

同等学力申请硕士学位论文

|  |  |
| --- | --- |
| 题目： | **P2P平台风控数据分析** |
|  | **系统的研究与实现** |

|  |  |
| --- | --- |
| 姓 名： | 殷艳梅 |
| 学 号： | S1548029 |
| 院 系： | 信息科学技术学院 |
| 专 业： | 计算机应用技术 |
| 导师姓名： | 刘学洋 |

二〇二〇年三月

**版权声明**

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

P2P借贷是一种个人对个人的信用借贷形式，作为中间介质撮合出借人和借款人成交，并签订借款合同。随着国外P2P先进行业经验的引进和国内互联网以及民间贷款的兴起，P2P网络借款已经成为一种新型的不可忽视的业务模式。与传统银行贷款不同，P2P借贷的线上进行、信息不对称、数据量大的特点增加了用户风险评估的难度。现阶段，借款方主要通过人工审核的方法对用户风险进行评估，人力消耗巨大且审批的时效性缓慢，且审批结果依赖审批人的经验。

本文旨在提高P2P用户风险评估的效率与准确度，探索自动化用户风险评估方法，提出了基于机器学习的用户风险评估技术，开发了相应的支持工具。具体来说，本文的主要研究内容包括：

1. **提出了基于机器学习的用户风险评估方法(MBRA)：**利用现有借款历史数据，通过特征集分割加权SVM算法结合,建立客户风险预测模型，根据训练出的模型预测用户的信用评分，然后将用户信用评分指导用户借款审核，降低了P2P平台放贷面临的风险和损失。
2. **开发了风险控制的数据分析系统**：将客户风险预测模型以微服务的形式嵌入到P2P平台。风控数据分析系统主要包括以下模块：数据采集模块，数据预处理模块，数据挖掘建模分析模块，统计分析模块。P2P平台嵌入风控数据分析系统后，可以线上完成数据采集和客户风险预测功能，使风控准确率和审批时效得到了大幅度提高。
3. **采用经验研究验证了MSBA方法的有效性：**利用真实的数据评估与比较了MBRA与传统方法的效率，评估结果表明MBRA将客户风险预测的准确度提高到96%，将贷款审批周期从原来的平均10个工作日缩短到4个工作日，并且参与贷款审核的工作人员下降到原来的三分之一。

本文提出的MBRA改进了传统用户风险评估方法的局限性，提高了用户风险评估的效率，开发的支持工具提高了MBRA的自动化程度。

关键词：风险评估，建模分析，算法

P2P PLATFORM RISK CONTROL DATA ANALYSIS SYSTEM RESEARCH AND IMPLEMENTATION

Yin Yanmei (Technology of Computer Application)

Directed by Mr.Liu Xueyang

# ABSTRACT

P2P lending is a form of personal to personal credit lending. P2P platform, as an intermediate medium, helps lenders and borrowers to conclude transactions and sign loan contracts. Risk control data analysis system is a very important part in P2P platform construction. Different from traditional bank loans, P2P lending has the following characteristics: online, asymmetric information, large amount of data, etc. The main process of P2P lending includes three stages: loan application, user risk assessment and loan approval. The current situation of user risk assessment and loan approval mainly includes multi-level manual audit. The manpower consumption is huge and the timeliness of the approval is slow. The approval result mostly depends on the experience of the approver, which has the problem of strong subjectivity and low accuracy.

In this paper, we use the existing loan history data of P2P platform, combined with the feature set segmentation weighted SVM algorithm, to build a customer risk prediction model, and use the model trained by the history data to predict the credit score of users. Then the user credit score is used to guide the user loan review, which reduces the risk and loss faced by P2P platform lending. The risk control data analysis system is designed to embed the customer risk prediction model into P2P platform in the form of micro service. The risk control data analysis system mainly includes the following modules: data acquisition module, data preprocessing module, data mining modeling and analysis module, statistical analysis module. After the P2P platform is embedded in the risk control data analysis system, it can complete the data collection and customer risk prediction functions online, which greatly improves the accuracy of risk control and the effectiveness of approval.

It can be seen from the evaluation results that the risk control data analysis system improves the accuracy of customer risk prediction to more than 96%, shortens the loan approval cycle from the original average of 10 working days to within 4 working days, and reduces the number of staff participating in loan review to one third of the original. In a word, the construction and use of risk control data analysis system greatly improves the accuracy of customer risk assessment, reduces the risk of P2P loan, and effectively saves manpower and time cost.

Keywords: risk assessment, modeling analysis, algorithm

# 目录

[摘要 I](#_Toc31982267)

[ABSTRACT II](#_Toc31982268)

[目录 IV](#_Toc31982269)

[1 绪论 1](#_Toc31982270)

[1.1 项目背景 1](#_Toc31982271)

[1.2 机器学习算法研究现状 2](#_Toc31982272)

[1.2.1  国内外机器学习算法研究现状 2](#_Toc31982273)

[1.2.2  机器学习算法优化 3](#_Toc31982274)

[1.3 研究目的和意义 3](#_Toc31982275)

[1.4 论文的组织结构 5](#_Toc31982276)

[2 相关技术研究 6](#_Toc31982277)

[2.1  数据分析流程 6](#_Toc31982278)

[2.2 机器学习相关算法研究 7](#_Toc31982279)

[2.2.1  SVM算法 7](#_Toc31982280)

[2.2.2 逻辑回归算法 9](#_Toc31982281)

[2.2.3 随机森林算法 11](#_Toc31982282)

[2.2.4 算法评价标准 11](#_Toc31982283)

[2.4 层次分析获取权重方法 12](#_Toc31982284)

[2.4.1 层次分析原理 12](#_Toc31982285)

[2.4.2 层次分析步骤 12](#_Toc31982286)

[2.5 本章小结 13](#_Toc31982287)

[3 系统需求分析和总体设计 15](#_Toc31982288)

[3.1 需求分析 15](#_Toc31982289)

[3.1.1 系统业务需求 15](#_Toc31982290)

[3.1.2 系统功能需求 17](#_Toc31982291)

[3.1.3 系统非功能需求 20](#_Toc31982292)

[3.2 系统总体设计 20](#_Toc31982293)

[3.2.1 系统架构设计 20](#_Toc31982294)

[3.2.2 系统功能设计 21](#_Toc31982295)

[3.2.3 风险评估模型设计 23](#_Toc31982296)

[3.3 系统数据库设计 24](#_Toc31982297)

[3.4 本章小结 26](#_Toc31982298)

[4 数据处理 27](#_Toc31982299)

[4.1 数据获取 27](#_Toc31982300)

[4.2 数据处理 27](#_Toc31982301)

[4.2.1 数据抽取 27](#_Toc31982302)

[4.2.2 异常处理 28](#_Toc31982303)

[4.3 特征指标工程 29](#_Toc31982304)

[4.3.1 特征指标量化 29](#_Toc31982305)

[4.3.2 特征指标标准化 30](#_Toc31982306)

[4.3.3 特征指标选择权值确定 31](#_Toc31982307)

[4.3 本章小结 33](#_Toc31982308)

[5 风控数据分析设计与实现 34](#_Toc31982309)

[5.1 风险评估模型实现 34](#_Toc31982310)

[5.1.1 算法筛选 34](#_Toc31982311)

[5.1.2 特征集分割 39](#_Toc31982312)

[5.1.3 模型优化 40](#_Toc31982313)

[5.1.3 模型评价 43](#_Toc31982314)

[5.2 风控数据分析设计与实现 44](#_Toc31982315)

[5.2.1 数据处理模块设计与实现 44](#_Toc31982316)

[5.2.2 数据挖掘建模分析模块设计与实现 47](#_Toc31982317)

[5.2.3 数据统计分析模块设计与实现 48](#_Toc31982318)

[5.3 本章小结 49](#_Toc31982319)

[6 总结与展望 50](#_Toc31982320)

[6.1 总结 50](#_Toc31982321)

[6.2 后续工作与展望 50](#_Toc31982322)

[参考文献 51](#_Toc31982323)

[本人攻读硕士学位期间发表的学术论文 53](#_Toc31982324)

[致谢 54](#_Toc31982325)

[北京大学学位论文原创性声明和使用授权说明 55](#_Toc31982326)

**表目录**

[表2- 1 混淆矩阵 11](#_Toc31982446)

[表3- 1 数据预处理功能表 17](#_Toc31982392)

[表3- 2 建设模型功能表 18](#_Toc31982393)

[表3- 3 风险评估功能表 18](#_Toc31982394)

[表3- 4 统计分析功能表 18](#_Toc31982395)

[表3- 5 主要数据库表 25](#_Toc31982396)

[表4- 1 数据源组成 27](#_Toc31982467)

[表4- 2 初步抽取特征指标集合 28](#_Toc31982468)

[表4- 3 学历编码表 29](#_Toc31982469)

[表4- 4 单位性质编码表 29](#_Toc31982470)

[表4- 5 特征指标标准化前 30](#_Toc31982471)

[表4- 6 特征指标标准化后 30](#_Toc31982472)

[表4- 7 初步抽取特征指标 31](#_Toc31982473)

[表5- 1 五大行业风险分析模型 34](#_Toc31982480)

[表5- 2 SVM算法训练结果 35](#_Toc31982481)

[表5- 3 逻辑回归算法训练结果 37](#_Toc31982482)

[表5- 4 随机森林算法训练结果 38](#_Toc31982483)

[表5- 5 算法筛选结果 39](#_Toc31982484)

[表5- 6 待分割特征指标集 39](#_Toc31982485)

[表5- 7 训练集模型评价结果 43](#_Toc31982486)

[表5- 8 训练集模型评价结果 44](#_Toc31982487)

**图目录**

[图1- 1 传统借款审批流程 1](#_Toc31967117)

[图1- 2 新客户借款审核流程 4](#_Toc31967118)

[图2- 1 SVM算法图 8](#_Toc31982495)

[图2- 2 Sigmoid 函数图 10](#_Toc31982496)

[图2- 3 层次模型图 13](#_Toc31982497)

[图3- 1 业务流程用例图 16](#_Toc31982506)

[图3- 2 业务流程图 16](#_Toc31982507)

[图3- 3 流程时序图 17](#_Toc31982508)

[图3- 4 模型创建流程图 19](#_Toc31982509)

[图3- 5 风险评估流程 19](#_Toc31982510)

[图3- 6 系统架构图 21](#_Toc31982511)

[图3- 7 系统功能设计图 22](#_Toc31982512)

[图3- 8 系统功能结构图 23](#_Toc31982513)

[图3- 9 风险评估模型构建过程 24](#_Toc31982514)

[图3- 10 实体关系图 24](#_Toc31982515)

[图3- 11 风控数据分析系统数据库设计 25](#_Toc31982516)

[图4- 1 借款金额分布图 28](#_Toc31982520)

[图4- 2 去异常借款金额散点分布图 29](#_Toc31982521)

[图4- 3 特征指标权重 32](#_Toc31982522)

[图5- 1 SVM算法训练流程 35](#_Toc31982527)

[图5- 2 逻辑回归算法训练流程 36](#_Toc31982528)

[图5- 3 随机森林算法训练流程 38](#_Toc31982529)

[图5- 4 加权SVM算法模型 41](#_Toc31982530)

[图5- 5 SVM算法优化流程 42](#_Toc31982531)

[图5- 6 模型评价流程图 43](#_Toc31982532)

[图5- 7 数据处理模块类图 44](#_Toc31982533)

[图5- 8 数据处理流程图 45](#_Toc31982534)

[图5- 9 数据源数据 45](#_Toc31982535)

[图5- 10 用户页面提示 46](#_Toc31982536)

[图5- 11 数据处理结 46](#_Toc31982537)

[图5- 12 数据挖掘建模分析模块类图 47](#_Toc31982538)

[图5- 13 风险评估流程 48](#_Toc31982539)

[图5- 14 风险评估结果 48](#_Toc31982540)

[图5- 15 数据统计分析模块类图 49](#_Toc31982541)

[图5- 16 统计分析页面 49](#_Toc31982542)

