基于大数据技术的P2P网贷平台风险预警模型

林春雨[1]，李崇纲[1]，许方圆[2]，许会泉[1]，石 磊[1]，卢祥虎[1]
1. 北京金信网银金融信息服务有限公司 北京 100101；2. 国网能源研究院 北京 100101

摘要
近几年，我国P2P网贷行业在高速发展的过程中出现了大量的“失联跑路”事件。为此，基于P2P网贷及大数据相关概念的深入剖析，创新性地将平台的风险预警与大数据技术结合，通过对海量数据采集、Spark分布式平台计算、机器学习建模等大数据技术的整合，构建一个有效的P2P网贷平台风险预警模型。该模型在多维度风险评价指标的基础之上，可以实现对网贷平台风险的实时、精准、全面监测，从而有效降低平台集资诈骗、恶意跑路等恶意事件的发生频率，维护广大投资人的资金安全及社会稳定。

关键词
互联网金融；P2P网贷；大数据；风险预警；机器学习

doi: 10.11959/j.issn.2096-0271.2015037

A Model of Pre-Warning Based on the Big Data Technology for P2P Lending Platform

Lin Chunyu[1], Li Chonggang[1], Xu Fangyuan[2], Xu Huiquan[1], Shi Lei[1], Lu Xianghu[1]
1. Beijing JinXinWangYin Financial Information Service Co., Ltd., Beijing 100101, China;
2. State Grid Energy Research Institute, Beijing 100101, China

Abstract
In recent years, P2P lending industry in China has appeared a lot of escape events in the process of its rapid development. Bases on deep analysis of the related concepts for P2P lending and big data, combining innovatively the risk pre-warning of platform with big data, an effective risk pre-warning model of P2P lending platform was constructed according to the collection of huge amounts of data, big data technology including Spark distributed computation and machine learning. Based on the establishment of multi-dimensional risk assessment, the model can be achieved on real-time, accurate, comprehensive monitoring for the risk of P2P lending, thus effectively reducing the frequency of financial fraud, escape malicious event, so as to the majority of investors' money to maintain security and social stability.

Key words
internet financial, P2P lending, big data, risk pre-warning, machine learning

1 引言

近几年，我国互联网金融发展十分迅速。一方面，互联网金融的发展可以很好地满足中小微企业、创新型企业及中低收入阶层个人的投融资需求，为“大众创新，万众创业”营造良好的资本环境；但另一方面，互联网金融在创新发展过程中也暴露出大量的问题及隐患。本文通过对其中的P2P网络借贷平台运营状况进行相关调查发现，截至2015年6月底，P2P网络借贷平台累计达到2 814家，其中问题平台为786家，比例高达27.93%，其不仅严重危害了人民的财产安全，也有碍互联网金融的健康发展。如何有效地监测到潜在的具有高风险的平台就成为一项非常有意义的研究。

P2P网络借贷平台发端于英国，成熟于美国。从资金流向来看，国外的P2P网络借贷资金主要流向小额信贷领域，借款主体主要为个人，其用途也是为了满足个人消费需求和补充个体户经营的流动资金需要。对于这些借款人，P2P平台仅需要通过个人征信报告确定其信贷违约风险，并将可公开的信息提供给投资人，最终由借贷双方直接达成借贷协议。因此，在完善的征信体系与政府监管环境下，可以通过行业自律等方式有效预防问题平台的出现。而在我国，P2P网络借贷不仅为个人服务，而且很大程度上也服务于中小微企业，在风控手段上必须依靠强化抵押与质押品的要求以及引进有实力的融资性担保机构对项目进行担保。这样，投资人的信贷风险不再主要取决于个体项目的违约风险，而主要取决于平台合作方的担保实力与抵押品的实际抵押能力[1]。这其中还存在平台与合作方相互勾结的风险，因此，由第三方对P2P网络借贷平台的风险进行监测预警势在必行。

目前国内关于网络借贷平台风险及预警的研究还处于初级阶段，研究的内容也都是从金融业务层面进行展开，比如黄叶苣、齐晓雯认为P2P面临的风险主要包括由于法律缺失导致的监管风险、用户导致的风险及借贷平台自身运营与网络技术带来的风险[2]；而胡旻昱、孟庆军基于系统科学理论辩证地分析了P2P平台所面临的风险，他们认为环境对系统有“压力”，即平台会受到金融危机、行业法律缺失、机构主管单位不明确等外界风险的影响，反过来系统自身会对环境有“污染”，网贷平台自身监管不到位、系统安全漏洞、担保机构与征信机制不完善等都是平台自身引发的风险[3]；余及尧等基于2013-2014年P2P网贷平台样本数据，运用logistic回归模型从企业性质、收益率及风控保证模式3个方面对平台发生财务困境的影响进行研究，结果表明其与短期收益率呈显著正相关[4]；随后马玉娟通过分析P2P的主要风险类型，同时综合相关专家的评审构建了包含信用评级、流动性、信息透明度、技术服务、品牌、杠杆率6个方面内容的指标体系，然后结合运用主成分分析和改进的KLR信号分析法建立了风险预警模型，最后对20家网贷平台进行了综合打分和排名，验证了模型的可行性和准确性[5]。但是由于数据量少，对于模型的准确性结论就有一定的局限性。王楚珺、刘会芳等人认为P2P网贷主要存在信用评估、业务监管及系统安全三大风险，并且提出可以将大数据引入P2P的风险控制工作[3]，但是他们并没有深入分析与研究，更没有提出一个具体的风险控制模型。

