	高风险	0.93	0.97	0.95
	Accuracy	—	—	0.95

表 3-12 显示，融合模型的准确率、召回率等方面相较于单模型有了明显的提升，M1、M2、M3、M4 在各个评价指标上效果相近，且相较于单模型，融合后的模型效果有了较大程度的提升。以 M1 为例，总体的准确率达到 95%，相比于单模型中效果最好的 RF，融合模型的精度提升超过 2%。模型对风险较高的用户预测查准率有了很大程度上的提升，由之前的 87%上升到 93%，在一定程度上解决了上述提到的问题。准确率、查准率、召回率的提升同样使得 F1 值相比于集成学习算法提升超过 2%。综上，经过多种算法的模型融合，整体的模型效果相较于集成学习单模型有了显著提升。

3.4.2 算法不同参数的模型融合

同种算法不同参数能训练出不同的模型，通过对算法的部分参数做小范围的扰动来训练生成不同的模型，加强融合个体的差异性，从而提升融合模型的最终效果。

通过算法参数的扰动提高模型融合效果的过程示意图如图 3-13 所示。从图 3-13 可以看出，对 XGBoost、GBDT、LightGBM 和 RF 的部分参数做小范围的扰动训练形成多个单模型，并使用 Stacking 进行模型融合。

算法的不同参数融合步骤如下：

Step1:对 GBDT、XGBoost 等算法参数做小范围的扰动，每个算法形成多组不同的参数。

Step2:使用对应的参数组合做单模型的训练，一个算法生成多个不同的模型。

Step3:使用 Stacking 算法对上述生成的模型做模型融合，并输出最终结果。

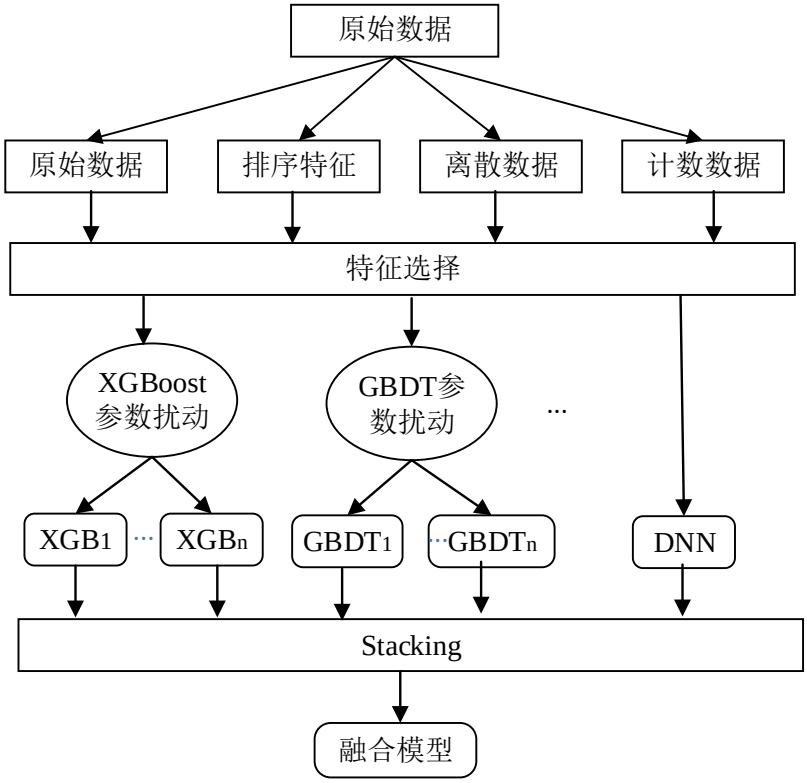


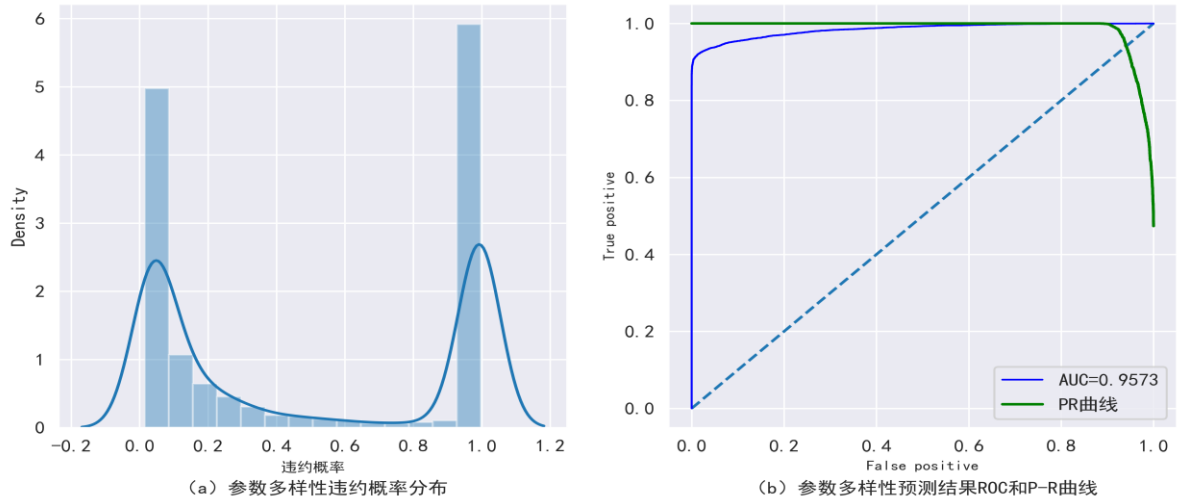
图 3-13 算法的不同参数的融合过程示意图

以 XGBoost 为例，对 XGBoost 的部分参数做小范围扰动生成多个模型，XGBoost 的参数调整如表 3-13 所示。

表 3-13 XGBoost 参数扰动值示例

模型名称	超参数	参数说明	参数值
XGBoost_1	learning_rate	学习率	0.1
	n_estimators	个体学习器个数	120
	max_depth	最大深度	12
XGBoost_2	learning_rate	学习率	0.15
	n_estimators	个体学习器个数	150
	max_depth	最大深度	15
XGBoost_3	learning_rate	学习率	0.08
	n_estimators	个体学习器个数	200
	max_depth	最大深度	10
XGBoost_4	learning_rate	学习率	0.12
	n_estimators	个体学习器个数	240
	max_depth	最大深度	8

对算法的参数做小范围扰动，最终模型的效果图如图 3-14 所示。（a）图为预测概率的分布图，横轴为用户违约的概率，纵轴为对应概率的密度大小，从图中可以看出，概率分布越来越趋于两端，尤其在[0.5，0.7]之间的分布很少，也从侧面证实了模型效果的提升。（b）图展示了模型预测结果的 P-R 曲线和 ROC 曲线，AUC 值为 0.9573，相比于单模型提升超过 2%。



（a）参数多样性违约概率分布　（b）参数多样性预测结果ROC和P-R曲线

图 3-14 算法的不同参数融合模型评价结果

算法的不同参数融合模型效果如表 3-14。表中 macro avg 为宏平均（对每个类别的精准、召回和 F1 加和求平均，如 macro avg 行对应 F1-Score 列代表的是对低风险和高风险预测 F1 值加和求平均），weighted avg 为加权平均（对宏平均的改进，考虑每个类别样本数量在总样本中比例后计算精准、召回和 F1）。

从表 3-14 模型预测结果可得，模型效果与多种算法融合模型基本持平，相较于单模型有较大的提升效果。同多种算法融合一样，显著地提升了模型整体的准确率和对高风险用户识别的查准率。

表 3-14 算法的不同参数融合模型预测结果

	Precision	Recall	F1-Score
低风险	0.98	0.94	0.96
高风险	0.93	0.97	0.95
accuracy	—	—	0.95
macro avg	0.95	0.95	0.95
weighted avg	0.95	0.95	0.95

3.4.3 时间成本分析

模型融合在整体的效果上相比于单模型如集成学习算法 XGBoost 等有了一定程度的提升，尤其对于高风险用户预测的查准率来说。但时间消耗的成本也随着个体学习器的增加而增加。表 3-15 为各单模型及融合模型的运行时间。

表 3-15 各模型运行时间表

模型	时间/秒	模型	时间/秒
XGBoost	70.22	DNN	640.88
GBDT	157.88	多种算法融合(M1)	2867.42
LightGBM	4.71	算法不同参数融合	11532.65
RF	45.58		

从表 3-15 可以看出，在单模型中，LightGBM 模型的运行速度远快于其他模型，因其使用了直方图算法和差加速等。模型融合的时间成本远大于单模型的时间，因做模型融合时采用了十折交叉验证的方法，因此消耗的时间大大增加，尤其对于算法不同参数的模型融合来说，其针对每个单模型进行了参数的扰动，进一步增加了模型的训练时间。在现实中，对算法的准确性和运行时间需做一个平衡。

3.5 本章小结

本章主要内容为介绍基于多模型融合的信贷风险预测模型研究的实验部分。第一部分是对数据的探索及预处理，基于可视化对数据的部分特征进行了分析。第二部分为特征工程的构建，主要包括基础特征和显著特征。第三部分为基于深度神经网络和集成学习算法的仿真实验对比研究。第四部分是差异性多模型融合的仿真实验，主要包含多种模型和模型的不同参数两方面。

第 4 章 基于改进的遗传算法与多模型融合的信贷风险预测模型

在基于深度神经网络、集成学习算法以及多模型融合的预测模型研究中取得了较好的预测效果。但在模型训练过程中参数组合较多，难以选择模型最佳的训练超参数，影响模型的效果。在本章研究中，首先针对标准遗传算法容易陷入局部最优以及早熟等问题，建立了改进的线性自适应遗传算法，并使用改进后的遗传算法对 GBDT、XGBoost、LightGBM 和 RF 模型进行智能调参，选取域空间内的最优训练超参数组合，并运用 Stacking 模型融合技术建立精准的信贷风险预测模型，提高整体模型的预测效果。