# 1 绪论

## 1.1 项目背景

随着互联网技术的发展，互联网金融应运而生。传统银行业务已经无法满足人们日益发展的金融支付需求。P2P金融平台登上历史舞台已有一段时间，世界第一家P2P金融平台的名字是Zopa，起源于英国，中国第一个P2P借贷平台是拍拍贷[1]。P2P平台给人们的生活提供了极大便利，平台建设同样也面临多方面的挑战。风险分析在P2P平台建设中是最重要部分。目前P2P平台客户风险分析主要面临以下挑战：

（1）客户风险识别的准确性低

由于线上交易，用户的数量增长迅速，用户信息种类繁多，用户信用难于量化评估等诸多原因，致使客户风险识别难度增加，传统的贷款审核过程主要借助平台数据统计分析，更多的靠专家经验人为判断用户的风险程度决定贷款批核结果。风险客户识别的准确率低造成P2P平台的贷款逾期、欺诈等案件频发。根据对平台历史数据的统计结果显示，风险客户审批拒贷率不足70%。

（2）客户贷款审核时效性

另外随着借款用户量不断增加，需要处理的数据量越来越大。传统方法是线下人工进行多级贷款审核。审核人工依赖性强，层层审批时效性差，已经无法支撑P2P平台的日益增加借款审核需要。传统审核流程如图1-1：



图1- 1 传统借款审批流程

这就需要建立风控数据分析系统，需要对影响用户信用的诸多因素进行提取，并且建立科学合理的风险评估流程。利用机器学习算法线上对客户的大量数据进行挖掘建模分析和风险评估，通过模型算法训练提高客户风险评估的准确性。以线上风险评估和借款审核代替传统的线下人工多层审核，为企业节约人力成本，提高贷款审批时效。帮助P2P平台过滤劣质不良客户，提升优质客户服务效率。

## 1.2 机器学习算法研究现状

### 1.2.1  国内外机器学习算法研究现状

风险分析是互金行业最重要的研究内容之一。最早在国外已经开展，并且应用于商业服务，也产生了效益。国外某些比较发达的国家拥有较完善的信用体系，这些国家的网贷平台风控技术水平较高。国外最初也采用经验分析法进行风险评估，常用的方法包括5C法和5P法。经验分析法根据专家的经验进行人工评估主观性较强，评估结果往往不够准确。随着科学技术的不断发展，挖掘建模分析方法大量应用于金融风险评估领域。 1980年，逻辑回归算法模型被Wiginton 应用于风险评估，对比人工经验逻辑，回归算法模型准确度更高。1996年，神经网算法被Desai 发现并应用于风险评分。研究发现在处理复杂非线性变量时，神经网络算法比逻辑回归算法的表现更佳。T.VanGestel于2003年建立了SVM算法评分模型，很大程度的提高了客户风险分析的准确性。2015年，Malekipirbazari用随机森林算法建立了风险评估模型，该模型擅长信誉较高借款人的信用评估，综合表现超越了IC和FICO等信用评估方法。总之，国外常用风险分析的算法有：SVM（支持向量机）、逻辑回归算法、决策树算法、神经网络算法等。

近年国内也大量关注风险分析的研究，并取得了一定成果。Fisher将风险分析总结为分组过程，把数据看作整体，而在整体上根据数据的不同特征值分组的方法。Thomas重点研究方向是处理数据的缺失问题，提出极大似然的方法，他的研究建立在真是有效数据基础之上，通过实验验证了该方法在提高模型的性能方面等的有效性[2]。

方兆本在信用风险评价体系的研究有着长远的意义。他提出了评价用户信用的诸多方法并且对方法的性能进行了研究和比较，常用方法例如最近邻法，回归分析法，神经网络方法等等。唐焕文等则专注于研究神经网络方法，发现神经网络方法应用于算法模型可以提高风险分析的准确率。李志辉等专注于研究商业银行信用风险建模分析，建设性选取了贷款不良率的指标应用于风险评估系统，同时结合了神经网络法，判别分析法应用于风险分析取得了显著成效。石庆焱等提出了风险分析组合模型，将神经网络方法和Logistic回归方法相结合，经过真实数据进行风险分析实验并得出结论，组合模型在稳定性上较神经网络方法有一定优势，并且组合模型的准确性较Logistic回归方法要高[3][4]。

### 1.2.2  机器学习算法优化

通过对风险分析现状的研究，发现互联网金融风险分析这一领域所提供的研究成果比较丰富，主要的风险分析算法有：SVM（支持向量机）、逻辑回归算法、决策树算法、神经网络算法等。建设风控数据分析系统首先需要从众多算法中筛选出在P2P平台客户风险分析中表现最佳的算法，验证算法的性能是否能够满足风控数据分析系统的准确度要求。还要应用训练数据集集对算法进行算法训练和算法优化，建设出性能更高的客户风险分析模型。

通过对国内外风险分析算法的研究，结合P2P平台历史数据的特点，发现目前已有算法性能还存在进一步提升的空间，算法优化过程中主要从以下两个方面入手：

1. 特征指标处理

P2P平台历史数据的客户信息，来源包括客户申请贷款时录入的基本信息，职业信息，房产信息，车辆信息，从第三方系统获取到的征信信息，电信信息，不良记录信息，法院强制执行信息等。每类信息都包括多个特征指标，如果不加选择直接应用于机器学习算法，会造成分析维度爆炸，影响机器学习算法性能。另外，每个特征指标对机器学习算法的影响程度不同，如果不加区分的对每个特征指标应用及其学习算法，会造成风险评估结果的准确性降低。

（2）空值填充误差

P2P平台历史数据的客户信息，会存在某些特征指标为空的情况。数据的预处理过程中，多数风险分析系统采取了平均值计算或者回归推理读取预测结果替代空值，给后续的分析带来了比较大的误差，需要采取优化措施，避免空值预测产生的误差。

（3）算法筛选和优化

可用于风险评估的算法较多，国内外普遍认可的存在成功案例的算法包括：SVM（支持向量机）、逻辑回归算法、决策树算法、神经网络算法等。需要根据运用P2P平台的历史数据进行算法筛选，优化，训练出符合P2P平台数据特点的高性能模型。

## 1.3 研究目的和意义

建设P2P平台风控数据分析系统的目标在于提高用户风险评估的准确性和客户贷款审核时效性。

（1）提高客户风险识别的准确性

为了提高风控数据分析系统客户风险评估的准确度，本文利用P2P平台的历史借款数据构建了客户风险评估模型。首先，从大量客户信息中选择出于客户风险密切相关的数据特征指标，进行数据清洗，数据标准化等一系列处理组成待评估的数据集[5]；然后，利用数据集对多种机器学习算法进行筛选，训练和优化；最后，训练出符合平台准确度和性能需求的机器学习算法，应用于客户风险评估模型。对新客户的风险评估和借款审核应用客户风险评估模型完成。客户风险评估模型的应用提高了用户风险分析的准确性。考虑到机器学习算法研究过程中总结的算法优化策略，主要从以下三个方面对机器学习算法做了优化：

1. 建设特征工程，完成风险评估特征指标的选择和加权处理。
2. 通过特征集分割，建设特征子集，屏蔽空特征填充造成的误差。
3. 对多种机器学习算法进行筛选，训练和优化，得到最优客户风险评估模型。

（2）缩短客户贷款审核时间

本文设计并实现风险数据分析系统，将客户贷款申请的审批过程由数据统计分析结合专家经验对客户进行多层审核，评估客户风险，升级为线上对客户的数据信息进行筛选，处理，应用风险评估模型对客户数据进行评估，得到客户风险评估结果来决定贷款申请的审核结果。系统只需要一位信审人员在线上页面进行数据处理，客户风险评估和贷款审核的处理。风控数据分析系统参与客户借款审核不但节省了人力成本，而且因为线上只需要一次风险评估就可以得出客户风险评估结果，大大缩短了用户借款审核的时间。提高了用户的借款体验满意度。依托风控数据分析系统的新客户借款审核流程如图1-2



图1- 2 新客户借款审核流程

## 1.4 论文的组织结构

本文完成了风控数据分析系统的研究和实现。以P2P平台风控数据信息系统的历史数据为依据，结合国内外先进的机器学习算法，通过算法筛选，训练和优化，构建客户风险评估模型。运用客户风险评估模型对客户的特征指标进行粉线评估，预测客户风险。为平台降低了客户违约带来的损失。风控数据分析系统替代了原有的客户风险专家逐层审核机制，提高了客户贷款审核的效率，为互联网金融公司节省了人力开支，同时提高了用户体验满意度。本文的组织结构如下：

第一章 绪论 介绍了项目背景，国内外研究现状，论文解决的问题，交代了论文组织结构。

第二章 相关技术研究，对数据挖掘理论研究，机器学习相关算法研究，层次分析获取权重方法，行业风险分析算法现状做了系统阐述。

第三章 系统需求分析和总体设计，内容包括需求分析，系统总体设计，系统数据库设计。

第四章 数据处理，介绍了数据获取，异常处理，特征指标选择，特征指标量化，特征指标权值确定。

第五章 风控数据分析设计与实现，是本次研究的核心内容，首先进行风险评估模型的实现，包括算法筛选、特征集分割、模型优化、模型评价。

然后对风控数据分析系统分模块进行设计与实现，包括数据处理模块，数据挖掘建模分析模块，数据统计分析模块。

第六章 总结与展望 总结了文章研究的相关内容，对P2P风控分析平台的研发前景和发展方向做了前瞻性分析，预测了后续发展的趋势。

# 2 相关技术研究

本章介绍了P2P平台风控分析系统在进行数据挖掘建模分析研究过程中涉及和应用到的相关理论。包括挖掘建模分析相关的定义和流程方法；机器学习以及相关算法的研究[6]；着重介绍常用机器学习算法SVM算法，逻辑回归算法，随机森林算法，介绍了算法性能评价指标。

接下来对层析分析获取特征指标的权重方法的原理和步骤进行了介绍。

## 2.1  数据分析流程

一、确定目标

数据分析的第一步是确定数据挖掘分析的目标，根据具体的用户需求和其他要求，明确数据分析的目标，只有明确分析目标，才能清楚从何处入手对什么样的目标数据进行研究分析，分析应用于什么领域，进而在相关领域中研究已有经验，为下一步的数据分析工作打下坚实基础[7]。

二、抽样

确定了数据挖掘目标后，需要从具体系统里面抽取与挖掘目标强相关的数据集合[8]。数据抽样过程中遵循的原则包括：可靠性，相关性，最新性。数据抽样中抽取的数据质量非常重要，如果抽样的数据质量不好，那想要从抽取的数据中研究分析出一般规律，得出准确结果几乎是不可能的。

三、探索数据

数据抽样之后，获得可作为样本的数据集，需要我们进一步探索数据，获取包含在数据中的有效信息，例如，数据与第一步确定的挖掘目标匹配与否，有没有实现预期效果，样本数据里所包含的大量信息在我们进行数据挖掘分析时是不是有用，假如没用我们应该怎样选择取舍，这之间有没有存在相互关系，这些都是探索数据过程中需要考虑的问题[9]。

四、数据预处理

探索数据完成后，紧接着是数据预处理这一关键的步骤。数据预处理又包括以下步骤：数据集成、数据清洗、数据变换、数据演绎等[10]。另外，我们所获取的数据，有不完整等情况存在，我们需要进行数据填充操作，并且有的数据存在重复需要我们进行去重。不同的分析维度所需要的数据指标也是不尽相同的，就需要我们进行离散处理或归一处理。具体进行何种操作则需视具体情况而定。

五、确定挖掘模式

确定数据挖掘的目标，经过了数据探索和数据预处理后，我们下一个步骤是为数据集选择合适的模式，将模式运用在数据集上。模式的选择范围包括：关联规则，聚类，分类等分析模式，寻找合适的模式是进行下一步的基础。

六、数据挖掘建模

数据挖掘建模首先需要知道我们所研究数据集的内部结构如何，清楚了数据集结构，选择出与结构相匹配的模型构建处理，这个步骤中，可以反映出数据集结构从而为以后分析的过程提供方便。模型种类多种多样，不同数据集可以匹配不同的模型。如何选择模型，选择正确模型是数据挖掘建模过程中极其重要的一步。并不是简单选择模型来完成这个步骤，而需要耐心将一个数据集在不同的模型反复进行实验，通过具体的实验数据比较，用数据结果确定哪个模型更加适合已有数据集，获得最优的模型效果。不同预测模型在这个步骤中的操作有所不同，关键的步骤有：选择模型，模型构建，模型训练，验证模型（更换数据集合），应用模型，得出预测结果[11]。