综上所述，已有的相关研究主要还是集中于理论上的探索，但是P2P网贷平台风险评估是一个跨领域和多数据来源的复杂问题，多方面的数据采集和多角度的特

征分析是最终模型能够完成准确预警的重要保障。基于大数据体量大、类型多、速度快、时效高的特点，可以大大扩宽用于最终模型训练的历史数据特征字段，因此本文将基于大数据及相关技术完成对P2P平台监测预警模型的构建。

2 基本理论与关键技术

2.1 P2P网络借贷

P2P网络借贷又称为点对点借贷，指非金融机构利用互联网或移动平台为民间借贷双方提供的借贷信息中介服务，包括信息发布、交易撮合以及为实现交易撮合而提供的风险评估、信用评价、投资咨询、交易管理及资金流转等服务[6]。2005年3月，全球第一家P2P网贷公司Zopa在英国伦敦成立，接着美国两大巨头网贷公司Prosper和Lending Club先后成立，而我国第一家网络贷款平台拍拍贷在2008年上线，直到2011年，我国的网贷平台迎来了高速增长时期。截至2015年6月底，全国累计平台数量达到2 814家。P2P网贷在我国经过探索创新，其主要运营模式主要包括3类：一是“纯线上中介”模式，此模式借贷双方通过相应网络平台发布信息，自行配对、自主成交，而P2P企业此时只充当交易撮合平台和资金划转平台，但不参与或较少参与借贷交易，没有线下审贷环节，也不对借款提供担保；二是“担保赔付”模式，即P2P网络借贷平台事先承诺，当借款人延迟付款时，在一定条件下由平台从风险拨备中先期垫付本金和利息，或由平台合作的担保机构垫付本金和利息，此举可以有效降低违约风险，吸引更多投资人；三是“线上+线下复合”模式，此模式将不止依附于自身网络平台，相关业务人员会直接到线下寻找投资者及借款人，并对借款人开展实地信用调查。由于激烈的行业竞争，目前国内平台大多将后两种模式相结合来最大限度地争取投资人、减少信用风险等。

2.2 大数据内涵

移动互联网、物联网和云计算技术的迅速发展，开启了移动云时代的序幕，大数据也越来越多地被人们所了解和利用。目前，对于大数据来说并没有一个明确的定义，李国杰等人认为大数据是指无法在一定时间内用常规机器和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合[7]；胡雄伟等人认为大数据是指数据量的大小超出了传统意义上的数据尺度，一般的软件工具难以捕捉、存储、管理和分析的数据[8]；而维基百科中将大数据定义为：所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理，并整理成为帮助企业经营决策更积极目的的资讯。对以上相关定义进行总结归纳，笔者认为大数据是指由于结构复杂、种类繁多、数量庞大而无法在一定时间内运用常规工具对其进行获取、存储、分析及感知的数据集合。对于大数据的特性，比较有代表性的是4V定义，即认为大数据需满足4个特点：规模性、价值密度低、多样性和高速性。目前，大数据在电信、智慧城市、电子商务及社交娱乐等行业已经出现规模化应用，随着网速的进一步提升，数据将迎来新一轮爆发式增长，今后能够快速获取、处理、分析海量、多样化的数据对政府及企业来说都是至关重要的。

2.3 基于Spark的分布式计算

分布式计算研究如何把一个需要巨

大的计算能力才能解决的问题分解成许多小的部分，然后把各个小的部分分给若干计算机同时进行处理，最后把这些计算结果综合起来得到最终结果[9]。之前运用较多的是传统的MapReduce框架，它将一个任务的执行过程划分为两个阶段，即map阶段和reduce阶段。在map阶段，每个map任务读取一个block，并调用map函数进行处理，然后将结果写到本地磁盘上；而在reduce阶段，每个reduce任务远程地从map任务所在节点上获取相关数据，并调用reduce函数进行数据处理，最后将结果写入HDFS（Hadoop distributed file system，Hadoop分布式文件系统）。但是这种方法在两个阶段计算的结果均要写入磁盘，因此系统性能降低，很难满足迭代编程的要求。为了解决迭代问题，Spark应运而生，它是基于MapReduce的新一代大数据分析框架，吸收了前者的所有优点，但Spark将计算的中间结果数据存储在内存中，通过减少磁盘I/O，使后续的数据运算效率更高。Spark的这种架构设计对于需要重复利用计算中间数据的机器学习、交互式数据分析等工作十分适用。