4.1 遗传算法原理概述

4.1.1 遗传算法思想

遗传算法（Genetic Algorithm，GA）是将进化论和遗传学说的思想模拟应用于计算机算法研究中，可作为一种优化和搜索域空间的方法。遗传算法可以高效、并行地控制搜索过程，通过交叉、变异等操作获得域空间内近似最优解[66]。

相比于其他优化算法，遗传算法对于问题本身并不要求很多限制，其可以直接对所需求解问题的对象操作，无须求导或梯度等限定，可直接用于寻找最优的超参数组合[67]。

在使用遗传算法搜索域空间时，需要求解问题的每一个初始可能解均被某种编码方式编码成个体，即进化论和遗传学说中的染色体。所有被编码的个体组成了初始解的种群。遗传算法在计算时，根据求解问题的评价标准选择适应度的计算方式，并确定其相应的适应度，仅保留个体适应度大于阈值的个体。再通过遗传算子等操作组合生成下一代的个体，如此迭代进行，该种方式可以保证逐渐向问题的近似最优解方向进化[68]。

4.1.2 编码及遗传算子

遗传算法求解最优值的过程中，涉及到两个重要的步骤，分别是编码和遗传算子的操作。其中，遗传算子包含选择、交叉、变异等。

（1）编码操作

在使用遗传算法求解最优化问题之前，首先需要明确在该问题中的函数形式以及

其中存在的各项变量，然后使用遗传算法中相应的编码方式对该问题中的各项变量进行编码[68]。

（2）选择操作

遗传算子中的选择操作指的是基于适应度函数和适应度值来选择保留或者淘汰种群中的个体。对于种群中的每个个体，保留大于适应度阈值的个体，其余全部舍弃[69]。

目前遗传算法中常用的选择算子包括轮赌盘方法等。以轮赌盘方法为例，构建赌盘时，需要计算每个个体的适应度占据总的所有个体适应度之和的比率，所有个体的比例总和为 1，如图 4-1 所示。

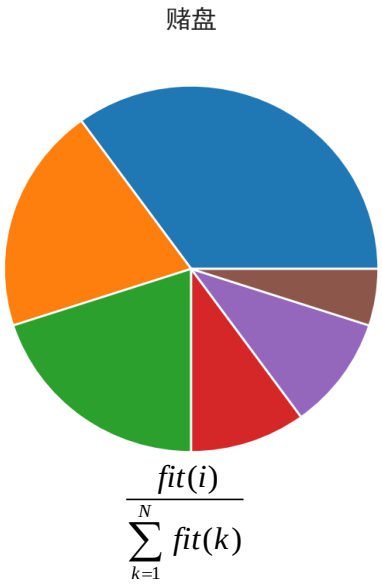


图 4-1 轮赌盘方法

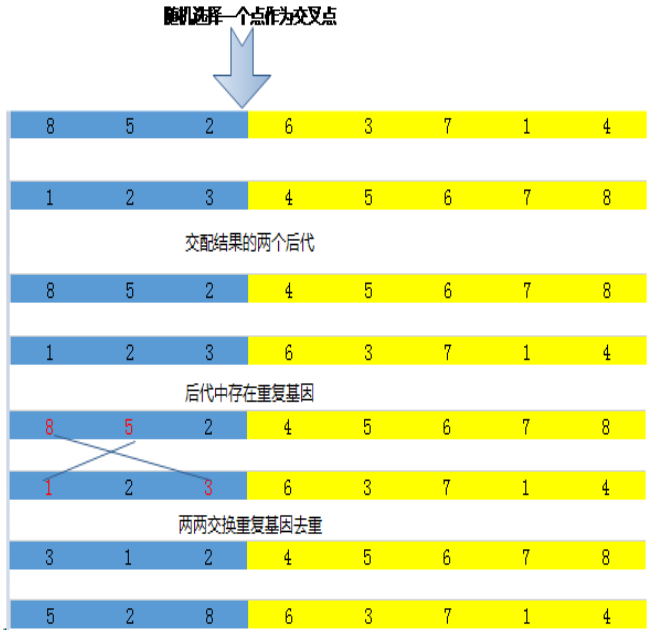


图 4-2 单点交叉示意图

（3）交叉操作

遗传算子中的交叉操作指的是将父代中的两个个体结构进行一定程度的交叉变化，以此在种群中生成新的个体。生成的新个体可以遗传到下一代。遗传算子的交叉操作，可以在遗传算法搜索过程中大大增加个体的数量，极大提升遗传算法搜索近似最优解的能力[70]。遗传算子中经常使用的单点交叉示意图如图 4-2 所示。对于两个个体，对其进行相互交叉，生成交配结果的两个后代，对于后代中存在重复基因，需两两交换去除重复基因。

（4）变异操作

遗传算子中的变异操作指的是在遗传算法运行过程中以一个很小的概率选择对个体某些基因值做一定程度的变化。遗传算法运行过程中通过交叉和变异操作相互配合使其具备全局的搜索能力[67]。变异操作还可以维持种群的多样性，防止出现未成熟收敛现象。

4.2 改进的线性自适应遗传算法 ILAGA

4.2.1 交叉率和变异率的改进

如上文阐述的遗传算法为标准遗传算法。在该遗传算法中，遗传算子均使用了不变的交叉概率以及变异概率。这对于简单的最优问题较为有效，但对于复杂问题其劣势也较为明显。其劣势容易导致较早的出现“早熟”以及收敛速度较慢等问题，最终结果容易陷入局部最优。

针对此类问题，Srinvas 提出了针对交叉概率和变异概率的自适应遗传算法。在该改进算法中，交叉概率和变异概率可适时变化，采用了线性函数进行自适应的变化调整，有效解决了早熟等问题。

在 Srinvas 的改进算法中，提出交叉率和变异率随着种群的平均适应度及种群个体中的最大适应度之间进行适时调整，交叉率和变异率的公式如 4-1 和 4-2 所示。

$$P_c=\begin{cases}\dfrac{k_1(f_{\max}-f')}{f_{\max}-f_{avg}} & f'\geq f_{avg}\\ k_2 & other\end{cases} \tag{4-1}$$

$$P_m=\begin{cases}\dfrac{k_3(f_{\max}-f)}{f_{\max}-f_{avg}} & f\geq f_{avg}\\ k_4 & other\end{cases} \tag{4-2}$$

其中，$f_{\max}$ 和 f_{avg} 分别是种群中个体的最大适应度以及所有个体的平均适应度，f 为种群中即将变异个体的适应度。f' 为在进行交叉算子操作过程中两个个体中较大的适应度。从公式 4-1 和 4-2 直观来看，对变异率和交叉率做了线性处理的调整，而不再固定。当根据适应度函数计算出个体的适应度低于平均适应度时，则说明该个体所代表的解效果较差，则根据遗传算法的思想，对其进行较大的进化，即采用较大的交叉率和变异率。若种群中的个体适应度较高，则根据公式 4-1 和 4-2 进行线性调整。交叉率和变异率曲线如图 4-3 所示。

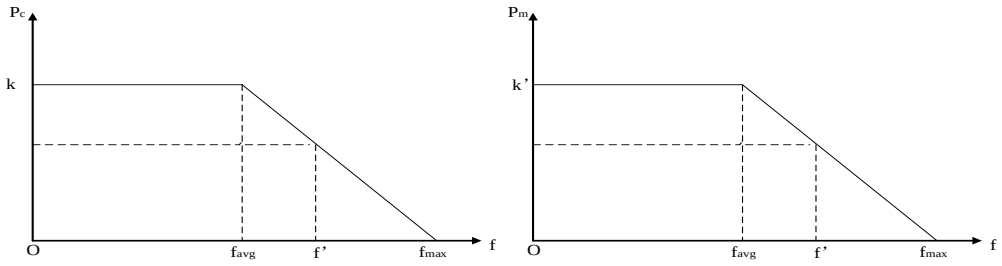


图 4-3 线性自适应遗传算法交叉率和变异率

上述对变异率和交叉率的改进可以显著提升模型寻优的能力。但存在这样一个问

题，当 f 与 $f_{\max}$ 相等时，此时根据公式 4-1 和 4-2 可知，P_m 和 P_c 均为 0，导致遗传算法计算初期时，种群中适应度高的个体只能发生较小的变化，容易陷入局部最优。因此，针对该问题，本文建立了一种改进的线性自适应遗传算法 (Improved Linear Adaptive Genetic Algorithm，ILAGA)，对遗传算子中的交叉率和变异率做进一步的优化。

对 P_c 和 P_m 做如下优化，交叉率和变异率计算公式如 4-3 和 4-4 所示。

$$P_c=\begin{cases} c1-\dfrac{(c1-c2)(f^{'}-f_{avg})}{f_{\max}-f_{avg}} & f^{'}\geq f_{avg} \\ c1 & other \end{cases} \tag{4-3}$$

$$P_m=\begin{cases} m1-\dfrac{(m1-m2)(f-f_{avg})}{f_{max}-f_{avg}} & f\geq f_{avg} \\ m1 & other \end{cases} \tag{4-4}$$