七、评价模型

对于从大量模型选择和数据结构匹配度最高，预测的准确性最好的模型，并且能让具体业务应用到选择的模型，使模型能够发挥最大应用价值，模型评价是尤为重要的。模型评价是需要新数据作为评价数据，新数据与构建模型所应用的数据集不存在任何交集，假如存在交集那么模型评价将会失去评价意义，同时可以体现我们所选择的模型在全新的数据集上的真实表现是怎样的。新数据就是测试集。到目前为止比较有效的方法包括保持法，交叉验证[12]，随机二次抽样等。

## 2.2 机器学习相关算法研究

数据建模分析是运用机器学习算法建模，对数据进行分析的过程。常用的机器学习相关算法包括：SVM算法，逻辑回归算法，随机森林算法，神经网络算法等[13]。下面我们对P2P平台风控数据分析系统相关的常见算法进行描述及介绍，在这个基础上进行之后的研究分析[14][15]。

### 2.2.1  SVM算法

（1）算法定义

SVN算法是分类算法的一种，分类问题需要进行建模，用于描述数据集[16]。分类算法分析数据集的特征，按照一定的规则将数据集进行分类[17]。

（2）SVM模型

SVM模型把已知的数据集视为复杂整体，整体中既有正值的数据集样本也有负值的数据集样本，他们混在一起[18]。我们要做的工作则是寻求一个所谓的“平面”，把正值和反值的样本分割开来，达到支持向量机模型的最终目的[19]。

（3）SVM算法原理



图2- 1 SVM算法图

对于线性可分的SVM模型就是找一个超平面如图实线所示使得每个样本点到该平面的距离最大，也就是说对于m个样本点来说超平面让最小几何间隔最大化。也就是所有样本点中离超平面最近的点到超平面距离最大化。

显然支持向量机模型的超平面一定是最靠近超平面的正实例样本点和负实例样本点到超平面具有相等的几何间距，并且最靠近超平面的样本点称为支持向量。​​求解问题的过程就是得到最优平面以及分类的决策函数的过程。

平面的公式如2-1所示：

**wx+b=0**

2-1

相应的分类函数式如2-2所示：

**f(x)=sgn(wx+b)**

2-2

假设训练集H={（x1，y1），（x2，y2），……（xn，yn）},寻找最优分割平面的问题就可以等价为二次凸优化问题求解如式2-3和2-4所示：

2-3

**s.t yi(wxi+b) -1≥0,i=1,2…n**

2-4

而线性不可分的数据集，不是所有实例都满足以上公式，如果要支持线性不可分的数据集，需要做一下改动。设存在样本点（xi,yi）不满足约束，我们引入松弛变量δi≥0使约束能够满足如公式2-5所示：

**yi(wxi+b)≥1-δi**

2-5

对每一个松弛变量需要代价δi，所以目标函数演变为式2-6：

2-6

算式中的C为代价参数，前面所述支持向量机都为线性模型。

解决线性不可分问题的另一种方法是非线性SVM。这种方法是吧数据集投射到高维的空间，这样做的目的是让数据集在新空间线性可分，这样就可以通过线性可分的SVM模型在新空间解决分类问题。假设原空间T⊂Rn，新的空间是Z，假设存在T到Z的映射如式2-7：

2-7

使所有（x,y）,有K(x,y)使得表达式如2-8

**K(x,y)=**

2-8

那么函数K(x,y)是核函数，是映射函数，这就是解决解决线性不可分问题的核函数方法。

### 2.2.2 逻辑回归算法

（1）逻辑回归算法定义

逻辑回归算法虽然算法名包含回归，但是一种用于处理分类问题的方法[20]。逻辑回归（Logistic Regression）是用于处理因变量为分类变量的回归问题，常见的是二分类或二项分布问题，也可以处理多分类问题，它实际是一种分类方法。

（2）逻辑回归算法模型

逻辑回归算法模型把输入值通过线性回归转化为预测值，通过 Sigmoid 函数得到影射值，作为x值，预测值y值是概率。该函数的取值范围为（-∞，+∞），值域为（0,1）。对于任意一个特征指标，都可通过sigmod函数将其映射到（0,1）之间的一个值，（0,1）就相当于我们数学中的概率值。函数的临界值为0.5，概率大于0.5时为一类，小于0.5时为另一类。概率接近1或0说明预测结果拟合度高，拟合度是指测试集得到的预测结果与训练集的结果与训练集得到的预测结果的符合程度。符合程度越高表示拟合度越高。

（3）逻辑回归算法原理

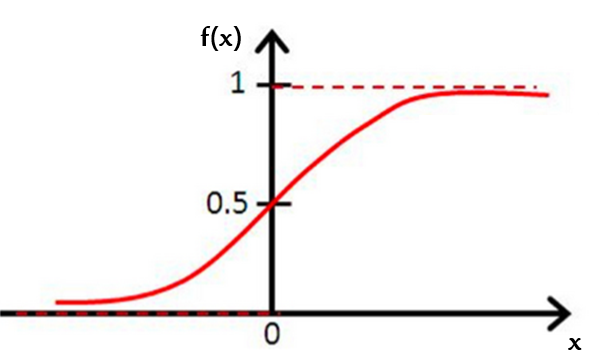


图2- 2 Sigmoid 函数图

Sigmoid 函数的表达式，δ为要求解的模型参数如式2-9：

**fδ(x)=**

2-9

逻辑回归函数采用对数似然函数，样本的输出y定义为1和0两类，对数似然函数表达式如式2-10：

**P(y=1|x)=**

**P(y=1|x)=**

2-10

其中x为特征输入，y为预测结果，对与给定的特征输入x逻辑回归函数进行优化求解，然后利用逻辑回归比较，最后根据得到的结果趋近于1或者趋近于0确定x的分类。

### 2.2.3 随机森林算法

（1）随机森林算法定义

随机森林算法是一个运用多个决策树进行分类的分类算法， 输出类别由个别决策树输出类别的众数确定。

（2）随机森林算法原理

随机森林通过建立多个决策树，将它们进行合并获得更加准确、稳定的预测。随机森林算法用bagging方法训练随机森林模型，这种方法随机并且有放回的进行训练数据的选择，用来进行分类器的构造，最后把每个决策树的效果组合得到优化的整体效果。

### 2.2.4 算法评价标准

为了得到定量数据挖掘分析结果，P2P平台风险数据分析系统会引入模型评价标准，预先设定阈值或者评价公式，用于评价数据挖掘分析结果。用于评价的主要参数包括：准确率、精准率、召回率、均方根误差等。下面我们对评价标准进行逐一介绍。

算法模型优劣通过混淆矩阵加以评价，混淆矩阵如图：

其中TP是将正例预测为正例数；TN是将负例预测为负例数；FP是将负例预测为正例数即误报；FN是将正例预测为负例数即漏报。

表2- 1 混淆矩阵

|  |  |  |
| --- | --- | --- |
| 类别 | 预测 | |
| 负例 | 正例 |
| 负例 | TN | FP |
| 正例 | FN | TP |

（1）准确度(accuracy)

准确度表示所有实例中被分类正确的实例的占比。假设实力总数为total，准确度计算公式如式2-11：

**accuracy=TN+TP/total**

2-11

（2）精准度（Precision）

精准度表示被分为正例的实例中实际的正例比例。Precision的计算公式可表示为2-12：

**Precision=TP/(TP+FP)**

2-12

（3）召回率（Recall）

召回率表示被分为正例的实例在所有真实正例中的占比。Recall的计算公式可表示为2-13：

**Recall=TP/(TP+FN)**

2-13

（4）F-measure

因为召回率和精准度会存在矛盾的情况，F-measure是基于前面描述的方法衍生出的方法，对召回率和精准度做了综合的考量。公式如2-14：

**Fα=(α2PR+PR)/ α2P+R**

2-14

公式2-15中，P是精准度，R是召回率，α为参数，α值为1，表示F1-measure:

**F1=2PR/(P+R)**

2-15

## 2.4 层次分析获取权重方法

### 2.4.1 层次分析原理

层次分析方法是一种复杂问题决策方法，适用于难于定量分析的相关问题求解。通过对多个影响特征的相对的重要性进行两相比较计算得到特征权重值。这种方法把特征进行分层，通过进行两两比较上下层特征的相对重要性，量化了主观判断作为特征权重值，这种方法相对于不加区分同等对待各个特征或者凭经验加权的做法更加科学合理，得到的结果也更加客观准确。

### 2.4.2 层次分析步骤

(1)建模

根据分析目的建立特征层次模型，根据分析统计等各种手段，确定影响评估目标的特征，建立一个由目标和特征组成的层次模型。



图2- 3 层次模型图

（2）建立特征矩阵

通过比较上下两层特征的重要程度，建立权重矩阵A,其中aij表示ai对aj的重要程度，取值范围是1~9及其倒数，1表示前者与后者相同重要，9表示前者与后者重要程度相差最大且前者重要。

**A**

2-16

（3）计算权重值

方根法获得下层特征对上层特征的相对重要性的数据序列值。1.得到各行元素的乘积。2.计算n次开方根。3.向量规范化，求近似权重值。4.计算得到最大的特征值λmax。

（4）校验一致性

一致性值计算公式CI=λmax-n/n-1 其中n是A的阶数，当A完全一致时，CI=0。λmax-n越大表示一致性越差。判断一致性是否满意计算公式:CR=CI/RI,RI是平均随机一致性的指标值。CR<0.1表示一致性符合要求。可以根据得到的特征向量值确定每个特征的权重。

## 2.5 本章小结

本章介绍了P2P平台风控分析系统研究过程中涉及的挖掘建模分析理论，包括定义，流程方法；对机器学习相关算法进行研究，并对相关的算法作了类比；针对常见的数据问题，介绍了数据预处理的具体方法。接下来对数据建模分析中风险分析领域算法应用的实际情况加以总结，得出风险分析领域相关的算法在互联网金融，电商，医疗，P2P，社交五个行业的应用情况并对其进行了归纳和总结。并根据P2P平台风控数据分析系统中数据特征集合的特点选择出来普适性强，应用最广，模型综合效果好的支持向量机算法为基础进行改进，以提高算法的性能，给后续的研究奠定基础。

# 3 基于机器学习的风险评估技术研究

本章首先分析改进现有风险控制方法的必要性，然后介绍提出的基于机器学习的风险控制技术，包括方法框架和算法。

## 基于机器学习的风险评估技术的研究动机

随着互联网技术的日新月异，互联网金融技术登上舞台，并成为一种不可或缺的革命性技术。传统银行业务已经无法满足人们日益发展的金融需求。P2P接待以其方便性、高效性受到社会的广泛关注。P2P平台给人们的生活提供极大便利，同时，面临多方面的挑战：(1) 由于线上交易，用户数量增长迅速，用户信息种类繁多，用户信用难于量化评估等诸多原因，致使客户风险识别难度增加.传统风险评估方法（风险评估专家利用经验对借款人进行评估）的准确率低，容易造成P2P平台的贷款逾期甚至引发欺诈案件；(2) 随着借款用户量不断增加，需要处理的数据量越来越大。传统方法需要大量的时间资源对借款人进行审核，无法支撑P2P平台的日益增加的借款审核需要。因此，提高风险评估的效率与准确度成为一个亟待解决的关键问题。

在智能化时代的背景下，智能算法成为驱动时代进步的动力。机器学习已成为新的边缘学科并在高校形成一门课程。它综合应用心理学、生物学和神经生理学以及数学、自动化和计算机科学形成机器学习理论基础。机器学习结合各种学习方法，取长补短的多种形式的集成学习系统研究正在兴起。各种学习方法的应用范围不断扩大，一部分已形成商品。归纳学习的知识获取工具已在诊断分类型专家系统中广泛使用。连接学习在声图文识别中占优势。分析学习已用于设计综合型专家系统。遗传算法与强化学习在工程控制中有较好的应用前景。研究者可以从大量存储的数据中识别、提取感兴趣的信息，为国家、企业和个人带来巨大的益处，具体来说：（1）算法自动从数据中获取知识，避免了面临新需要频繁改动代码逻辑的工作，节约了大量的时间；（2）