2.4 文本挖掘技术

由于本文是基于大数据的建模，原始数据中包含了大量的新闻报道、社交文本等非结构化数据，必须运用相应的文本挖掘技术对其进行排重、分词、分类等一系列的结构化处理。

2.4.1 文档分布式排重

排重技术是指根据词语的抗篡改能力及语义信息等特征生成词语指纹，然后根据词语指纹对不同文本进行检测以排除相似性文档。具体过程分为两个层次，即粗排重和细排重，粗排重是对一篇文档只生成一个指纹来进行初步的排重，而细排重则是在前者的基础上，针对更细分的主题对文档生成一组指纹来进行更加精准的排重。由于网络信息发布主体的去中心化，相同信息（尤其是较为敏感的负面信息）会被多个主体进行报道，同时这些报道还会在论坛、微博等社交网络中进行转载和评论，致使网络中出现大量的重复信息。因此，对其进行自动排重将会大大提高后面工作的效率及准确性。

2.4.2 自动分词技术

自动分词是计算机针对一段文本，按照词性、语义等将其自动切分成单个词汇的过程。人们通过大脑识别文本中的词汇是依赖于对语言的理解和积累而形成的思维，但对于机器来说显然是不具备此种思维的，因此利用机器进行准确分词是比较困难的，其涉及的主要问题包括分词规范、歧义词切分及新词识别。经过相关学者的探索，目前主要的自动分词方法包括机械分词算法、基于统计的分词算法及基于知识的分词算法。其中基于知识的分词算法是通过计算机模拟人类对句子的认知过程来达到分词的目的，但是这种方法目前还处于研究阶段。另外两种方法相对已经比较成熟，但各有优缺点，基于统计的分词方法通过判断相邻字同时出现的频率将共现频率高的字当成一个词汇分离出来，但在实践中发现这种方法准确率较低。本文采用的分词技术是机械分词算法，它利用一定策略将待分词文本与预先准备的语料库进行匹配来达到分词的目的，虽然这种方法使用简单、实用性强，但是其语料库词汇往往会少于实际应用中遇到的词汇量。为了解决这一问题，笔者研究团队制作了近10万个词的基础分词词典，同时通过定期与客户交流建立客户词典来

进行有效补充。本文将利用自动分词技术来抽取新闻、社交等文本信息中各类主题的关键词，以达到文本分类的目的。

2.5 机器学习

1997年Mitchell T M给出了一个机器学习的经典定义，即计算机利用经验改善系统自身性能的行为[10]。人类具有学习能力，其学习行为背后具有非常复杂的逻辑判断过程，机器学习正是以此过程中人脑对信息的处理机制为理论依据，利用计算机来模拟实现人类获取知识的过程，再通过不断地创新、重构已有知识，最终提升计算机处理问题的能力[11]。在大数据环境下，只有运用机器学习的方式才能帮助人们从各式各样的海量数据中挖掘出其中所蕴藏的价值。因此，本文试图利用机器学习法对预处理后的大量特征字段进行反复的训练，以找出真正与平台高风险相关的指标及精准的预警模型。

3 研究假设

通过全面分析当前P2P网贷平台出现风险的原因，本文总结提出以下4条假设。

3.1 H1：运营数据异常程度与平台风险呈正相关

P2P平台的运营数据主要包括借贷人数、借贷金额、预期收益率及平台标的信息等。上述运营数据在行业内通常会有一个合理的取值区间，当某些数据脱离此区间太远时，平台可能会产生相关问题。例如平台预期收益率远远高于行业平均水平，而平台中显示的标的数量却很少，则此时该平台很有可能出现“资金池”现象。平台运营数据是与其风险关联最为直接的指标，数据越偏离合理区间，其面临的风险就越大。

3.2 H2：网络负面舆情数量与平台风险呈正相关

网络舆情是指由于各种事件刺激而产生的，并通过互联网传播和形成的人们对于该事件的所有认知、态度、情感和行为倾向的集合[12]。网络舆情来自于现实世界，同时又会从正面或负面反作用于现实世界，尤其是一些涉及民生、政风等负面敏感事件，网络会迅速将其变为全民热议的公共话题。因此，基于网络舆情传播的及时性与广泛性等特点，将有关P2P平台的负面舆情比例作为其风险预警指标是十分有效的。