其中，$c1$、$c2$ 是交叉率的最大值和最小值，$m1$、$m2$ 为变异率的最大值和最小值。

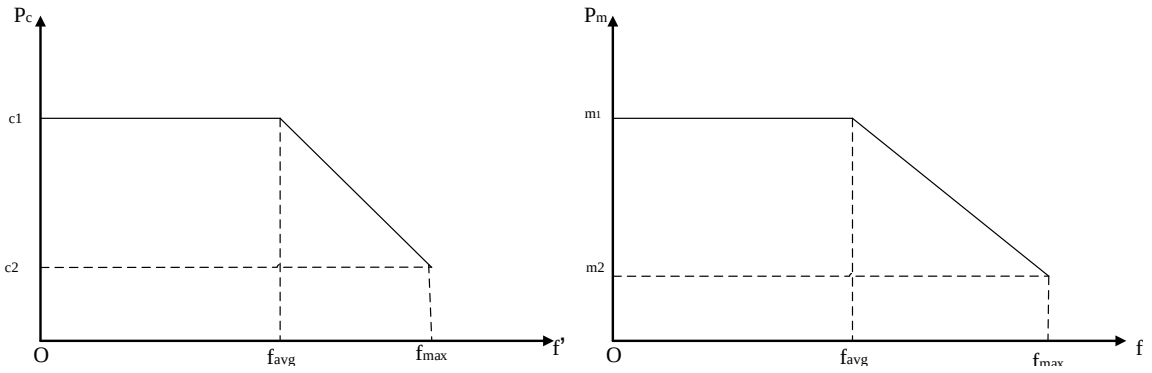


图 4-4 改进的线性自适应遗传算法交叉率和变异率

图 4-4 为改进的线性自适应遗传算法的交叉率和变异率曲线。从图可以看出，交叉率和变异率均有上下限的限制。通过上述公式的改进，在种群中，具备最大适应度的所有个体的其变异率和交叉率也不为 0，使得优质个体（即适应度高的个体）不再停止更新，进一步增加优质个体的后代，增强遗传算法寻优的能力。

4.2.2 改进的线性自适应遗传算法实现步骤

改进的线性自适应遗传算法的流程图如图 4-5 所示。

基于交叉率和变异率的改进线性自适应遗传算法详细步骤如下：

Step1：编码。确定实际问题的参数集后针对欲解决问题的变量进行某种形式的编码，编码需能够反映问题的解空间。

Step2：初始种群的生成。随机产生 N 个初始串结构，每个串结构数据称为一个个体。N 个个体，构成了一个种群。GA 以这 N 个串结构数据作为初始点开始迭代。

Step3：根据实际问题的优化目标，确定问题的目标函数和适应度函数，如在回归中可以使用 RMSE 作为目标函数，RMSE 的倒数作为计算适应度的函数。

Step4：适应度计算。将种群中的个体代入优化问题的目标函数和适应度函数，计算每个个体的适应度值。并由计算出的适应度值来评价染色体的优劣，若满足问题的优化指标或达到最大迭代次数，则输出问题的解，否则继续对染色体进行遗传操作（step5-step7），升级种群。

Step5：选择操作。将选择算子作用于种群。选择的目的是把优化的个体直接遗传到下一代或通过配对交叉产生新的个体传到下一代。选择操作是建立在种群中个体的适应度评估基础上的，如选择具有最大适应度的前 m 个个体。

Step6：交叉操作。对于第五步选出的较优个体按照一定的方式进行交叉操作产生新的个体，使得种群更加多样化。对于交叉率的设定，使用公式 4-3 的 P_c 。

Step7：变异操作。对交叉操作完成的部分染色体再进行变异操作，即对种群中个体串的某些基因值做变动，进一步扩展种群的多样性。对于变异率的设定，使用公式 4-4 的 P_m 。

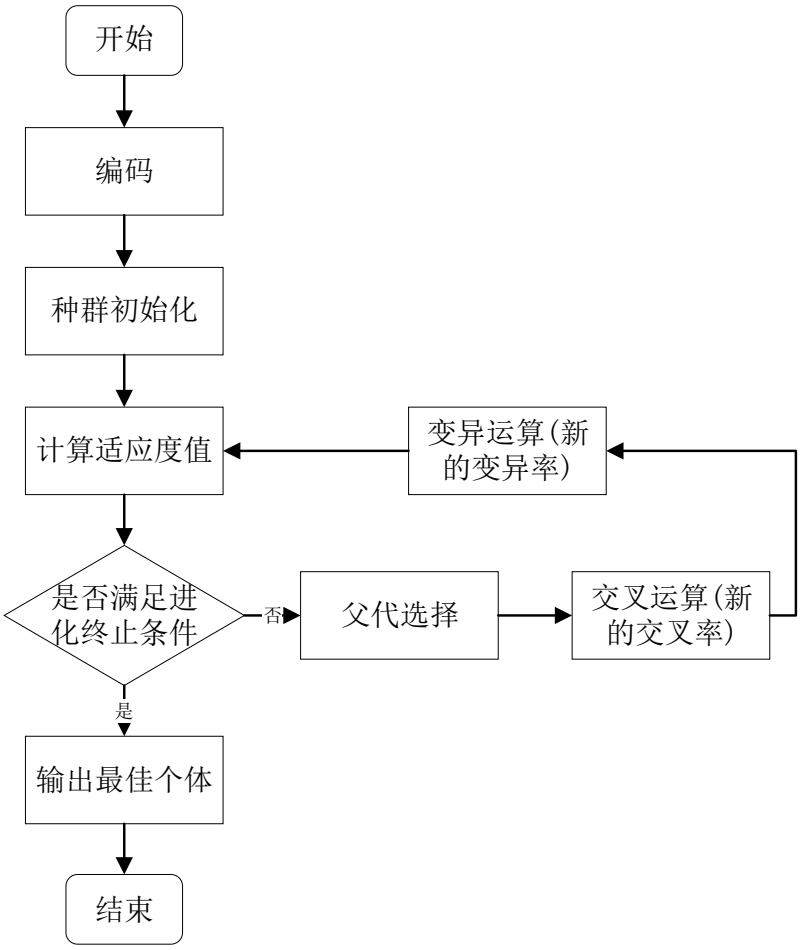


图 4-5 改进的线性自适应遗传算法流程图

4.3 单模型算法改进仿真实验及对比分析

4.3.1 基于 ILAGA 优化集成学习模型的设计与实现

遗传算法作为一种具有随机性、并行性的优化方法，其个体适应度函数能自动化确定和缩小最优参数组搜索的方向与规模。同时，遗传算子中的交叉和变异操作能有效帮助跳出局部搜索范围，避免落入局部最优的情况。最后，多个搜索信息点的同步执行，可高效确定问题的近似最优解，从而获得模型最优参数组。

本文利用上述基于交叉率和变异率的改进遗传算法对四种集成学习算法（GBDT、RF、XGBoost、LightGBM）进行调参优化，在其超参数域空间内进行寻优，寻找模型的近似最优解，解决集成学习模型在调参时出现的因低效率、慢收敛以及局部最优导致的低准确率的问题。以集成学习算法中的 LightGBM 为例，利用改进的遗传算法优化 LightGBM 的执行步骤如图 4-6 所示。

从图 4-6 可以看出利用改进的遗传算法优化 LightGBM 步骤如下：

Step1:数据集的划分。将整个数据集划分为三个部分，分别是训练集、验证集以及测试集。训练集用于寻优过程中模型的训练，验证集是模型在获得遗传算法搜索的参数后进行评价的数据集。测试集是用来测试优化后模型的效果。

Step2:遗传算法初始化。确定整个种群的大小、子代的规模、交叉率、变异率以及 LightGBM 的参数，其中交叉率 P_c 和变异率 P_m 采用优化后的计算公式 4-3 和 4-4。

Step3:模型训练。使用 LightGBM 参数和训练集进行模型训练，并在验证集上预测。

Step4:求解目标函数值。本次研究以最大化模型在验证集上的准确率为目标，因此确定目标函数为计算模型在验证集上的准确率。使用定义的目标函数求解当前种群中各个体的目标函数值，并对种群内各个体进行二进制编码，即对每个 LightGBM 待寻优参数组中的参数进行编码；

Step5:计算种群中各个体的适应度函数值。本次研究中使用的适应度函数与目标函数一致，即计算模型在验证集上预测的准确率。种群内各个体适应度计算完成后找出其中适应度函数值最大对应的参数组，同时记录其结果；