机器学习算法可以从海量的数据中提取被人们忽略或者不可能掌握的信息。

基于上述分析，机器学习技术可以有效地帮助企业从大量的历史借款数据中提取有关知识，帮主专家对借款人的信用进行评估，降低风险。

## 基于机器学习的风险评估技术 (MBRA)

本节首先介绍基于机器学习的风险评估框架，然后介绍3种风险评估算法。

### 基于机器学习的风险评估框架

机器学习是研究怎样使用计算机模拟或实现人类学习活动的科学，是人工智能中最

具智能特征，最前沿的研究领域之一。自 20世纪80年代以来，机器学习作为实现人工智能的途径，在人工智能界引起了广泛的兴趣，特别是近十几年来，机器学习领域的研究工作发展很快，它已成为人工智能的重要课题之一。机器学习不仅在基于知识的系统中得到应用，而且在自然语言理解、非单调推理、机器视觉、模式识别等许多领域也得到了广泛应用。一个系统是否具有学习能力已成为是否具有“智能”的一个标志。机器学习的研究主要分为两类研究方向：第一类是传统机器学习的研究，该类研究主要是研究学习机制，注重探索模拟人的学习机制；第二类是大数据环境下机器学习的研究，该类研究主要是研究如何有效利用信息，注重从巨量数据中获取隐藏的、有效的、可理解的知识。本文主要利用机器学习的前沿算法，收集P2P平台的借贷数据并训练模型，得到相应的模型，然后将新的借款人信息输入训练后的模型，辅助专家对风险进行评估。图3.1展示了基于机器学习的风险评估框架，主要包含3个部分：收集并预处理历史数据集，训练风险评估模型，将用户的借贷信息输入模型中得到评估结果并输出相应的风险报告。

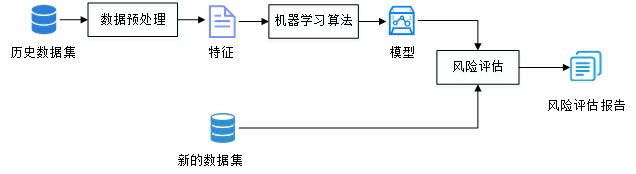


图3.1 基于机器学习的风险评估框架

1. **数据预处理：**现实世界中数据具有不完整，不一致的特点，无法直接从数据中获取或者获取不恰当的知识。通过对数据进行预处理得到规则的数据进而提高训练模型的准确度。数据预处理有多种方法：数据清理，数据集成，数据变换，数据归约等。这些数据处理技术在数据挖掘之前使用，大大提高了数据挖掘模式的质量，降低实际挖掘所需要的时间。
2. **特征提取：**特征没有精确的定义，往往由问题或者应用类型决定，反应出影响问题的关键因素。因此一个算法是否成功往往由它使用和定义的特征决定。因此特征提取最重要的一个特性是“可重复性”。本文首先考虑影响风险评估结果的关键因素，然后对得到的关键因素（特征）进行选择（特征提取），然后将对应的数据输入带训练的模型。
3. **模型训练：**该组件主要将规则的数据集输入到选择的机器学习算法中，然后选中的算法自动地从数据中学习知识。本文地模型训练过程分为两个阶段：(1)模型选择阶段，主要是筛选适合P2P借贷风险评估地算法；(2) 将处理好的数据(规则数据)输入到选择的模型中，得到决策模型。
4. **风险评估：**该组件主要将新的数据输入到训练好的模型中，并自动对其进行风险评估，并输出风险评估报告。

### 基于支持向量机的风险评估算法

|  |
| --- |
| **Algorithm 1** |
| **Input**: *S*, , *, D, M;* |
| **Output**: ; |
| 1. Obtaining regular data by preprocessing Data *S;* |
| 1. Identify and extract data features of ; 2. **for** i=1 to n 3. Set ; 4. **for** j=1 to m 5. Set 6. Training M using with c and k; 7. **end\_for** |
| 1. **end\_for** |

|  |
| --- |
| **Algorithm 2** |
| **Input**: *S*, , *,, D, M, i, j, ;* |
| **Output**: ; |
| 1. Obtaining regular data by preprocessing Data *S;* |
| 1. Identifying and extract data features of ; |
| 1. Initializing |
| 1. **for** to y 2. Set ; 3. **for** to n 4. Set ; 5. **for** to m 6. ; 7. Training M using with w, c, and k; |
| 1. **end\_for** 2. **end\_for** 3. **end\_for** |

|  |
| --- |
| **Algorithm 3** |
| **Input**: *S*, , *,, D, M, i, j, ;* |
| **Output**: ; |
| 1. Obtaining regular data by preprocessing Data *S;* |
| 1. Identifying and extract data features of ; |
| 1. Initializing |
| 1. **for** to y 2. Set ; 3. **for** to n 4. Set ; 5. **for** to m 6. ; 7. Training M using with w, c, and k; |
| 1. **end\_for** 2. **end\_for** 3. **end\_for** |

# 3 系统需求分析和总体设计

P2P平台风控数据分析系统在P2P借贷过程中起着至关重要的作用，风控数据分析系统的数据处理流程，包括数据抽取，数据预处理，数据挖掘建模分析，统计分析和结果输出等。本章介绍系统需求分析，系统总体设计，主要模块详细设计和数据库设计。

## 需求分析

系统研发基于系统的需求，系统设计的首要任务就是明确系统需求，明确系统业务需求就可以确定系统应当具备的功能，功能需求确定后也确定了系统使用的技术以及实现方式。

### 系统业务需求

P2P平台风控数据分析系统旨在通过对业务人员录入的客户信息和第三方系统获得的客户信息对借款客户的风险进行准确有效的挖掘建模分析，获得客户最终的风险评分，运用风险评分帮助借款审核人员快速高效完成借款审核。

业务流程用例如图3-1所示，风控数据分析系统的参与者主要有贷款客服，信审人员和第三方系统三种角色。其中：

1. 贷款客服：负责发起用户借款申请并向第三方系统发起获取数据的请求，客户基本信息录入保存。
2. 信审人员：负责查询待审批借款请求，根据风控数据分析系统的客户风险评分进行放款审批。
3. 第三方系统：包括银行征信等第三方系统，提供第三方系统中客户的征信信息等。



图3- 1 业务流程用例图

P2P平台风控数据分析系统的业务流程如图3-2：贷款客服给贷款客户进行贷款申请和信息录入，后台系统将客户信息和第三方获取到的客户信息数据进行入库。信审人员进行放款审批时，风控数据分析系统调用客户风险评估功能，对客户风险进行评估然后信审人员再通过客户风险评估结果给出审批结果，最终确定贷款是否获批，进入下一步借款流程。流程时序图如图3-3。



图3- 2 业务流程图



图3- 3 流程时序图

### 系统功能需求

P2P平台风控数据分析系统基于P2P平台历史数据训练风险评估模型，运用模型对增量客户进行风险评估，评估结果应用于用户贷款审核。平台需事先的功能如下：

（1）数据预处理功能：贷款客服将客户基本信息录入数据和贷款申请时获取的第三方数据组成客户风险评估的数据源，当源数据的数据量较大，那么数据中就会存在一些问题，如：数据不完整，数据噪声和存在离群点，数据不一致或数据冗余。数据预处理就是解决这些数据质量问题，完成以上列举的“脏数据”的转换、清洗以及离散化处理。另外还需对数据进行离散，归一化等处理。详细功能见表3-1：

表3- 1 数据预处理功能表

|  |  |  |
| --- | --- | --- |
|  | 功能 | 说明 |
| 数据预处理 | 数据校验 | 对数据的规范行，完整行，数据类型，数据格式进行校验 |
| 异常处理 | 对异常数据进行辨认和处理 |
| 数据标准化 | 对不同的数据类型进行标准化处理 |

（2）模型建设功能：创建风险评估模型，利用 P2P平台历史数据进行风险评估模型的筛选、训练和优化，得到理想效果的评估模型，将模型版本，配置参数等保存成模型文件，需要进行新客户的风险评估时再读取模型文件，供系统对新客户进行风险评估。

表3- 2 建设模型功能表

|  |  |  |
| --- | --- | --- |
|  | 功能 | 说明 |
| 模型建 设 | 获取历史数据 | 获取建设模型的模型训练数据 |
| 模型建设 | 训练客户风险评估模型 |
| 模型管理 | 模型文件保存和版本管理 |

（3）风险评估功能：新客户发起贷款申请时，获取客户信息并读取客户风险评估模型文件，通过风险评估模型对新数据进行风险评估，将评估结果返回。

表3- 3 风险评估功能表

|  |  |  |
| --- | --- | --- |
|  | 功能 | 说明 |
| 风险评 估 | 获取评估模型 | 获取模型文件构建风险评估模型 |
| 客户风险评估 | 应用风险评估模型对新客户进行风险评估 |
| 返回评估结果 | 将评估结果保存数据库，返回给信审人员进行放款审核 |

（4）统计分析功能：对申请贷款用户进行基本信息和用户风险评估结果进行统计分析，对借款产品和审批期限进行统计分析，对借款交易详情和借款金额等进行统计分析。

表3- 4 统计分析功能表

|  |  |  |
| --- | --- | --- |
|  | 功能 | 说明 |
| 统计分 析 | 用户分析 | 进行用户分析并保存结果 |
| 产品分析 | 进行产品分析并保存结果 |
| 交易分析 | 进行交易分析并保存结果 |

P2P平台风控数据分析系统主要的流程包括：

1. 模型创建流程如图3-4



图3- 4 模型创建流程图

1. 客户风险评估流程如图3-5



图3- 5 风险评估流程

### 系统非功能需求

（1）系统性能

支持用户数：系统需支持千级用户访问数，高峰值能够支持万级用户数。

系统响应速度：系统响应速度控制在平均2s响应，最大响应时间控制在10秒内。

（2）系统可靠性

系统运行时间：7\*24小时，系统可用性99.99%

（3）系统准确性

模型算法准确率：高于95%

模型算法召回率：高于92%

## 系统总体设计

P2P平台风控数据分析系统设计是基于比较通用的分层系统架构，这种分层架构将风控数据分析系统划分成多个不同层次，数据通过接口流转于各个层次间，系统各个模块间实现了弱耦合，这样日后升级维护成本比较低。

### 系统架构设计

P2P平台风控数据分析系统使用Spring框架研发，应用MVC模型，3-5系统架构图：



图3- 6 系统架构图

控制层： 接受用户的请求，将用户请求发送至逻辑层，并接受返回结果。

逻辑层：连接控制层和数据层，接受控制层消息分发至数据层进行数据库操作。主要用于原始数据管理，存储，挖掘建模分析，统计分析等处理过程。

数据层：负责接受上层指令并操作数据库，数据层是业务逻辑的最终实现。

数据查询层：数据查询层是对SQL的封装。

### 系统功能设计

风控数据分析平台的功能包括：数据获取，数据预处理，挖掘建模分析，统计分析。P2P平台风控数据分析系统基于P2P相关领域开源数据，对具体的数据特征进行建模分析。主要的模块包括，数据获取，数据预处理，统计分析，挖掘建模，各模块内部包含具体实现过程，后台Java开发，数据环境应用spark2.0。系统功能设计图如3-7：



图3- 7 系统功能设计图

（1）数据获取：门店客服录入的客户基本信息和第三方系统返回的客户征信相关信息组成了客户风险分析的数据源。数据获取是从数据源中选取与数据挖掘相关的特征指标，剔除与数据挖掘分析无关的内容。数据抽取遵循的原则包括3个：第一，特征指标必须做到可量化，这是进行数据分析的前提条件。第二，特征指标尽量做到全面，只有这样才能从多个角度全方位进行数据分析和数据评价。第三，特征指标需要做到线性独立，抽取的特征指标尽量做到无相关性，这样可以避免数据相关的一系列弊端。

（2）数据处理：当源数据的数据量较大，那么数据中就会存在一些问题，如：数据不完整，数据噪声和存在离群点，数据不一致或数据冗余。数据预处理就是解决这些数据质量问题，完成以上列举的“脏数据”的转换、清洗以及离散化处理。另外还需对数据进行离散，归一化等处理。

（3）挖掘建模分析：挖掘建模分析分析模块包括风险分析模型创建，模型训练，风险评估，模型管理功能。

为了定量分析数据挖掘建模分析结果，还会引入模型评价标准，预先设定阈值或者评价公式，用于对挖掘结果进行分析。常见用于评价的参数有：准确率和召回率，平均绝对误差和均方根误差。我们对算法进行了优化就需要试验验证通过这些参数来证明算法优化的效果。

（4）统计分析：基于统计学方法，设计分析主题，根据不同分析维度最终给出直观的统计结果。P2P风险数据分析平台现阶段重要分析维度包括：（1）用户分析：包括的功能有用户特征分析，用户信用评分分布等。（2）交易分析：功能包括按照申请人和审批时间分析借款处理周期等。



图3- 8 系统功能结构图

### 风险评估模型设计

通过对行业风险分析现状的研究，发现P2P平台风险分析这一领域所提供的研究成果还不足以直接应用于我们P2P平台风险分析中，需要考虑的问题包括以下两点：第一，目前的机器学习算法准确性指标不能达到系统要求。第二，数据的预处理过程中，多数风险分析系统采取了平均值计算或者回归推理读取预测结果替代空值，给后续的分析带来了比较大的误差。

我们研究了互联网金融，P2P，O2O电商，社交，医疗五大行业应用的“风险分析”相关模型算法，结合每种算法的优势和不足以及适用的场景，发现在金融风险分析领域取得显著成果的算法包括SVM算法，Logistic回归算法，随机森林算法。我们需要将已有算法应用在历史数据客户风险分析模型中进行效果比较，完成模型算法筛选。

选择出风险评估效果好的模型算法后需要进行算法优化一提高准确率和召回率，使之符合系统可靠性需求。在现有算法基础上进行改进，提高算法性能主要从两方面对算法性能进行提升。第一，不同的特征指标对客户风险评估的影响程度不同，需要通过科学方法确定特征指标权重。第二，采用特征集分割方法规避预测结果替代空值造成的误差。



图3- 9 风险评估模型构建过程

经过筛选，优化训练后的算法模型以模型文件形式保存到数据库，当有新用户数据集需要进行风险评估时，只需读取模型文件对新的数据即进行风险评估将评估结果返回给信审人员即可。风险评估模型的构建过程如图3-9。

## 系统数据库设计

P2P平台风控数据分析系统的实体包括：风险评估模型，评估结果，待评估数据，数据源，模型评价结果。风险评估模型是我们训练得到的风险评估模型；数据源是从P2P平台录入的客户信息和第三方系统返回的征信相关信息；待评估数据是可以直接进行模型评估的数据；评估结果是模型对新的用户数据进行风险评估的结果；模型评价结果是训练阶段对模型进行评价的结果。实体关系图如图3-10所示：



图3- 10 实体关系图

我们把处理好的数据文件作为输入源，数据库仅用于存储结果自动化，这样可以充分利用分布式环境并行计算能力。我们运用Mysql数据库，优点是：免费，开源，速度快，安全性比较高，支持事务。本系统查询复杂度并不是特别高，应用环境比较适合Mysql数据库。主要数据库表设计如下：

表3- 5 主要数据库表

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 文件名 | 存储形式 | 备注 |
| 1 | risk\_ model | mysql | 风险评估模型表 |
| 2 | risk\_ evaluate | mysql | 模型评价结果表 |
| 3 | risk\_ result | mysql | 风险评估结果表 |
| 4 | risk\_ data | mysql | 待评估数据表 |
| 5 | loan\_customer | mysql | 借款客户表 |
| 6 | credit\_info | mysql | 第三方征信信息表 |
| 7 | house\_info | mysql | 房产信息 |
| 8 | car\_info | mysql | 车辆信息 |
| 9 | company\_ info | mysql | 单位信息 |
| 10 | analysis\_customer | mysql | 用户分析记录表 |
| 11 | analysis\_loan | mysql | 交易分析记录表 |



图3- 11 风控数据分析系统数据库设计

## 本章小结

本章首先对系统的业务需求和功能需求做了描述。明确了需求之后，又对系统的数据分析流程做了阐述，完成了系统的总体设计。着重介绍了P2P平台风险数据分析，数据建模分析模块设计。最后介绍了P2P平台风险数据分析系统的数据库设计。

# 4 数据处理

数据处理部分主要解决数据挖掘建模分析前的数据源获取和数据预处理工作，经过数据处理得到待评估数据。数据处理分为三部分实现：数据源阶段获取P2P平台门店客服录入的客户基本信息以及第三方系统的客户征信信息；数据预处理阶段进行数据清理和数据转换；特征指标选取加权完成特征选择和特征q值确定。

## 4.1 数据获取

数据源阶段获取P2P平台门店客服录入的客户基本信息以及第三方系统的客户征信信息后将数据返回风控数据分析系统，组成风控数据分析系统的原始数据主要有一下几类：个人信息，房产信息，职业信息，第三方征信信息，车辆信息等。数据源的数据组成详情如表4-1：

表4- 1 数据源组成

|  |  |
| --- | --- |
| 类型 | 详情 |
| 个人信息 | 姓名、性别、年龄、配偶信息、学历等 |
| 房产信息 | 房屋数量、地址、面积、是否贷款、房产价值、产权人等 |
| 职业信息 | 单位性质、行业、职位、收入、工作年限等 |
| 第三方征信信息 | 银行征信信息、有无逾期记录、信用卡信息、资产处置记录、电信记录、工商企业信息、不良记录等 |
| 车辆信息 | 车辆数量、是否贷款、车辆价值等 |
| 借款信息 | 借款额、借款产品、申请时间、还款方式等 |

在数据源获取阶段，返回的数据中并不是所有数据都与贷款的审批有直接关系，其中例如借款信息、手机号码、地址等是可以人为判断对客户风险评估结果无影响的信息，可以把这类信息剔除。保留与判断客户风险性有相关性的信息。

## 数据处理

### 4.2.1 数据抽取

数据抽取是从数据源中选取与数据挖掘相关的特征指标，剔除与数据挖掘分析无关的内容。数据抽取遵循的原则包括3个：第一，特征指标必须做到可量化，这是进行数据分析的前提条件。第二，特征指标尽量做到全面，只有这样才能从多个角度全方位进行挖掘建模分析，保证分析结果准确性。第三，特征指标需要做到线性独立，抽取的指标尽量做到无相关性，这样可以避免数据相关的一系列弊端。

初步抽取到的特征如表4-2：

表4- 2 初步抽取特征指标集合

|  |  |  |  |
| --- | --- | --- | --- |
| 特征指标字段 | 特征指标含义 | 特征指标字段 | 特征指标含义 |
| amount | 借款金额 | car\_num | 车辆数 |
| age | 年龄 | car \_value | 车辆价值 |
| education | 学历 | loan\_num | 贷款账户数 |
| company\_type | 单位性质 | over\_times | 逾期次数 |
| wage | 月薪 | card\_num | 信用卡账户数 |
| house\_num | 房产数 | other\_wage | 其他收入 |
| house\_value | 房产价值 | have\_loan | 贷款额度 |
| phone\_history | 电信记录 | enforcement | 强制执行记录 |

### 4.2.2 异常处理

由于人工录入失误或者数据但未变化等原因会造成异常值的存在。识别异常值的方法包括：标准差法，连续值异常值的识别是把距离其均值标准差2倍以上的数据定义为异常值；聚类法，离散值异常值的识别是采用聚类的方法，如果数值没有落在类里面就定义为异常值。

我们以借款金额为例描述去异常的过程：对借款金额进行异常值分析，P2P平台的借款用户借款金额是散列值，可以尝试用聚类法识别异常值，现在对数据源数据利用数据处理工具以数据源编号为横坐标以借款金额为纵坐标得到借款金额的散点分布图4-1，可以总结借款金额数据类的特点是金额在100000到500000范内。

图4- 1 借款金额分布图

数值落在借款范围外的两条数据（1412，990000），（1617，500）为异常数据，借款金额异常需要将数据删除。删除异常数据后，得到去异常借款金额散点分布图4-2

图4- 2 去异常借款金额散点分布图

## 特征指标工程

### 4.3.1 特征指标量化

有些特征指标为枚举类型，包含有限的选项。在将这些特征指标进行量化过程中，我们选取了独热编码。独热编码即 One-Hot 编码，又称一位有效编码，其方法是使用N位[状态寄存器](https://baike.baidu.com/item/%E7%8A%B6%E6%80%81%E5%AF%84%E5%AD%98%E5%99%A8)来对N个状态进行编码，并且在任意时候，其中只有一位有效。这样的他正指标包括：学历，单位性质，年龄段等。量化结果如表4-3学历编码表，表4-4 单位性质编码表

表4- 3 学历编码表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高中及以下 | 本科 | 研究生 | 博士及以上 |
| 数据集1 | 1 | 0 | 0 | 0 |
| 数据集2 | 0 | 0 | 1 | 0 |
| 数据集3 | 0 | 1 | 0 | 0 |

其中数据集数据集1学历编码1000代表高中及以下，数据集2学历编码0010代表研究生，数据集3学历编码0100代表本科。

表4- 4 单位性质编码表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 国企 | 事业单位 | 私营 | 外企 | 其他 |
| 数据集1 | 1 | 0 | 0 | 0 | 0 |
| 数据集2 | 0 | 0 | 1 | 0 | 0 |
| 数据集3 | 0 | 1 | 0 | 0 | 0 |

### 4.3.2 特征指标标准化

数据预处理时，还会有对数据进行标准化的操作，即对原始数据进行缩放。不同属性值的范围是不同的，在计算时值域较大属性可能会影响计算结果，使得其他因素的影响被弱化而使计算结果偏差较大。所以标准化步骤是不可或缺的，标准化方法有一下几种：

（1）极差法是将原始数据值映射[0,1]之间，这种方法能消除属性单位不同给结果带来的影响，保证属性对结果产生的影响均等。计算公式4-1：

**A\*=A-MIN/MAX-MIN**

4-1

（2）Z-SCORE法是利用数据的平均值和标准差进行标准化。经过Z-SCORE法处理后数据大约有一半为正，一半为负，并且符合标准的正态分布。计算公式：

**A\*=A-  https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s=13/sign=fc592cb0738b4710ca2ff9cfc1ce63c9/63d9f2d3572c11dfef146bb2622762d0f603c25e.jpg/  https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s=10/sign=58e9702ea1ec08fa220017a758ee9a39/9c16fdfaaf51f3de1b656df595eef01f3b297948.jpg （其中https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s=13/sign=fc592cb0738b4710ca2ff9cfc1ce63c9/63d9f2d3572c11dfef146bb2622762d0f603c25e.jpg是均值 https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s=10/sign=58e9702ea1ec08fa220017a758ee9a39/9c16fdfaaf51f3de1b656df595eef01f3b297948.jpg 是标准差）**

4-2

我们采用Z-SCORE法对借款金额，房产价值，车辆价值，逾期账户数，薪资，其他收入等进行标准化。标准化前后的数据如表4-5和表4-6所示

表4- 5 特征指标标准化前

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | amount | car \_value | house\_value | wage | over\_times | have\_loan |
| 数据集1 | 400000 | 78 | 138 | 25000 | 0 | 200 |
| 数据集2 | 500000 | 138 | 145 | 18000 | 0 | 300 |
| 数据集3 | 300000 | 145 | 78 | 45000 | 9 | 200 |
| 数据集4 | 200000 | 106 | 148 | 100000 | 6 | 400 |
| 数据集5 | 200000 | 148 | 106 | 180000 | 3 | 500 |

表4- 6 特征指标标准化后

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | amount | car \_value | house\_value | wage | over\_times | have\_loan |
| 数据集1 | 0.6896 | -1.7236 | 0.5745 | -0.2061 | -0.9272 | -0.9718 |
| 数据集2 | 1.5517 | 0.5745 | 0.8426 | -0.2842 | -0.9272 | -0.1724 |
| 数据集3 | -0.1724 | 0.8426 | -1.7236 | 0.0168 | 0.5417 | -0.9718 |
| 数据集4 | -0.9718 | -0.2681 | 0.9576 | 0.6298 | 0.4378 | 0.6896 |
| 数据集5 | -0.9718 | 0.9576 | -0.2681 | -0.2842 | 0.0517 | 1.5517 |