Step6:选择操作。运用轮盘赌算法进行自然选择，保留被选中的参数组，淘汰部分个体；

Step7:交叉和变异操作。对剩余的参数组按上述改进的交叉率 P_c 和变异率 P_m 进行交叉和变异操作，本文使用的是多点交叉；

Step8:更新参数组。对记录的适应度最大的二进制参数组进行解码操作，更新当前最优参数组；

Step9:判断条件。判断计算结果是否达到停止进化条件，若未达到，重复上述第（3）

步到第（8）步的操作，若达到，则输出最优参数组和验证集上预测结果的准确率。

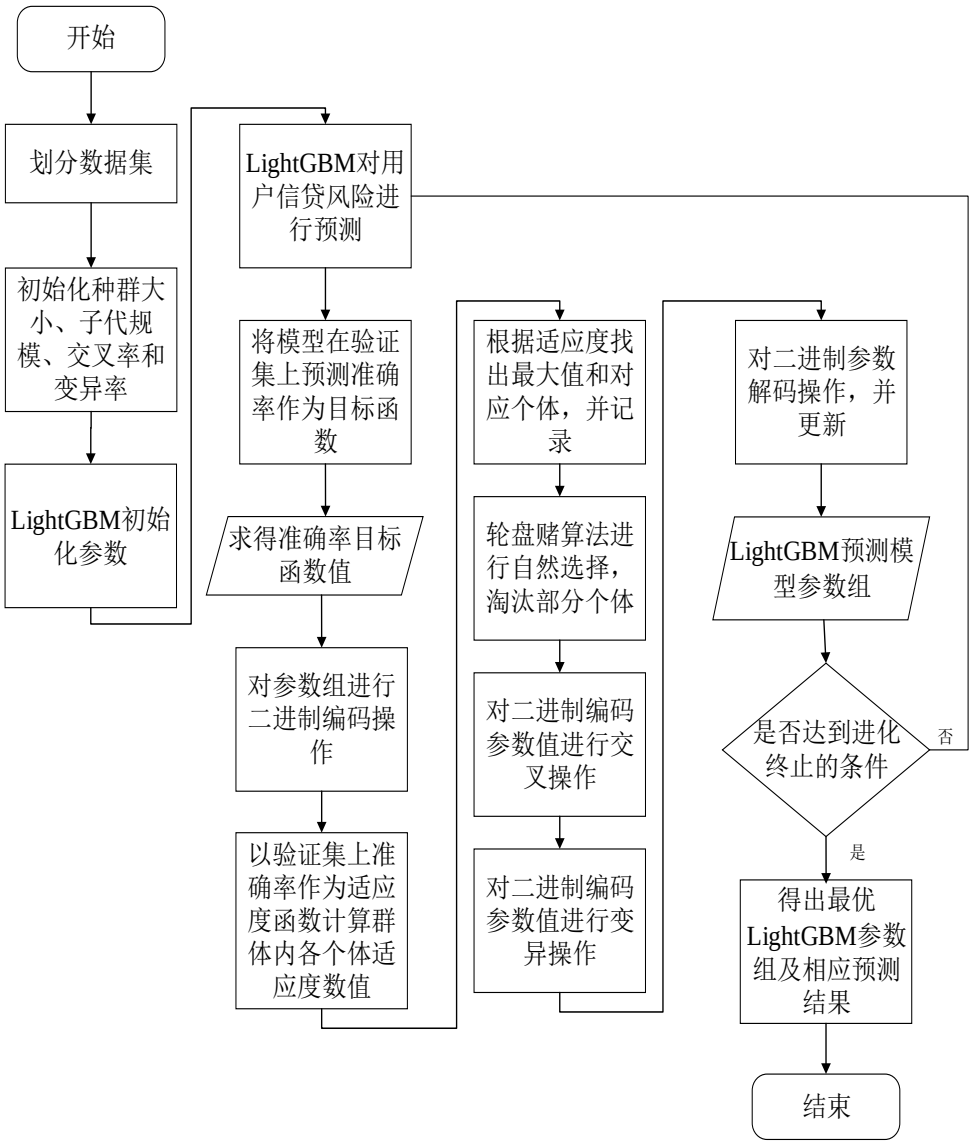


图 4-6 遗传算法改进 LightGBM 流程图

4.3.2 集成学习模型超参数域定义

同特征工程一样，模型的参数调节是非常重要的一项工作。在本文的研究中，主要利用改进后的遗传算法对 XGBoost、GBDT、LightGBM 以及 Random Forest 进行优化，为模型寻找最优超参数组合，使模型具有更好的泛化能力。

通过第二章对 GBDT 等集成学习算法的介绍以及第三章单模型调参研究发现，learning_rate，n_estimators，max_depth 等参数对模型的准确率影响较大。根据经验，模型学习率的参数范围一般设置在[0.01, 0.8]之间，学习率过大或过小都会影响模型训练的最终效果。树模型的最大深度过小或过大都会极大影响模型最后的效果。因此，确定本文研究各单模型的超参数域空间如表 4-1 所示。

表 4-1 遗传算法改进的单模型超参数域空间

模型名称	超参数	参数说明	参数范围
XGBoost	learning_rate	学习率	[0.01，0.8]
	n_estimators	个体学习器个数	[10，2000]
	max_depth	最大深度	[1，15]
	min_child_weight	子节点最小权重和	[0，10]
GBDT	learning_rate	学习率	[0.01，0.85]
	n_estimators	个体学习器个数	[10,1500]
	max_depth	最大深度	[1,15]
	subsample	子样本比例	[0.01,0.9]
LightGBM	learning_rate	学习率	[0.01,0.8]
	n_estimators	个体学习器个数	[10,1500]
	max_depth	最大深度	[1，20]
	colsample_bytree	列采样比例	[0.1, 1]
RF	n_eatimators	个体学习器个数	[8，2000]
	max_depth	最大深度	[1， 20]
	min_samples_leaf	叶子节点最少的样本数	[30, 400]

4.3.3 基于 ILAGA 优化集成学习模型的实验结果分析

通过使用改进后的线性自适应遗传算法对 GBDT 等单模型的参数域空间寻找最优超参数组合，达到优化单模型的效果，进而改善模型融合最终的效果。利用上述改进后的遗传算法及实验步骤对集成学习算法进行优化，经过仿真实验，最终得到的四组最优参数组合如表 4-2 所示。

表 4-2 改进后遗传算法改进的单模型超参数组合

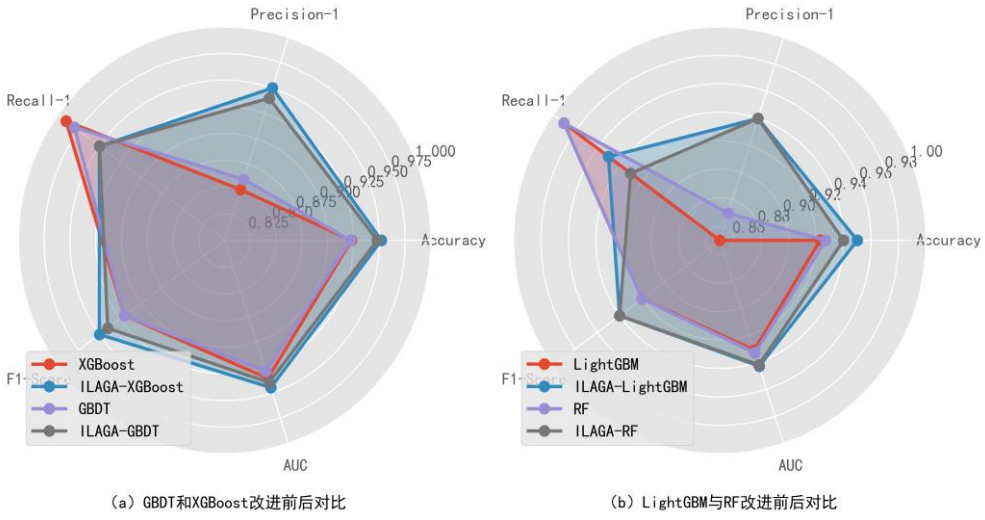
模型名称	超参数	参数说明	参数值
XGBoost	learning_rate	学习率	0.13
	n_estimators	个体学习器个数	198
	max_depth	最大深度	9
	min_child_weight	子节点最小权重和	4.75
GBDT	learning_rate	学习率	0.21
	n_estimators	个体学习器个数	296
	max_depth	最大深度	12
	subsample	子样本比例	0.83
LightGBM	learning_rate	学习率	0.19
	n_estimators	个体学习器个数	142
	max_depth	最大深度	8
	colsample_bytree	列采样比例	0.76
RF	n_estimators	个体学习器个数	106
	max_depth	最大深度	14
	min_samples_leaf	叶子节点最小样本数	78

集成学习算法改进前以及利用标准遗传算法、ILAGA 算法优化后在测试集上预测结果对比如表 4-3 所示。表 4-3 中 Precision-0 是该模型下预测低风险用户的查准率，Recall-0 是该模型下预测低风险用户的召回率，F1-score-1 为该模型下预测高风险用户结果的 F1 值，ILAGA-XGBoost 为利用改进的自适应遗传算法优化 XGBoost 模型，GA-XGBoost 为使用标准遗传算法优化 XGBoost 模型。从上表可以看出，改进后的模型识别高风险用户的查准率上有显著提升，总体准确率相比未改进前的模型提升 2%左右。值得一提的是无论是使用标准遗传算法还是改进的自适应遗传算法，对高风险用户的识别查准率达到了 93%以上，虽对于低风险用户的识别有了一些下降，但从研究意义上来说，对于高风险用户的识别具有更大的价值。此外，对利用标准遗传算法和 ILAGA 优化单模型效果作对比分析，在查准率、准确率、召回率方面后者明显优于前者。

表 4-3 遗传算法改进前后模型效果对比

模型名称	Accuracy	Precision-0	Recall-0	Precision-1	Recall-1	F1-score-1
XGBoost	0.923	0.99	0.87	0.85	0.99	0.92
ILAGA-XGBoost	0.952	0.95	0.95	0.95	0.95	0.95
GA-XGBoost	0.944	0.95	0.94	0.94	0.95	0.94
GBDT	0.922	0.99	0.87	0.86	0.98	0.92
ILAGA-GBDT	0.948	0.95	0.95	0.94	0.95	0.94
GA-GBDT	0.942	0.94	0.95	0.94	0.94	0.94
LightGBM	0.923	0.99	0.87	0.85	0.99	0.92
ILAGA-LightGBM	0.950	0.96	0.95	0.94	0.95	0.94
GA-LightGBM	0.941	0.95	0.94	0.93	0.94	0.93
RF	0.927	0.99	0.88	0.87	0.99	0.92
ILAGA-RF	0.946	0.95	0.94	0.94	0.94	0.94
GA-RF	0.940	0.94	0.93	0.94	0.93	0.93

上述指标以及 AUC 值对比雷达图可视化如图 4-7 所示。



（a）GBDT和XGBoost改进前后对比　　（b）LightGBM与RF改进前后对比

图 4-7 遗传算法改进前后各指标结果对比雷达图

从图 4-7 雷达图可得，图中共有五个维度，分别是 Precision-1，Recall-1，F1-score，Accuracy 以及 AUC，五个维度的含义如前所述。从雷达图可以明显看出利用改进的遗传算法优化前后的差别，改进后识别信贷高风险用户的查准率、低风险用户的召回率等维度有了显著提升。