### 4.3.3 特征指标选择权值确定

经过初步筛选后的特征指标数量仍然比较多，需要分析的指标数量越多，机器学习算法的处理效率就越低。如果不加区分同等对待各个特征或者凭经验对特征进行加权，很容易造成结果失准。这就需要我们选择更加科学合理的加权算法进行特征的加权量化。进行特征指标选择和权值确定是进行数据挖掘建模的基础，首先从大量的特征指标中选择出于客户扶风县最密切相关的指标，能够最大程度的缩减指标数量，从而有效避免面维数灾难问题。其次，去除与客户风险相关性小的指标，也能降低机器学习算法处理难度，提高数据处理效率。

常用的特征选择方法有专家调查表法，敏感性分析法和层次分析法。其中专家调查表法是建立在专家调查定性分析的基础上；而敏感性分析法是把最为敏感的因素作为概率分布的关键变量，这两种方法显然不适合我们用于用户风险分析的特征集选择方法。P2P借贷平台有大量的用户借款记录，从可靠性方面考虑：这些历史数据都是真实的借款数据，可靠性有保障；从相关性考虑，我们选取借款数据中对用户的风险评估存在影响的用户信息作为特征集合；从最新性考虑，我们随机选取近一年的一万份数据进行分析，保证了分析数据的最新性。因此我们最终选用层次分析法来进行特征分析和选择。对选取的2000份借款数据进行资料分析，把数据抽取得到的特征指标分为三大类如表4-7所示：

表4- 7 初步抽取特征指标

|  |  |  |
| --- | --- | --- |
| 目标层 | 特征类层 | 特征指标层 |
| 风险评估 | 个人信息 | age |
| amount |
| education |
| 职业信息 | company\_type |
| wage |
| other\_wage |
| 资产信用信息 | house\_value |
| house\_num |
| car \_value |
| car\_num |
| have\_loan |
| over\_times |
| loan\_num |
| card\_num |
| phone\_history |
| enforcement |

根据选择的第一层特征包括a1（个人信息），a2（职业信息），a3（资产信用信息）。构造q值判断矩阵4-3：

**A**

4-3

采用第二章层次分析获取权重方法中介绍的方根法计算出矩阵的λmax=3.008，特征向量的值（0.075，0.333，0.592），一致性值CR=0.014<0.1，一致性符合要求。因此可以将特征向量结果作为特征类的权重。运用相同方法对每个特征类构造q值判断矩阵，得到每个特征指标的q值如图4-3所示。

图4- 3 特征指标权重

我们从特征指标的权重计算结果选取权重值大于0.02的特征指标作为客户风险分析的特征集合，组成客户风险挖掘建模分析的待评估数据。最终选择的特征指标和权重q值如表4-8

加权待测试数据集

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | amount | education | company\_type | car\_value | house\_value | wage | over\_times | have\_loan |
| q值 | 0.0290 | 0.0444 | 0.0833 | 0.1395 | 0.2861 | 0.2498 | 0.0868 | 0.0395 |
| 数  据  集 | 0.6896 | 0010 | 00100 | -1.7236 | 0.5745 | -0.2061 | -0.9272 | -0.9718 |
| 1.5517 | 0001 | 00010 | 0.5745 | 0.8426 | -0.2842 | -0.9272 | -0.1724 |
| -0.1724 | 1000 | 10000 | 0.8426 | -1.7236 | 0.0168 | 0.5417 | -0.9718 |
| -0.9718 | 0100 | 01000 | -0.2681 | 0.9576 | 0.6298 | 0.4378 | 0.6896 |
| -0.9718 | 0010 | 00100 | 0.9576 | -0.2681 | -0.2842 | 0.0517 | 1.5517 |

## 本章小结

本章对数据处理的实现做了详细描述。明确了数据源的获取，数据预处理，特征指标量化，选择和加权的一般方法和过程。着重介绍了特征指标选择加权形成待评估数据的处理过程。数据处理是从数据源获取待测试数据的过程，为数据挖掘建模分析模块提供了数据基础。

# 5 风控数据分析设计与实现

## 5.1 风险评估模型实现

### 5.1.1 算法筛选

我们首先调查了互联网金融，P2P，O2O电商，社交，医疗五大行业应用的“风险分析”相关模型，对其中涉及到的算法新行了统计和分析：

表5- 1 五大行业风险分析模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型算法 | 互金 | P2P | O2O | 社交 | 医疗 |
| 支持向量机算法 | 是 | 是 | 是 | 是 | 是 |
| 决策树算法 | 是 | 否 | 是 | 是 | 否 |
| 神经网络算法 | 是 | 否 | 否 | 是 | 否 |
| 遗传算法 | 否 | 是 | 否 | 是 | 否 |
| Logistic回归算法 | 是 | 是 | 否 | 是 | 否 |
| Naïve Bayes算法 | 否 | 否 | 否 | 是 | 是 |
| 其他算法 | 否 | 是 | 是 | 否 | 否 |

统计结果显示，支持向量机算法，决策树算法，Logistic回归算法是互联网金融和P2P领域中的主流算法，并且在其他领域也有广泛应用[21]。其中支持向量机算法的优势是适合小样本；用于非线性问题；不存在局部极小值问题（相对神经网络等其他算法）；善于处理高维数据集；泛化能力强；不足在于缺失数据敏感；于二分类领域；文本分类，图像识别等场景[22]。决策树算法的优势在于易于解释理解，容易提取规则；可同时处理标称和数值型数据，运行速度快；便于数据库扩展，独立于数据库大小；不足在于缺失数据处理困难；容易出现过拟合；忽略数据集属性关联；主要应用于企业投资决策；企业管理实践[23]。逻辑回归算法的优势在于计算代价不高，易于理解实现；不足在于容易产生欠拟合，分类精度不高，主要应用于根据分类概率排名的领域，如搜索排名等[24]。

算法筛选过程中，我们将目前互联网金融领域的优秀算法应用于训练数据集，通过调整算法的各项参数使算法达到最优状态，对各种算法进行比较筛选出应用于风险评估模型的基础算法。

（1）SVM算法

对待评估数据集应用SVM算法时，需要考虑SVM算法的两个核心参数：惩罚系数和核函数。惩罚系数C的取值包括0.1，0.2，0.5，1，2，4，10，核函数kernel的类型包括让RBF, Linear, Poly。对待评估数据集应用SVM算法过程中要遍历所有参数组合，找到最佳参数组合并返回评估结果。流程如图5-1



图5- 1 SVM算法训练流程

参数遍历完成，得到算法最终评价结果，结果集如表5-2

表5- 2 SVM算法训练结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | kernel | C | accuracy | Recall |
| 1 | RBF | 10 | 0.8725 | 0.9298 |
| 2 | RBF | 4 | 0.8850 | 0.9206 |
| **3** | **RBF** | **2** | **0.9001** | **0.9068** |
| 4 | RBF | 1 | 0.9102 | 0.8953 |
| 5 | RBF | 0.5 | 0.9129 | 0.8856 |
| 6 | RBF | 0.2 | 0.9120 | 0.8657 |
| 7 | RBF | 0.1 | 0.9149 | 0.8510 |
| 8 | Linear | 10 | 0.8454 | 0.6123 |
| 9 | Linear | 4 | 0.8454 | 0.6123 |
| 10 | Linear | 2 | 0.8454 | 0.6123 |
| 11 | Linear | 1 | 0.8454 | 0.6123 |
| 12 | Linear | 0.5 | 0.8454 | 0.6123 |
| 13 | Linear | 0.2 | 0.8454 | 0.6123 |
| 14 | Linear | 0.1 | 0.8454 | 0.6123 |
| 15 | Poly | 10 | 0.9436 | 0.6921 |
| 16 | Poly | 4 | 0.9445 | 0.6898 |
| 17 | Poly | 2 | 0.9438 | 0.6838 |
| 18 | Poly | 1 | 0.9435 | 0.6456 |
| 19 | Poly | 0.5 | 0.9440 | 0.6668 |
| 20 | Poly | 0.2 | 0.9442 | 0.6356 |
| 21 | Poly | 0.1 | 0.9446 | 0.6359 |

由实验结果可知在核函数维度进行比较可以明显看出，Linear的效果无论在准确率还是召回率表现都不理想，泛化程度低；Poly核函数似然有较高的准确度，但召回率明显偏低；RBF核函数相对其他核函数而言准确率和召回率表现都比较突出。综合考虑算法准确率和召回率两方面因素，选中SVM算法参数组合C=2，kernel= RBF，得到Precision=0.9001，Recall=0.9068。

（2）Logistic回归算法

对待评估数据集应用Logistic回归算法时，需要考虑Logistic回归算法的两个核心参数：分类权重和惩罚系数C。惩罚系数C是正则因数lambda的倒数，随着正则因数的增大，模型变简单，泛化能力变弱。反之模型变复杂泛化能力变强。惩罚系数C的取值包括0.1，0.2，0.5，1，2，4，10，核函数kernel的类型包括让RBF, Linear, Poly。对待评估数据集应用Logistic回归算法过程中要遍历所有参数组合，找到最佳参数组合并返回评估结果。流程如图5-3



图5- 2 逻辑回归算法训练流程

参数遍历完成，得到算法最终评价结果，结果集如表5-3

表5- 3 逻辑回归算法训练结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | class\_weight | C | accuracy | Recall |
| 1 | 1.5 | 10 | 0.9224 | 0.7096 |
| 2 | 1.5 | 4 | 0.9252 | 0.7092 |
| 3 | 1.5 | 2 | 0.9314 | 0.7087 |
| 4 | 1.5 | 1 | 0.9342 | 0.7082 |
| 5 | 1.5 | 0.5 | 0.9365 | 0.7065 |
| 6 | 1.5 | 0.2 | 0.9374 | 0.7052 |
| 7 | 1.5 | 0.1 | 0.9380 | 0.7046 |
| 8 | 2.2 | 10 | 0.9382 | 0.7059 |
| 9 | 2.2 | 4 | 0.9383 | 0.7059 |
| 10 | 2.2 | 2 | 0.9383 | 0.7059 |
| 11 | 2.2 | 1 | 0.9383 | 0.7059 |
| 12 | 2.2 | 0.5 | 0.9385 | 0.7059 |
| 13 | 2.2 | 0.2 | 0.9380 | 0.7054 |
| 14 | 2.2 | 0.1 | 0.9383 | 0.6962 |
| 15 | 2.4 | 10 | 0.8686 | 0.8713 |
| 16 | 2.4 | 4 | 0.8686 | 0.8713 |
| **17** | **2.4** | **2** | **0.8986** | **0.8854** |
| 18 | 2.4 | 1 | 0.8686 | 0.8697 |
| 19 | 2.4 | 0.5 | 0.8346 | 0.8713 |
| 20 | 2.4 | 0.2 | 0.8634 | 0.8784 |
| 21 | 2.4 | 0.1 | 0.8460 | 0.8659 |

选中Logistic回归算法参数组合class\_weight =2.4，C=2，得到Precision=0.8986，Recall=0.8854。

（3）随机森林算法

对待评估数据集应用随机森林算法时，需要考虑随机森林算法的两个核心参数：最大深度max\_depth，最大特征数max\_features。需要注意最大深度max\_depth取值越大时，模型性能随之提高，但是当max\_depth数量增大到一定程度后，模型性能的提高就变得很有限。max\_depth取值包括10，50，100，200，max\_features取值4，5，6，7，8。对待评估数据集应用随机森林算法过程中要遍历所有参数组合，找到最佳参数组合并返回评估结果。流程如图5-3



图5- 3 随机森林算法训练流程

参数遍历完成，得到算法最终评价结果，结果集如表5-4

表5- 4 随机森林算法训练结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | max\_depth | max\_features | accuracy | Recall |
| 1 | 10 | 4 | 0.9224 | 0.7096 |
| 2 | 10 | 5 | 0.9252 | 0.7092 |
| 3 | 10 | 6 | 0.9314 | 0.7087 |
| 4 | 10 | 7 | 0.9342 | 0.7082 |
| 5 | 10 | 8 | 0.9365 | 0.7065 |
| 6 | 20 | 4 | 0.9374 | 0.7052 |
| 7 | 20 | 5 | 0.9380 | 0.7046 |
| 8 | 20 | 6 | 0.9382 | 0.7059 |
| 9 | 20 | 7 | 0.9383 | 0.7059 |
| 10 | 20 | 8 | 0.9383 | 0.7059 |
| **11** | **50** | **4** | **0.9400** | **0.8022** |
| 12 | 50 | 5 | 0.9400 | 0.8003 |
| 13 | 50 | 6 | 0.9394 | 0.7985 |
| 14 | 50 | 7 | 0.9391 | 0.7989 |
| 15 | 50 | 8 | 0.9386 | 0.7989 |
| 16 | 100 | 4 | 0.8686 | 0.8713 |
| 17 | 100 | 5 | 0.8986 | 0.8854 |
| 18 | 100 | 6 | 0.8686 | 0.8697 |
| 19 | 100 | 7 | 0.8346 | 0.8713 |
| 20 | 100 | 8 | 0.8634 | 0.8784 |

选中随机森林算法参数组合max\_depth=50，max\_features =4，得到Precision=0.9400，Recall=0.8022。

统计算法筛选结果如表5-5，可以得出结论SVM算法综合表现最佳，所以后续风险评估模型在SVM算法基础上进行优化和训练。

表5- 5 算法筛选结果

|  |  |  |
| --- | --- | --- |
| 算法名称 | accuracy | Recall |
| SVM算法 | 0.9001 | 0.9068 |
| Logistic回归算法 | 0.8986 | 0.8854 |
| 随机森林算法 | 0.9400 | 0.8022 |

单独运用SVM算法进行客户风险评估，从得到的算法筛选结果中可以看出算法的准确度和召回率不足以满足系统性能需求，考录从以下两方面对算法进行改进。分析算法应用过程中造成算法性能降低的原因主要有:并没有根据算法特征对客户风险的影响程度大小评估客户风险，而是不加区分的对待所有特征指标；在处理空特征时，直接运用了业界普遍采用的空特征填充方法，取平均值填充，这样降低了风险评估的准确性。针对以上原因我们从以下两个方面对算法进行改进，优化风险评估模型：第一对包含特征值为空值的特征集合先进行特征集分割，然后对分割后的N个子特征集合循环使用SVM算法进行客户风险评估；第二，风险评估过程中对特征指标进行加权计算。

### 5.1.2 特征集分割

表5- 6 待分割特征指标集

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | amount | education | comp\_type | car\_val | house\_val | wage | over\_times | have\_loan |
| 1 | 0.6896 | 0010 | 00100 | -1.7236 | 0.5745 | -0.2061 | -0.9272 | -0.9718 |
| 2 | 1.5517 | 0001 | 00010 | Null | 0.8426 | -0.2842 | -0.9272 | -0.1724 |
| 3 | -0.1724 | 1000 | 10000 | 0.8426 | -1.7236 | 0.0168 | Null | -0.9718 |
| 4 | -0.9718 | 0100 | 01000 | -0.2681 | 0.9576 | Null | 0.4378 | 0.6896 |
| 5 | -0.9718 | 0010 | 00100 | 0.9576 | -0.2681 | -0.2842 | 0.0517 | 1.5517 |

建立模型的第一步，通过给每个特征集选择合适的特征子集来消除空值的影响，提高风险分析的准确性。目前业界普遍采取计算方法得到平均值，或者逻辑推理方法填充空缺值，这样的做法会给后续的风险分析带来原始误差。针对这个问题，我们在构建特征集的时候把空值保留下来。通过屏蔽相同空特征的数据实例，建立合适的特征子集，构建混合模型，不仅解决了空特征的问题，而且相对业界普遍采取的空特征填充方法，提高了风险分析的准确率。在表5-5中的数据集实例里，我们用Null表示缺省值，空白表示正常特征。通过前述构建合适的特征集自己的方法，我们可以构建如下子集：

1、由所有特征和编号1，5组成的数据集

2、由特征amount，education，comp\_type，house\_val，have\_loan和编号2，3，4组成的数据集

3、由特征car\_val，wage和编号3组成的数据集

4、由特征wage，over\_times和编号2组成的数据集

5、由特征car\_val，over\_times和编号4组成的数据集

创建个体分类子集的方法，加入有k个特征和m个数据集，可得H=k╳m其中Hij表示数据集i特征j，矩阵如下式5-1：其中0表示空值，1表示特征值。可以通过矩阵子集划分的方法获取到N个子集的SVM学习机。例如上面的待分割特征集表达式如式5-1所示，我们就划分得到了5个SVM学习机划分后每个学习机的表达式如5-2所示。

H=

5-1

5-2

### 5.1.3 模型优化

经过特征集分割，得到N个SVM学习机，数据处理过程中，我们得到了每个特征指标的权重，那么可以对每个SVM学习机循环运用SVM算法直至达到客户防线评估最终结果。根据这个思路我们得到了优化后的加权SVM算法模型，模型的具体实现步骤如图5-4。



图5- 4 加权SVM算法模型

对每个用户结果集根据q值加权算法计算得到的L(i)获得预测结果通过与边界值A.B比较，得出风险分析结论，小于A为低风险，大于B为高风险。



上述过程优化后的算法可以用伪代码表示如下

|  |
| --- |
| **Requlre: The decision threshold A and B;**  **Ensure: D∈ Good,Bad;**  **1:Calculate the test statistic Li by the SVM learner;**  **2:Get the decision D;**  **3:if B<=Li then**  **4: H1<- D,declaring prediction result is Good;break;**  **5:else**  **6: if Li <=A then**  **7: H0<- D,declaring prediction result is Bad;break;**  **8: else**  **9: Accept new value obtained by next SVM learner,go to step1 and update Li;**  **10: end if**  **11:end if** |

运用计算边界值A和B的公式，根据我们期望的正确率计算获得A.B公式如5-2：

**A=log(1-TP/1-TN) & B= log(TP/TN)**

5-2

首先，得到第一个SVM（支持向量机）学习机的风险评估值L1,如果B<L1<A那么混合模型不能做出判断。然后，通过下一个学习机更新L(i)表示第i个学习机获得的评估值。我们通过q值加权算法和聚合已有学习机获得L(i)， SVM算法求平面问题演变为求加权平均问题，根据计算公式5-3：

**Li=wl=w0l0+ w1l1+...+ wklk**

5-3

其中权重w的计算公式5-4：

**wk=qk/Qi**

5-4

这样检验统计量的迭代方法可根据下面公式5-5进行，可显著降低算法复杂度：

**Li+1 = (l0+ ql1+...+ qi+1li+1)/Qi+1 = (Qi /Qi+1) Li+ ( qi+1li+1)/Qi+1**

5-5

将优化后的加权SVM算法模型应用于历史数据集，选中SVM算法参数组合C=2，kernel= RBF，得到Precision=0.9503，Recall=0.9368。流程如下图5-5



图5- 5 SVM算法优化流程

通过优化加权SVM算法模型得到的客户风险分析评价的结果，与进行算法筛选时所得到的基础算法的评价结果进行综合比较，可以得出结论，优化加权SVM算法模型的性能的到了提高，同时也满足了系统性能方面的需求。训练集模型评价结果如表5-7

表5- 7 训练集模型评价结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型版本 | 模型算法 | Precision | Recall |
| V1.0.1 | SVM算法 | 0.9001 | 0.9068 |
| V1.0.2 | Logistic回归算法 | 0.8986 | 0.8854 |
| V1.0.3 | 随机森林算法 | 0.9601 | 0.8004 |
| V2.0.1 | 优化加权SVM算法 | 0.9863 | 0.9422 |

### 5.1.3 模型评价

模型评价是根据模型产生的客户评估结果对模型性能进行评估的过程，我们选取了精准度和召回率两个指标对算法模型预测结果进行评价。采用的评价方法是交叉验证法。交叉验证的基本思想将原始数据进行分组,一部分为训练集(t\_set),另一部分为验证集(v\_set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标。

本文采用10折交叉验证，原始数据分割成10个子样本，保留一个子样本作为验证集，其他9个样本模型训练。重复10次交叉验证，对每个[子样本](https://baike.baidu.com/item/%E5%AD%90%E6%A0%B7%E6%9C%AC)进行验证，平均10次的结果，最终得到评估值。这个方法可以重复运用随机产生的子样本进行训练和验证。模型评价流程如图5-6所示



图5- 6 模型评价流程图

经过交叉验证得到模型评价结果如表5-7所示

表5- 8 训练集模型评价结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型版本 | 模型算法 | accuracy | Recall |
| V1.0.1 | SVM算法 | 0.9003 | 0.9028 |
| V1.0.2 | Logistic回归算法 | 0.8974 | 0.8798 |
| V1.0.3 | 随机森林算法 | 0.9524 | 0.8132 |
| V2.0.1 | 优化加权SVM算法 | 0.9834 | 0.9396 |

## 5.2 风控数据分析设计与实现

### 5.2.1 数据处理模块设计与实现

数据处理模块是由源数据经过数据处理得到待评估模型数据的过程。其中涉及到的类包括源数据类，数据处理接口类DataProInterface，数据处理实现类DataProImpl，数据异常类，数据处理模块类图设计如图5-7



图5- 7 数据处理模块类图

（1）DataProInterface：数据处理接口

（2）DataProImpl:数据处理实现类

1)collectData（String customerId）：数据获取方法，返回源数据。

2)checkData（SourceData data）：数据校验方法，返回布尔类型。

3)proData（SourceData data）：数据处理方法，返回待风险评估的模型数据。

（3）SourceData：源数据类。

（4）ModData：待评估模型数据类。

（5）NullValueException：必填值为空异常类。

（6）TypeNotMatchException：数据类型不匹配异常类。

数据处理的流程包括数据获取，异常处理，特征指标量化，标准化，特征选择等一系列过程。数据处理流程如图5-8



图5- 8 数据处理流程图

数据获取是从借款客户表，第三方征信信息表，房产信息，车辆信息，单位信息表中获取源数据。包括个人信息，职业信息和资产信用信息三个类别。数据库与客户风险评估相关性较小的会在数据筛选过程中剔除。如图5-9所示



图5- 9 数据源数据

异常处理的输入为数据获取得到的数据源，对数据源中数据类型进行校验，如果类型不匹配则接口返回类型不匹配异常，如果必填项目为空则接口返回类型不匹配异常，用户页面提示如图5-10，如果存在异常值不影响其他数据校验则将异常数据剔除。

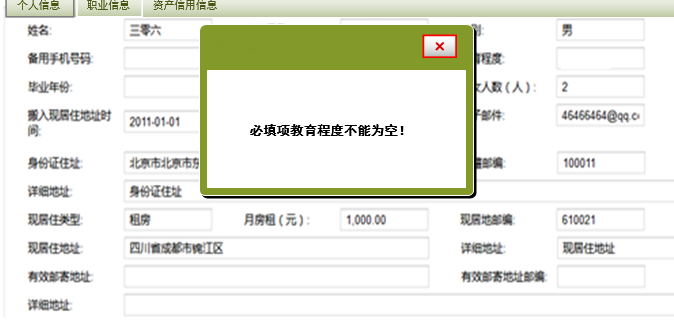


图5- 10 用户页面提示

数据标准化的过程是获得异常处理后的数据，对数据中的特定特征指标运用数据标准化公式，计算得到标准化数据。经过数据处理后得到待评估的模型数据，类型ModData。得到数据处理结果如图5-11



图5- 11 数据处理结

### 5.2.2 数据挖掘建模分析模块设计与实现

数据挖掘建模分析模块主要功能是模型训练和客户风险评估。其中类之间的关系图如图5-12



图5- 12 数据挖掘建模分析模块类图

（1）DataModAnalysis:数据挖掘建模分析接口。

（2）DataModAnalysisImpl：数据挖掘建模分析实现类。

trainMod（）：模型训练接口，训练算法模型并将模型文件和模型版本号入库。

loadMod（String version）：模型加载接口，根据模型版本号读取模型文件并完成模型创建。

modAnalysis（String cusId）：客户风险评估接口，根据客户id读取客户的待风险评估模型数据并运用选择的模型版本进行风险评估，返回风险评估结果。