四种集成学习算法改进前后的 ROC 曲线对比如图 4-8 所示，由 ROC 曲线可以看出，在评价指标 AUC 维度改进后模型效果提升了 1%到 2%。

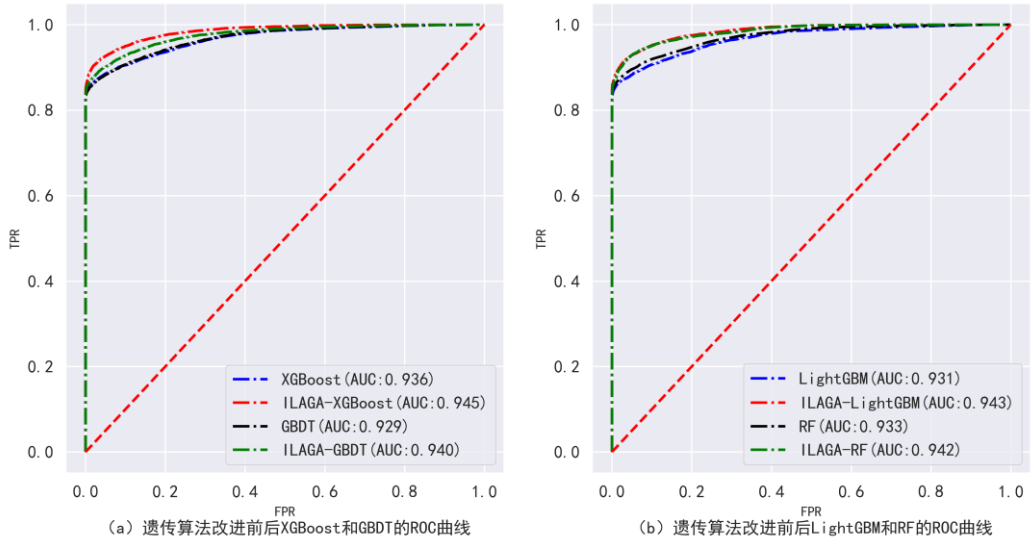


图 4-8 四种集成学习算法的遗传算法改进前后的 ROC 曲线对比

从各个模型预测的概率分布情况来看，四个集成学习算法模型的预测逾期概率往 0 和 1 两极聚集，中间段较少，证明模型对于是高风险或低风险用户具有较好的可分性，且各个模型的预测结果概率分布比较类似。各模型预测概率的分布小提琴图如图 4-9 所示。

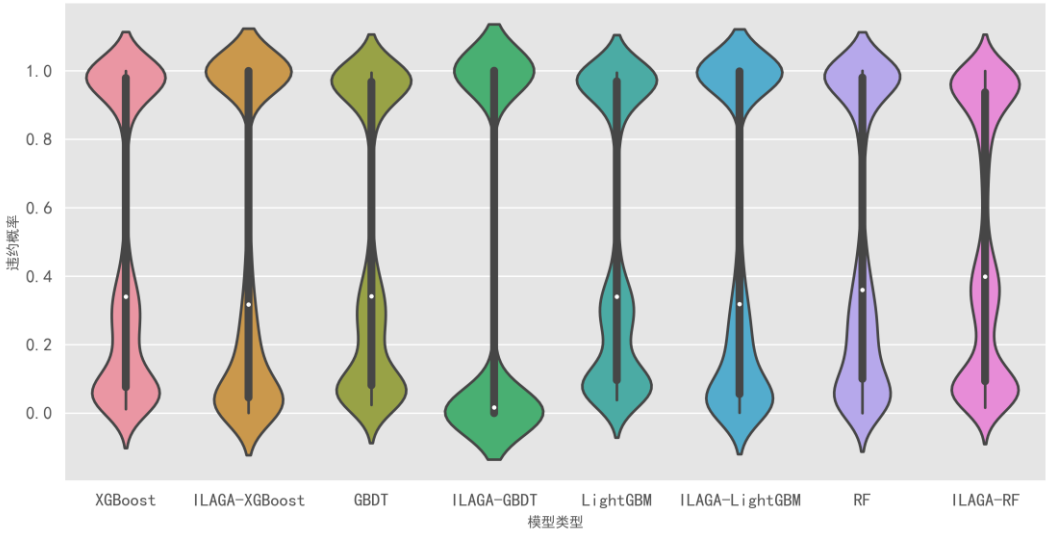


图 4-9 遗传算法改进前后概率分布图

4.4 随机搜索和网格搜索实验及对比分析

4.4.1 随机搜索和网格搜索理论概述

（1）网格搜索

网格搜索算法是指定参数值的一种穷举搜索方法，将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。即将各个参数可能的取值进行排列组合，列出所有可能的组合结果生成网格用于模型的训练。

（2）随机搜索

随机搜索算法由加拿大两位学者 Bergstra 和 Bengio 于 2012 年提出。当训练模型的超参数比较多的时候，采用网格搜索寻优的时间和计算资源会以指数级上升。而随机搜索是从抽样分布中抽取随机的参数组合，具有更高的搜索效率，在搜索次数相同时，随机搜索相对于网格搜索会尝试更多的参数值。

4.4.2 不同优化方式的实验参数结果

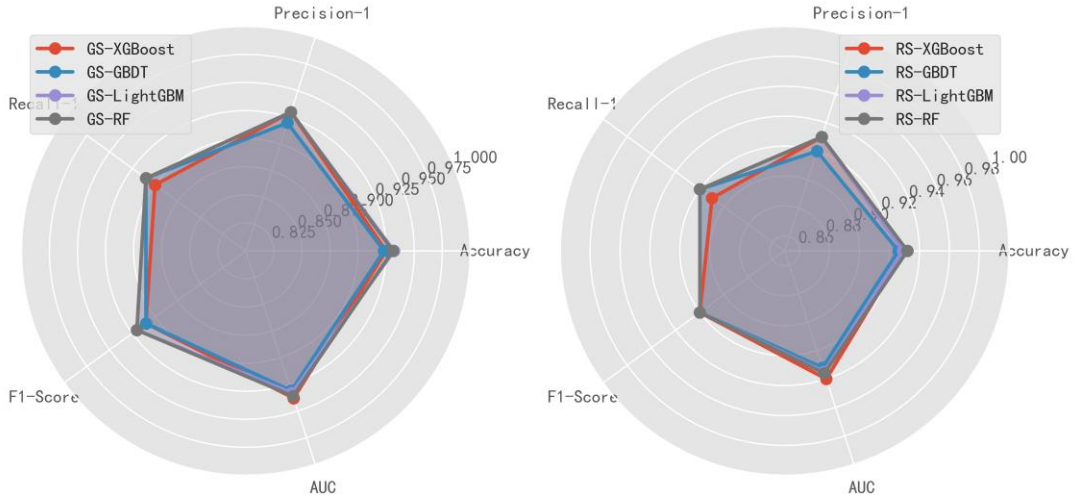
使用随机搜索和网格搜索对四种集成学习算法（GBDT、XGBoost、LightGBM、RF）进行参数寻优，搜索的参数域空间如表 4-1 所示。经过网格搜索和随机搜索在训练集和验证集上的仿真实验，获得的各预测模型超参数取值如表 4-4 所示。

表 4-4 网格搜索和随机搜索的参数结果

模型名称	超参数	参数说明
GS_XGBoost	learning_rate=0.20	max_depth=11
	n_estimators=196	min_child_weight=3.52
GS_GBDT	learning_rate=0.15	max_depth=12
	n_estimators=264	subsample=0.9
GS_LightGBM	learning_rate=0.16	max_depth=10
	n_estimators=386	colsample_bytree=0.86
GS_RF	n_estimators=136	max_depth=14
	min_samples_leaf=96	
RS_XGBoost	learning_rate=0.18	max_depth=10
	n_estimators=178	min_child_weight=3.2
RS_GBDT	learning_rate=0.16	max_depth=8
	n_estimators=508	subsample=0.9
RS_LightGBM	learning_rate=0.15	max_depth=12
	n_estimators=354	colsample_bytree=0.76
RS_RF	n_estimators=186	max_depth=14
	min_samples_leaf=69	

4.4.3 不同优化方式的实验结果对比

网格搜索和随机搜索效果对比结果如图 4-10 所示。结合算法改进前的单模型，由网格搜索和随机搜索优化得到的模型效果有了小范围的提升，在高风险用户查准率指标方面有了较大的提升。同时，将网格搜索及随机搜索的指标结果同遗传算法优化的结果进行对比，后者效果明显好于前者，也证明了使用遗传算法优化集成学习方法的有效性。



（a）网格搜索改进效果对比图　　（b）随机搜索改进效果对比图

图 4-10 网格搜索和随机搜索效果对比图

具体地，使用改进后的遗传算法、网格搜索算法、随机搜索算法这三种优化方式的预测结果如表 4-5 所示。

表 4-5 网格搜索和随机搜索的参数结果

模型名称	Accuracy	Precision-0	Recall-0	Precision-1	Recall-1	F1-score-1
GS-XGBoost	0.925	0.92	0.92	0.93	0.90	0.91
ILAGA-XGBoost	0.952	0.95	0.95	0.95	0.95	0.95
RS-XGBoost	0.930	0.92	0.93	0.93	0.91	0.92
GS-GBDT	0.923	0.92	0.92	0.92	0.91	0.91
ILAGA-GBDT	0.948	0.95	0.95	0.94	0.95	0.94
RS-GBDT	0.926	0.93	0.92	0.92	0.92	0.92
GS-LightGBM	0.932	0.92	0.91	0.93	0.91	0.92
ILAGA-LightGBM	0.950	0.96	0.95	0.94	0.95	0.94
RS-LightGBM	0.929	0.93	0.92	0.93	0.92	0.92
GS-RF	0.931	0.93	0.91	0.93	0.91	0.92
ILAGA-RF	0.946	0.95	0.94	0.94	0.94	0.94
RS-RF	0.932	0.93	0.92	0.93	0.92	0.92

由表 4-5 可以看出，使用网格搜索和随机搜索优化单模型的效果基本接近，对于低风险用户的预测查准率在 92%左右。对于高风险用户预测的平均查准率在 93%左右，

相比于集成学习单模型提升 6%左右，提升效果显著。实验证明，经过网格搜索和随机搜索的优化，模型在整体的准确率以及对于高风险用户预测的查准率均有了一定程度的提升，证明了网格搜索和随机搜索对模型优化的有效性。

此外，基于网格搜索优化后的集成学习模型在准确率指标下，效果最好的是 GS-LightGBM，其准确率为 0.932，相较于 LightGBM 单模型提升了近 1%。使用随机搜索优化后的 RS-RF 模型相较于 RF 提升了 0.5%。利用改进后的遗传算法优化集成学习方法准确率最高的是 ILAGA-XGBoost，其准确率达到了 0.952，对高风险用户预测的查准率达到了 0.95。相较于网格搜索和随机搜索，优化后的效果提升显著，验证了使用 ILAGA 算法优化集成学习方法的有效性。

4.5 算法改进后的模型融合仿真实验及对比分析

4.3 节使用基于交叉率和变异率的改进遗传算法对 GBDT、XGBoost、LightGBM 以及 Random Forest 模型进行优化，在其超参数域空间内进行寻优，寻找模型的近似最优解，建立了 ILAGA-GBDT、ILAGA-XGBoost、ILAGA-LightGBM 和 ILAGA-RF 模型，对比改进前的单模型各项评价指标有了一定幅度的提升。

为了进一步提升整体模型预测信贷风险的能力，使用 Stacking 算法对深度神经网络和 ILAGA-GBDT 等算法等做进一步融合。从多种算法和算法的不同参数两个维度出发建立信贷风险预测模型。

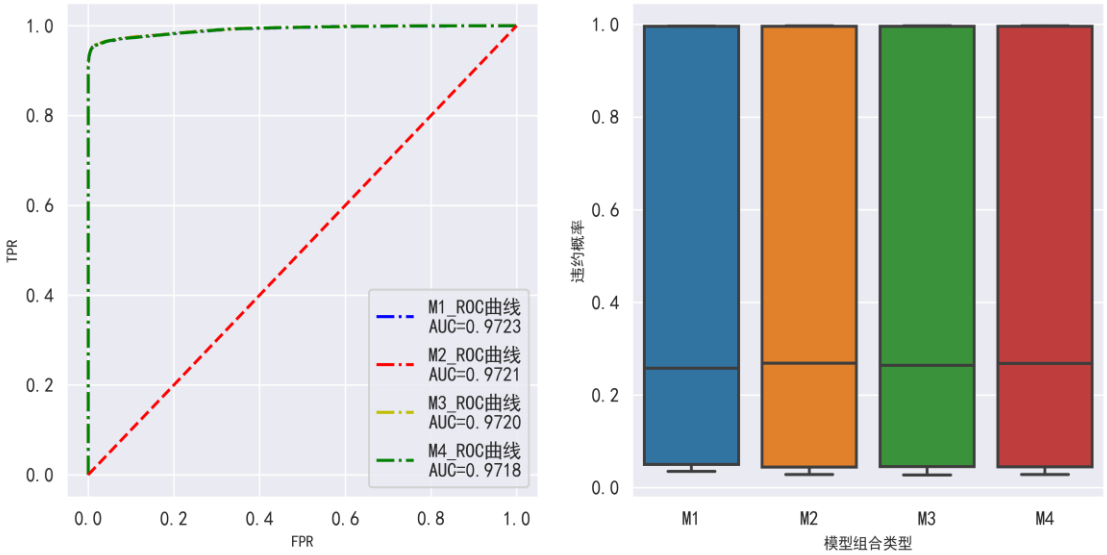
4.5.1 多种算法的模型融合

通过对不同的模型组合进行融合加强差异性，提升模型的效果。本文对多种模型的组合进行仿真实验，基学习器包括使用遗传算法改进前后的模型。随机选取两个及以上的单模型进行 Stacking 模型融合，分别比较各融合模型的效果。以表 4-6 中的四种组合为示例。

表 4-6 多种算法的组合示例

模型	单模型组合
M1	DNN、ILAGA-GBDT、ILAGA-XGBoost、ILAGA-LightGBM、ILAGA-RF、XGBoost、GBDT、LightGBM、RF
M2	DNN、ILAGA-GBDT、ILAGA-XGBoost、ILAGA-LightGBM、ILAGA-RF
M3	ILAGA-XGBoost 、ILAGA-GBDT、ILAGA-LightGBM、ILAGA-RF、RF
M4	DNN、ILAGA-XGBoost、ILAGA-RF、LightGBM

多种算法模型融合的实验方案同 3.4 节的多种算法实验步骤一样。模型预测结果的 AUC 值和 ROC 曲线以及概率分布效果如图 4-11 所示。



（a）各融合模型ROC曲线　　（b）各融合模型预测结果概率分布

图 4-11 M1、M2、M3、M4 模型的 ROC 曲线和违约概率分布

从图 4-11 的 ROC 曲线可知，AUC 的值均超过 0.97，相比于算法改进前的模型融合，AUC 值提升了超过 0.02，相比于集成学习单模型，AUC 值提升超过 0.05 以上，提升效果明显。四种组合的 ROC 曲线已基本重合，且与坐标轴围成一个近似正方形，均达到了较高的水平。同时，从 4-11 中的箱图看出，无论是预测概率分布箱图的中位数，还是上界、下界，多种算法融合生成的模型之间相似度较高。M1、M2、M3、M4 各模型的预测结果如表 4-7 所示。

表 4-7 多种算法的模型融合预测结果

		precision	recall	F1-score
M1	低风险	0.99	0.96	0.97
	高风险	0.96	0.98	0.97
	Accuracy	—	—	0.97
M2	低风险	0.98	0.96	0.97
	高风险	0.96	0.98	0.97
	Accuracy	—	—	0.97
M3	低风险	0.98	0.97	0.97
	高风险	0.96	0.98	0.97
	Accuracy	—	—	0.97
M4	低风险	0.98	0.96	0.97
	高风险	0.96	0.98	0.97
	Accuracy	—	—	0.97

经过多种算法的模型融合建立的 M1 到 M4 在查准率、准确率等方面均有明显的提升，平均提升相较于改进前单模型提升 5%以上，且总体模型准确率超过 97%。高风险用户预测的查准率和召回率都有了大幅度提升，尤其对于查准率来说，从单模型的 87%提升到 96%。预测低风险用户的召回率也从 86%提升到 96%，模型融合效果理想。

4.5.2 算法不同参数的模型融合

同种算法不同参数能训练出不同的模型，通过算法参数小范围扰动加强个体学习器之间的差异性。利用算法的不同参数做融合的实验方案同 3.4 节，除了由遗传算法搜索出的参数组合，另对其他四种集成学习算法参数做小范围扰动，训练生成不同的模型，并使用 Stacking 融合。算法的不同参数融合模型的预测结果如表 4-8。

表 4-8 算法的不同参数融合模型

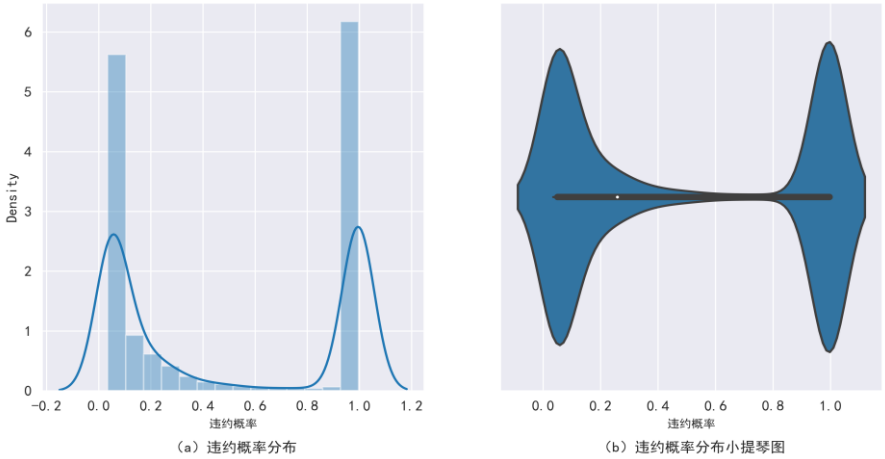
	precision	recall	F1-score
低风险用户	0.99	0.96	0.97
高风险用户	0.96	0.98	0.97
accuracy	—	—	0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97

从表 4-8 的预测结果来看，多种算法和算法的不同参数生成的融合模型效果基本持平，在测试集上预测效果相近。通过对算法的不同参数做模型融合，信贷风险预测模型的效果显著提升。

4.6 信贷风险预测模型的最终建立

通过上述实验对比，选出泛化性能最优的模型，确定为信贷风险的预测模型，即通过 Stacking 对 DNN、ILAGA-GBDT、ILAGA-XGBoost、ILAGA-LightGBM、XGBoost 等九种模型进行模型融合建立最终模型。最终建立的模型预测概率分布如图 4-11。

从图 4-12 中的柱状图和小提琴图可以看出模型预测的概率集中在两端，中间较少，尤其违约概率在[0.4，0.8]之间。概率分布在接近 0 和 1 的样本较多，证明了模型对分类结果有较高的确信度，也可以说明模型对是否违约风险有较为准确的判定，可以将正负样本较为“安全”的识别出来，证明了模型具有较好的泛化性能。