（3）RiskModle：训练得到的算法模型，用模型版本号进行区分。

（4）ModData：待评估数据。

（5）ModLoadFieldException：模型加载失败异常。

（6）AnalysisFieldException：风险评估失败异常。

风险评估模型训练的过程在本章第一节已经做了详细介绍。对客户进行风险评估的过程是首先输入客户编号和模型的版本号。然后进行数据处理获取到待评估模型数据，根据模型版本加载模型文件并完成模型的构建，最后运用构建的模型对待评估数据进行风险评估得到客户放线评估结果。风险评估流程页面如图5-13，风险评估结果页面如图5-14。



图5- 13 风险评估流程



图5- 14 风险评估结果

### 5.2.3 数据统计分析模块设计与实现

数据统计分析模块主要实现用户分析和交易审批时间分析。其中类之间的关系图如图5-15

（1）DataStatistic：数据统计分析接口。

（2）DataStatisticImpl：数据统计分析实现类。

UserStatistic（）：用户分析方法，分析时间段内高风险用户和低风险用户占比。

LoanStatistic（）：贷款审批时间分析，分析时间段内贷款的平均审批时间。

（3）StatisticFieldException：统计分析失败异常。

（4）StatisticResult：统计分析结果。



图5- 15 数据统计分析模块类图

统计分析模块完成用户指定时间段的客户风险评估结果的统计和用户贷款审批时间的统计，用户在页面选择时间范围，点击统计分析可以查看统计分析的结果，页面实现如图5-16，我们对系统上线后的一段时间的客户借款申请进行统计分析，得到的平局审批时间为3天，这与风控数据分析系统上线前平均审批时间10个工作日相比大大提供了审批效率。

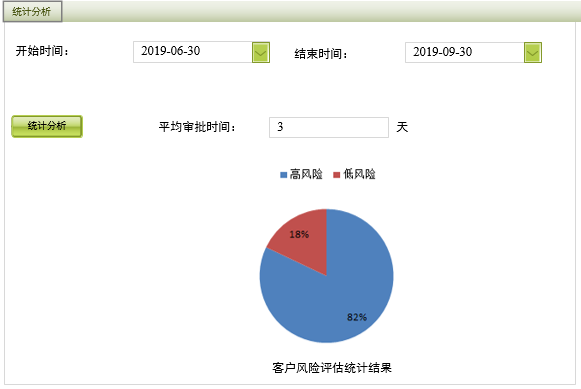


图5- 16 统计分析页面

## 5.3 本章小结

本章首先详细介绍了风险评估模型的设计实现过程；然后对风控数据分析系统的数据处理模块，数据挖掘建模分析模块，数据统计分析模块逐一介绍了接口和类的构造，说句处理的流程，具体的系统页面和功能实现过程。最后通过统计分析结果显示风控数据分析系统提高了贷款审批的效率。

# 6 总结与展望

## 6.1 总结

P2P借贷平台的研究在当今飞速发展的社会环境下具有深远意义。在当今互联网金融飞速发展的大环境下，本文研究并且实现了一套能够很好支持国内排名前几位的互联网金融公司的借贷系统的风控数据分析系统。满足了系统对风险控制的需求，缩短了贷款审批时间。就本文完成的工作作出总结如下:

目前系统建设需要解决的问题包括客户风险识别的准确性低和客户贷款审批时效慢。

为了提高风控数据分析系统客户风险评估的准确度，本文利用P2P平台的历史借款数据构建了客户风险评估模型。建设特征工程，完成风险评估特征指标的选择和加权处理；通过特征集分割，建设特征子集，屏蔽空特征填充造成的误差；对多种机器学习算法进行筛选，训练和优化，得到最优客户风险评估模型。最终得到客户防线评估模型的预测准确率在95%以上，召回率控制在94%以上。

为了提高客户贷款审核时效，将客户贷款申请的审批过程由数据统计分析结合专家经验对客户进行多层审核，评估客户风险，升级为线上对客户的数据信息进行筛选，处理，应用风险评估模型对客户数据进行评估，得到客户风险评估结果来决定贷款申请的审核结果。系统只需要一位信审人员在线上页面进行数据处理，客户风险评估和贷款审核的处理。风控数据分析系统参与最终将贷款的平均审核时间有10个工作日缩短到4个工作日之内。提高了用户的借款体验满意度。

## 6.2 后续工作与展望

随着互联网金融的不断发展和应用，P2P平台处理的业务量将来也会发生巨大增长，对风控机制会有更高的要求[25]。P2P借贷平台对相关的技术也会有更高的要求。本文虽然已取得了一定的的研究成果，但从长远的发展角度看来，仅是对P2P平台风险数据分析的初步探索。本文只是给出了P2P平台P2P平台风险数据分析的一种实现，解决了大数据环境下传统银行等金融系统相对难处理的一些问题，但此P2P平台在总体功能方面，性能方面还是存在巨大的优化空间风控系统的实时性准确性的优化，等等。

在互联网大数据的背景下，随着大数据处理和批量任务以及风控技术的进一步成熟，人工智能技术驱动的金融服务系统将成为必然趋势，下一步研究的内容主要是[26]：

1）目前P2P平台还远远未满足风控实时性的要求，需要进行研究网络风控大数据处理能力如何能进行提高，大数据计算和处理的算法怎样才能进一步得到优化；

2）进一步研究大数据分析和人工智能技术，在大数据和互联网环境下，建立如网络人行分析模型[27] 、感知预测态势模型，等等，通过海量事件的行为分析智能推测用户需求做到精准营销和智能化服务[28]。

# 参考文献

1. 钱金叶,杨飞.中国P2P网络借贷的发展现状及前景[J].金融论坛,2015,12(2):21-25
2. Shackleford D. SANS Security Analytics Survey [M]. 2013.

1. [田杰](http://kns.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e7%94%b0%e6%9d%b0&scode=28180869&acode=28180869" \t "knet), [郭紫嫣](http://kns.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e9%83%ad%e7%b4%ab%e5%ab%a3&scode=41146509&acode=41146509), [靳景玉](http://kns.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e9%9d%b3%e6%99%af%e7%8e%89&scode=10101387&acode=10101387).我国P2P网贷平台生存状况影响因素研究——基于3842家P2P网贷平台数据的实证分析[J]. [西部论坛](http://kns.cnki.net/kns/NaviBridge.aspx?bt=1&DBCode=CJFD&BaseID=CQSY&UnitCode=&NaviLink=%e8%a5%bf%e9%83%a8%e8%ae%ba%e5%9d%9b),2019,3: 71-72
2. 王冬吾. 中国P2P大数据风险控制现状分析[J]. 河北金融,2019, 8(9): 7-13.
3. McKinsey G lobal Institute .Bigdata:THE next frontier for innovation ,competition,and productivity.2016-06
4. 刘惠彬.数据挖掘及大数据分析技术在反网络欺诈中的应用[J].中国新通信 ,2019, 4(2): 5-12.
5. Wrona Ronald M. Medical data mining: The search for knowledge in workers' compensation claims[J]. American journal of industrial medicine,2019, 6(6):3-12.
6. 刘汝元.数据挖掘在人工智能中的应用分析[J]. 信息与电脑(理论版),2019, 5(9): 23-25.
7. Shackleford D. SANS Security Analytics Survey [M]. 2013.
8. Xiao Zhu. Research on Preprocessing Method of Performance Monitoring Data in Cloud Environment[J]. Atlantis Press,2019, 9(12): 22-35.
9. 黄志刚,刘志惠,朱建林. 多源数据信用评级普适模型栈框架的构建与应用[J]. 数量经济技术经济研究,2019
10. 张慧霞. 常用数据挖掘算法的分析对比[J]. 河南科技,2014, 11(3): 2-8.
11. 陈洁. 数据挖掘分类算法的改进研究[D]. 南京邮电大学,2018
12. Herzenste in,M.,Andrews,R.L.,Dholak ia,U.M.,and Lyandres,E. The democratization of personal consumer loans?Determinante of success in online peer-to-peer lending commuities[J].Working Paper,University of Delaware,Wilmingto n,DE,2008,(2):50-71
13. 邓自立.网络借贷平台P2P模式探索[J].中国流通经济,2014, 5(5): 12-16.
14. 郭勋诚. 朴素贝叶斯分类算法应用研究[J]. 通讯世界,2019, 8(9): 8-15.
15. 李云帆, 胡晧程, 康佳乐.朴素贝叶斯算法的应用[J].电脑编程技巧与维护,2018, 3(5): 22-34.
16. 赖莹. 支持向量机在P2P借款人信用风险评估中的应用[D].电子科技大学,2018
17. 彭晓冰,朱玉全. 基于特征内相关和互信息的加权SVM算法[J]. 计算机科学,2018, 7(6): 18-33.
18. 蒲杰方, 卢荧玲. 基于聚类算法和神经网络的客户分类模型构建[J]. 软件,2018, 12(4): 97-99.
19. 刘婷婷. 基于ALS算法的个性化推荐系统的应用研究[D]. 大连交通大学,2018
20. Shichao Zhang ,Chengqi Zhang,Qiang Yang,Data preparation for data mining [J] Applied Artifical Intelligence,2003 ,17:315-381
21. 林信良. Spring2.0技术手册[M]. 北京：电子工业出版社,2002..32-38
22. McGovern J, Adatia R, Fain Y, et al. Java2 Enterprise Edition 1.4 (J2EE 1.4) Bible [M]. John Wiley & Sons, 2011.
23. Freeman JW, Darr TC, Neely RB. Risk assessment for large heterogeneous systems; Proceedings of the Computer Security Applications Conference, 1997 Proceedings,13th Annual, F, 1997 [C]. IEEE
24. 李国杰. 大数据研究的科学价值 [J]. 中国计算机学会通讯, 2012, 8(9): 8-15.
25. 谭莹, 王丹. 基于流量的网络行为分析模型的设计与实现; 2010 年全国通信安全学术会议论文集, 中国云南昆明, F, 2010 [C].
26. 陈涛, 龚正虎, 卓莹. 综合网络态势分析模型研究; 中国通信学会第六届学术年会论文集, 中国广东广州, F, 2009 [C].

# 本人攻读硕士学位期间发表的学术论文

1. 殷艳梅.论预付费系统项目整体管理.商情,2019,16(4):171-172.

# 致谢

研究生的求学历程接近尾声，在这段重要的人生历程中我收获了很多，良师益友，知识能力，要感谢的人也很多。首先，要感谢我的导师刘学洋老师，他严谨朴实的教风、勤恳务实的原则，都潜移默化的影响了我，让我受益匪浅。在我的学术论文和学位论文编写过程当中，刘老师自始至终高标准，严要求，并且及时解答我遇到的各种困惑，对我的毕业论文给予悉心指导，大大提高了我论文的质量。

其次，我要感谢参与我论文评审工作的各位老师，从各个方面给我的论文提供参考意见，帮助我把论文完成的更加优质完善。

我要感谢我的班主任张霞老师，在研究生阶段给我做多支持的是班主任张霞老师。有了她的帮助我才顺利通过课程考试和毕业。感谢陈向群教授、王韬副教授，感谢他们在授课过程中教授我的知识，他们用治学态度给我熏陶和引导。感谢汪国平教授、刘志敏副教授、陈立军副教授、屈婉玲教授、王捍贫教授，感谢他们对我上课过程中耐心的指导，让我能顺利完成各门课程。

我还感谢我的家人，是他们在我研究生阶段做我的坚强后盾，让我免受家庭琐事困扰，给我陪伴和支持。我要感谢我未满周岁的女儿，虽然因为学业牺牲了很多陪伴她的时间，但她给予了我很多的精神鼓励，支持我顺利完成了学业。

我要感谢友信金服公司的贾艳斌总监，感谢项目团队的范林林、王佩佩、赵祥、张亚南，感谢他们在我编写论文过程中给我的帮助和支持。

我要感谢我研究生阶段的同学袁明明、李凤霞、李丹丹，感谢他们在我编写论文过程中的支持和帮助。感谢朱雨、罗文、邓守秀、翟英威、张昊炜，感谢他们在完成课程和作业期间的协作和帮助。

最后，再次感谢我的导师刘学洋老师，感谢评委专家教授们，感谢所有帮过我的良师益友，亲朋同事，祝愿各位幸福安康！

# 北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

**学位论文使用授权说明**

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日