（a）违约概率分布　　（b）违约概率分布小提琴图

图 4-12 最终模型预测结果的概率分布

最终模型在测试集上预测结果如表 4-9 所示。从表 4-9 看，模型整体的查准率、准确率、召回率以及 F1-Score 均达到 96%以上，提升效果显著。对于高风险用户识别效果准而全，对于低风险用户的识别同样如此。这不仅有助于帮助平台发现高风险用户，减少贷款带来的损失，同时也可以挖掘优质客户，增加平台的收益。

表 4-9 最终模型预测结果

	precision	recall	F1-score
低风险	0.99	0.96	0.97
高风险	0.96	0.98	0.97
accuracy	—	—	0.972
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97

最终模型预测结果的 ROC 曲线和 P-R 曲线如图 4-13，最终模型预测结果的 AUC 值达到 0.9725，相较于改进前的单模型，提升超过 5%，模型的泛化性能较优。从 P-R 曲线来看，其与坐标轴围成的面积已接近 1，红色虚线与 P-R 曲线的交点即平衡点也已接近最优。

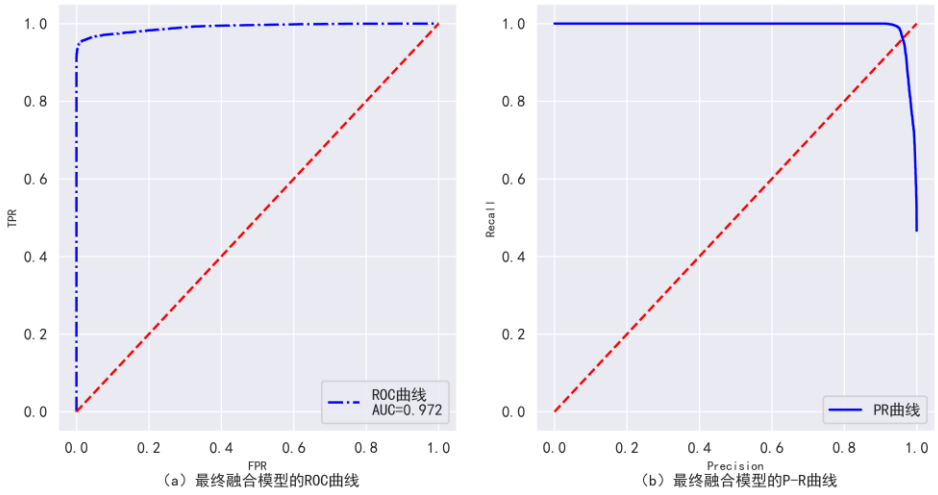


图 4-13 最终模型预测结果的 ROC 曲线和 P-R 曲线

4.7 本章小结

本章重点介绍了改进的线性自适应遗传算法的过程以及使用该算法优化 GBDT、XGBoost、LightGBM、Random Forest 这四个单模型超参数的过程。改进的遗传算法 ILAGA 针对四个模型确定不同的域空间，优化不同的超参数组合，有效解决了该类模型容易陷入局部最优等问题。完成并提升了单模型的信贷风险预测的准确性和泛化能力。接着对比了网格搜索和随机搜索优化四种集成学习算法的效果，实验证明，使用遗传算法优化的效果明显优于网格搜索和随机搜索。紧接着从多种算法和算法的不同参数增加模型之间的差异性，提升模型融合的效果。最终经过多个维度的比较，选定通过

Stacking 对深度神经网络 DNN、ILAGA-GBDT、ILAGA-XGBoost、ILAGA-LightGBM、XGBoost 等九种模型进行模型融合建立最终的模型，并对模型预测的结果和概率的分布进行了可视化研究。

总结与展望

本文总结

本文探讨建立了一种基于改进的遗传算法与多模型融合的信贷风险预测模型。信贷风险管理一直是金融服务行业重要的组成部分，尤其对于互联网金融服务尤为重要。而在当前传统的个人及企业信用风险评估形式下，主要依赖于中国人民银行的征信系统。如前所述，央行的征信系统存在大量的数据缺失，会大大限制评估客户的范围及评估的准确性。对于互联网金融企业，信贷业务由于用户的宏观因素变化等问题，导致用户风险模型预测精确度低，客户评级不准确，导致不良贷款余额和不良贷款率的增加。基于此问题，本文通过分析国内外的研究动态，从用户的基本信息、银行卡消费记录、信用卡账单以及用户的浏览行为出发，挖掘信用风险相关特征，并利用深度神经网络、集成学习方法以及多模型融合技术建立了个人信贷风险预测模型。同时，针对遗传算法容易陷入局部最优的缺陷，基于交叉率和变异建立了改进的线性自适应遗传算法。并使用改进后的遗传算法优化集成学习方法，寻找其最优参数组合。本文的研究工作可以概括为以下几个方面：

1、本文根据脱敏的用户行为数据，挖掘信贷风险关联因素，并基于建立的序列浮动双向搜索算法选择了 198 维显著性特征，构建了基于深度神经网络 DNN、集成学习算法的用户信贷风险预测模型。

2、引入多模型融合技术，确定各基学习器预测结果的最佳权重。通过单模型和融合后模型的结果对比分析，多模型融合比单模型的预测结果有显著提升。

3、本文建立了基于交叉概率和变异概率的改进线性自适应遗传算法，并使用改进后的算法对四种集成学习方法进行优化，寻找最优的模型训练超参数组合。同时，从不同模型及模型的不同参数两方面构建融合模型，在测试集上取得了较好的预测效果。

进一步的研究方向

1、本论文研究中使用的训练集数据量较小，且存在大量的数据缺失，可能存在泛用性问题。若需要对模型推广使用，需要对模型做进一步的研究。

2、本文引入的遗传算法只能以一定的概率找到近似最优的超参数组合。同时，遗传算法在面对大规模数据集上表现较差。

致 谢

论文至此已接近尾声，回顾过去的三年，依然历历在目。三年的时光，匆匆从时间的缝隙里流走。有人说，时间就像是一笔贷款，无论是谁，无论其信用再好，借了也还不起。依稀还记得第一次来到交大的校园是 2018 年复试的场景，怀着忐忑和紧张情绪从学校的南大门进入，穿过银杏大道，被巍峨高大的图书馆所震撼。应该说，三年的大部分时光均在实验室度过。在这里，我享受着学习的快乐时光，享受着与老师、同门一起探讨专业知识的时刻，不仅收获了相关知识，也收获了更多的友谊。在此，我希望借这个机会对所有帮助过我的人表示由衷的感谢。

在交大的三年时间，首先非常感谢我的研究生导师楼新远老师。无论是在学习上、生活上，还是在找工作阶段，都给予了我极大的帮助和谆谆教导。从每周的学术研讨会，到与师兄参与科技竞赛，再到论文的发表，都给予了悉心的指导。楼新远老师博学多闻、治学严谨，对我三年的学习生涯有着深远的影响。最后本篇学位论文的撰写，从题目的选定、文献的查找、技术路线的确定、关键实验到最后论文的完成，都给出了详细深入的指导意见，再次感谢楼老师。

在此，我还想要感谢我的女朋友赫倚以及我的家人在生活、工作等各方面否给予了我大力支持。让我对未来充满了期待，让我更有动力、更从容的走向社会。同时，我还要感谢实验室的同门涂钰、王越楚、徐晶晶，师兄师姐以及师弟师妹等，在学习方面，我们共同探讨，一起分享科研知识，生活上，我们一起锻炼，一起玩耍，感谢你们！

再次由衷感谢所有关心和帮助我的人，祝愿大家前程似锦，生活美满！

参考文献

[1] 张承钊. 浅谈互联网金融如何更好地为现代经济发展服务[J]. 中国集体经济, 2020(24):91-92.

[2] Mauromicale G, Portis E, Acquadro A, et al. An integrated model to accelerate t he development of seed-propagated varieties of globe artichoke[J]. Crop breeding and applied biotechnology, 2018, 18(1):72-80.

[3] 张鑫, 曹帅. 基于 BP 神经网络的 P2P 网贷信用风险甄别研究——以拍拍贷为实例[J]. 现代商业, 2020(011):105-106.

[4] 无[1]. 我国征信业的改革与发展[J]. 中国银行业, 2019(008):19-21.

[5] Zuckerman M, Knee C R, Kieffer S C, et al. What individuals believe they can and cannot do: explorations of realistic and unrealistic control beliefs[J]. Journal of Personality Assessment, 2004, 82(2):215-232.

[6] Jia X, Heidhues F, Zeller M. Credit rationing of rural households in China[J]. Agricultural Finance Review, 2010, 70(1):37-54.

[7] 沙希杜·伊斯兰姆, 和红梅, 基肖尔·库梅尔·班萨克. 孟加拉国视角下的“一带一路”及孟中印缅经济走廊建设[J]. 南亚东南亚研究, 2018(03):89-100.

[8] 闫真宇. 关于当前互联网金融风险的若干思考[J]. 浙江金融, 2013(12):40-42.

[9] 王裕粟. 基于数据挖掘的客户信用评级模型的设计与实现[D]. 电子科技大学, 2012.

[10] 洪娟, 曹彬, 李鑫. 互联网金融风险的特殊性及其监管策略研究[J]. 中央财经大学学报, 2014(09):42-46.

[11] 周贤. 新形势下银行信贷风险管理问题研究[J]. 中国市场, 2020(06):14+65.

[12] 刘佳蒙. 商业银行信贷风险管理存在的问题与对策研究[J]. 农村经济与科技, 2020(04):166-167.

[13] 徐溪蔓. 基于时间序列模型的商行信贷规模与风险管理分析[J]. 现代商业, 2019(24):136-137.

[14] 李帅鹏. 基于贝叶斯决策规则的商业银行信贷风险研究[D]. 淮北师范大学, 2020.

[15] Booker K M, Gadgil A J, Winickoff D E. Engineering for the global poor: The role of intellectual property[J]. Science & Public Policy, 2014, 39(6):775-786.

[16] Sandoval, Leonidas. Structure of a Global Network of Financial Companies Based on Transfer Entropy[J]. Entropy, 2014, 16(8):4443-4482.

[17] Mcquay T, Cavoukian A. A pragmatic approach to privacy risk optimization: privacy by design for business practices[J]. Identity in the Information Society, 2010, 3(2):379-396.

[18]李长征. 完善我国互联网金融风险管理机制的建议[J]. 商场现代化, 2019(04):92-94.
[19]章豪. 借款人未按协议履行分期还款义务应视为预期违约[J]. 商, 2014(018):136-137.
[20]林荫. 基于大数据分析的银行不良信贷风险模型[J]. 工程经济, 2015(06):112-118.
[21]陈晓兰, 任萍. 基于 Logistic 混合模型的企业信用风险评价研究[J]. 山东财政学院学报, 2011(02):90-93.
[22]喻光丽. 基于 Logistic 回归模型的 P2P 网络借贷平台借款人信用风险评估研究[D]. 兰州大学, 2017.
[23]呼振凯. P2P 网络借贷中借款人的信用风险评估研究[D]. 哈尔滨工业大学, 2016.
[24]Wong W E, Qi Y U. BP NEURAL NETWORK-BASED EFFECTIVE FAULT LOCALIZATION[J]. International Journal of Software Engineering and Knowledge Engineering, 2011, 19(4):573-597.
[25]高海兵, 高亮, 周驰, 喻道远. 基于粒子群优化的神经网络训练算法研究[J]. 电子学报, 2004(09):1572-1574.
[26]Wu W, Jian W, Cheng M, et al. Convergence analysis of online gradient method for BP neural networks[J]. NEURAL NETWORKS -OXFORD, 2011, 24(1):91-98.
[27]Bishop, C. Novelty Detection and Neural Network Validation[J]. IEE Proceedings - Vision Image and Signal Processing, 1994, 141(4):217-222.
[28]方先明, 熊鹏, 张谊浩. 基于 Hopfield 神经网络的信用风险评价模型及其应用[J]. 中央财经大学学报, 2007(008):34-40.
[29]张佳维. 基于模糊神经网络的个人信用风险评估[D]. 内蒙古大学, 2014.
[30]于玲, 吴铁军. 集成学习:Boosting 算法综述[J]. 模式识别与人工智能, 2004(01):52-59.
[31]Lary D J, Alavi A H, Gandomi A H, et al. Machine learning in geosciences and remote sensing[J]. Geoence Frontiers, 2016, 7(1):3-10.
[32]Cmv A, Jie D B. Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data[J]. Neural Networks, 2020, 128:268-278.
[33]Venter A, Laurie D P. A Doubly Adaptive Integration Algorithm Using Stratified Rules[J]. BIT, 2002, 42(1):183-193.
[34]Liu, Jiaming,Wu, Chong．A gradient-boosting decision-tree approach for firm failure prediction: an empirical model evaluation of Chinese listed companies[J]．The Journal of Risk Model Validation, 2017,11(2):43-64.
[35]尚朝轩, 王品, 韩壮志,等. 基于类决策树分类的特征层融合识别算法[J]. 控制与决策, 2016, 31(006):1009-1014.

[36]徐鹏, 林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009(10):2692-2704.

[37]曹礼园, 李深洛. 基于多目标最优化的最小代价决策树构建与实现[J]. 计算机与数字工程, 2019(12):3020-3024.

[38]付冬梅, 练丁槫. 基于广度优先遍历的关键路线生成树算法[J]. computer science&application, 2012(2):51-56.

[39]李超, 张文辉, 李然,等. 基于 Stacking 集成学习的恒星/星系分类研究[J]. 天文学报, 2020(002):102-111.

[40]赵国强, 王会进. 一种用于大规模数据集的决策树采样策略[J]. 微型机与应用, 2010(021):5-6.

[41]Wang D , Yang Z, Yi Z. LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients[C]. The 2017 International Conference on Computational Biology and Bioinformatics, Nha Trang. ICCBB, 2017:7-11.

[42]谢勇, 项薇, 季孟忠. 基于 Xgboost 和 LightGBM 算法预测住房月租金的应用分析[J]. 计算机应用与软件, 2019(09):151-155+191.

[43]熊苏生. 基于改进 LightGBM 的交通模式识别算法[J]. 计算机与现代化, 2018(10):72-77+130.

[44]王慧芳, 张晨宇. 采用极限梯度提升算法的电力系统电压稳定裕度预测[J]. 浙江大学学报（工学版）, 2020(003):606-613.

[45]王玉晶, 莫建麟. 基于 TS 特征选择的生理情感状态分类[J]. 齐齐哈尔大学学报:自然科学版, 2013(3):19-19.

[46]佟为明, 刘喆. 快速变化的动态信号的一种采样方法[J]. 微处理机, 1995(02):43-45.

[47]X. Wang, X. Y. Lou, S. Y. Hu and S. C. He. Evaluation of safe driving behavior of transport vehicles based on K-SVM-XGBoost[J]. Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020:84-92.

[48]简艺恒, 余啸. 基于数据过采样和集成学习的软件缺陷数目预测方法[J]. 计算机应用, 2018(09):2637-2643+2659.

[49]Hussain M, Wajid S K, Elzaart A, et al. A Comparison of SVM Kernel Functions for Breast Cancer Detection[C]. Eighth International Conference Computer Graphics, Hefei. IEEE Computer Society, 2011:145-150.

[50]Solaiman B, Debon R. Information fusion: application to data and model fusion for ultrasound image segmentation.[J]. IEEE Transactions on Biomedical Engineering, 1999, 46(10):1171-1175.

[51]Tarabalka Y, Fauvel M, Chanussot J, et al. SVM and MRF-Based Method for Accurate Classification of Hyperspectral Images[J]. IEEE Geoscience & Remote Sensing Letters, 2010, 7(4):736-740.

[52] Jeong H K, Lee Y P, Lahaye R J W E, et al. Evidence of Graphitic AB Stacking Order of Graphite Oxides[J]. Journal of the American Chemical Society, 2008, 130(4):1362-1366.

[53] Baurle R A, Tam C J, Edwards J R, et al. Hybrid simulation approach for cavity flows: Blending, algorithm, and boundary treatment issues[J]. Aiaa Journal, 2003, 41(8):1463-1480.

[54] 张慧敏, 宋东, 郭勇,等. 故障预测模型的评价方法研究[J]. 测控技术, 2013(005):121-124,129.

[55] 汪云云, 陈松灿. 基于 AUC 的分类器评价和设计综述[J]. 模式识别与人工智能, 2011(01):64-71.

[56] Grant T, Kluge A G. Data exploration in phylogenetic inference: scientific, heuristic, or neither[J]. Cladistics-the International Journal of the Willi Hennig Society, 2003, 19(5):379-418.

[57] Lee M L, Hsu W, Kothari V. Cleaning the spurious links in data[J]. Intelligent Systems IEEE, 2003, 19(2):28-33.

[58] Lekadir K, Merrifield R, Yang G Z. Outlier Detection and Handling for Robust 3-D Active Shape Models Search[J]. IEEE Transactions on Medical Imaging, 2007, 26:212-222.

[59] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1):208-221.

[60] Liu H, Che W, Liu T. Feature Engineering for Chinese Semantic Role Labeling[J]. Journal of Chinese Information Processing, 2007, 21(1):79-84.

[61] Chandrashekar G, Sahin F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014, 40(1):16-28.

[62] Bing X, Zhang M, Browne W N, et al. A Survey on Evolutionary Computation Approaches to Feature Selection[J]. IEEE Transactions on Evolutionary Computation, 2016, 20(4):606-626.

[63] 陈岩, 来海锋, 王清,等. 基于 filter-wrapper 的两步特征变量提取方法[J]. 机电工程, 2010, 27(4):67-71.

[64] 周传华, 柳智才, 丁敬安,等. 基于 filter+wrapper 模式的特征选择算法[J]. 计算机应用研究, 2019(007):1975-1979,2010.

[65] Mistry K, Zhang L, Neoh S C, et al. A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition[J]. IEEE Transactions on Cybernetics, 2017, 47(6):1496-1509.

[66] Yao L, Sethares W A, Kammer D C. Sensor placement for on-orbit modal identification via a genetic algorithm[J]. AIAA Journal, 2012, 31(10):1922-1928.

[67] Yang J, Honavar V. Feature subset selection using a genetic algorithm[J]. IEEE Intelligent Systems & Their Applications, 2002, 13(2):44-49.

[68] Aickelin U, Dowsland K. Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem[J]. Journal of Scheduling, 2015, 3(3):139-153.

[69] Moradi M H, Abedini M. A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems[J]. International Journal of Electrical Power & Energy Systems, 2010, 34(1):66-74.

[70] 李建锋, 彭舰. 云计算环境下基于改进遗传算法的任务调度算法[J]. 计算机应用, 2011, 31(001):184-186.

攻读硕士学位期间发表的论文及科研成果

1. 已发表的论文：

[1] X. Wang, X. Y. Lou, S. Y. Hu and S. C. He. Evaluation of safe driving behavior of transport vehicles based on K-SVM-XGBoost[J]. Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020:84-92.

2. 参与科研项目：

[1]四川省重点项目：面向汽车服务生命周期（SLM）的第三方制造服务云平台研发与应用示范，项目编号：2015GZ0076