本文是通过各大新闻及行业协会网站、论坛、微博等搜集P2P平台的相关文本信息，然后通过文本分类整理出其中所包含的负面信息（非法、虚假宣传、投诉等），这些负面信息可以及时、全面地揭示平台当前存在的问题，问题越多，面临的风险也就越大。

3.3 H3：平台及相关法人信用状况与平台风险呈负相关

P2P平台发生风险形成跑路的原因有两种：一种是自身运营不当，一种是恶意集资诈骗。现实过程中，很多平台以无风险、高收益等虚假宣传来吸引客户进行投资理财，实则是建立资金池以便自用。以上便涉及了平台及法人的信用问题，通过查询平台关联企业及相关法人的信用信息和涉诉信息来对其信用度进行判断，其信用度越高、涉诉牵连越少，平台风险就越低。

3.4 H4: 平台背景实力与平台风险呈负相关

P2P平台的背景实力主要包括其注册与实缴资本金数量、合作担保及资金托管机构、关联企业背景等。一些拥有国资上市公司背景的平台一般不存在跑路、非法集资等恶性事件，另外其在资金和管理团队方面具有一定的优势，能够较好地应对平台中产生的逾期与坏账。所以，平台背景实力越强，其拥有的风险将越低。

在接下来的建模过程中，将会针对每一条假设建立相应指标字段，从而对其进行验证。

4 风险预警模型的建立

整个模型构建过程：首先是运用不同的方法对大量原始数据进行采集；然后需要对其进行缺失值修补、异常值检测等一系列的数据预处理，使原始数据格式规范统一，以满足训练模型的要求；接着将处理后的数据分成训练样本和测试样本两部分，将训练样本带入多种模型进行机器学习，同时利用测试样本来验证不同模型的准确性，并通过增减原始字段及进一步的预处理来不断优化改进模型的准确性；最后则是平台功能实现的展示。具体流程如图1所示。

4.1 数据采集

平台自身的运营不善及相关人员的恶意欺诈是P2P网贷平台重要的风险构成因素，这两个因素在平台的日常运营、诚信记录、涉诉情况及相关网络舆情等方面均会有所表现，因此这些信息可以作为风险预警的判断依据，信息集合如图2所示。

本文经过深入研究，最终确立了与P2P平台风险大小紧密相关的六大特征集合（如图2内环所示），即企业基本特征、运营状况特征、模式与制度特征、平台诚信记录、运营者信用信息及平台宣传信息，这六大特征集合完整地描述了相关平台的背景实力、风险保障、标的及利率、企业与个人信用、网络新闻、社交舆情、涉诉等内容。这些信息的来源（如图2中环所示）主要包括工商注册信息、平台网站信息数据、宣传信息数据、征信数据、银行数据及其他数据。针对不同的数据来源还需要运用不同的方

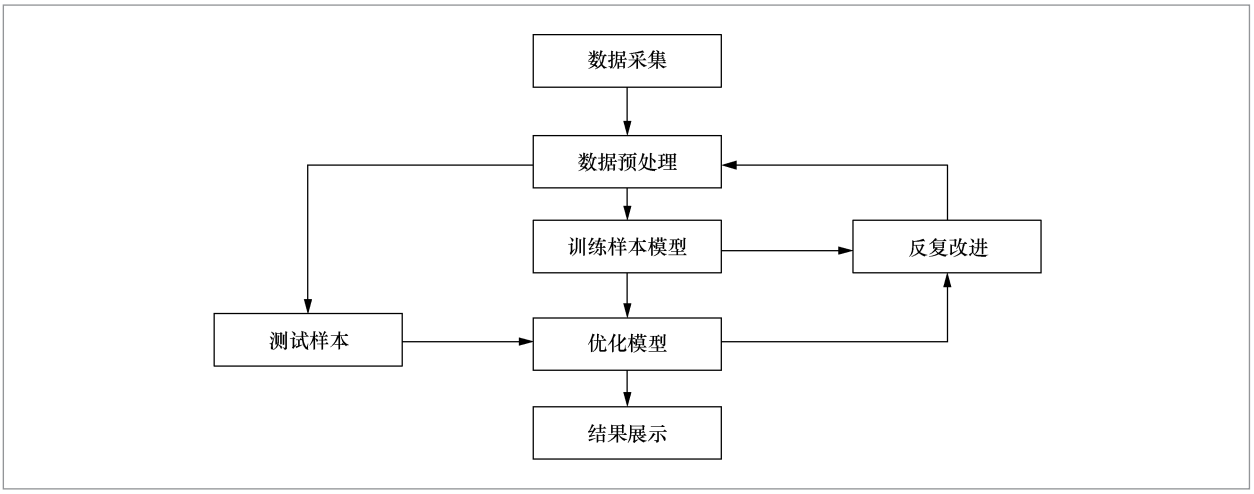


图1 模型建立流程

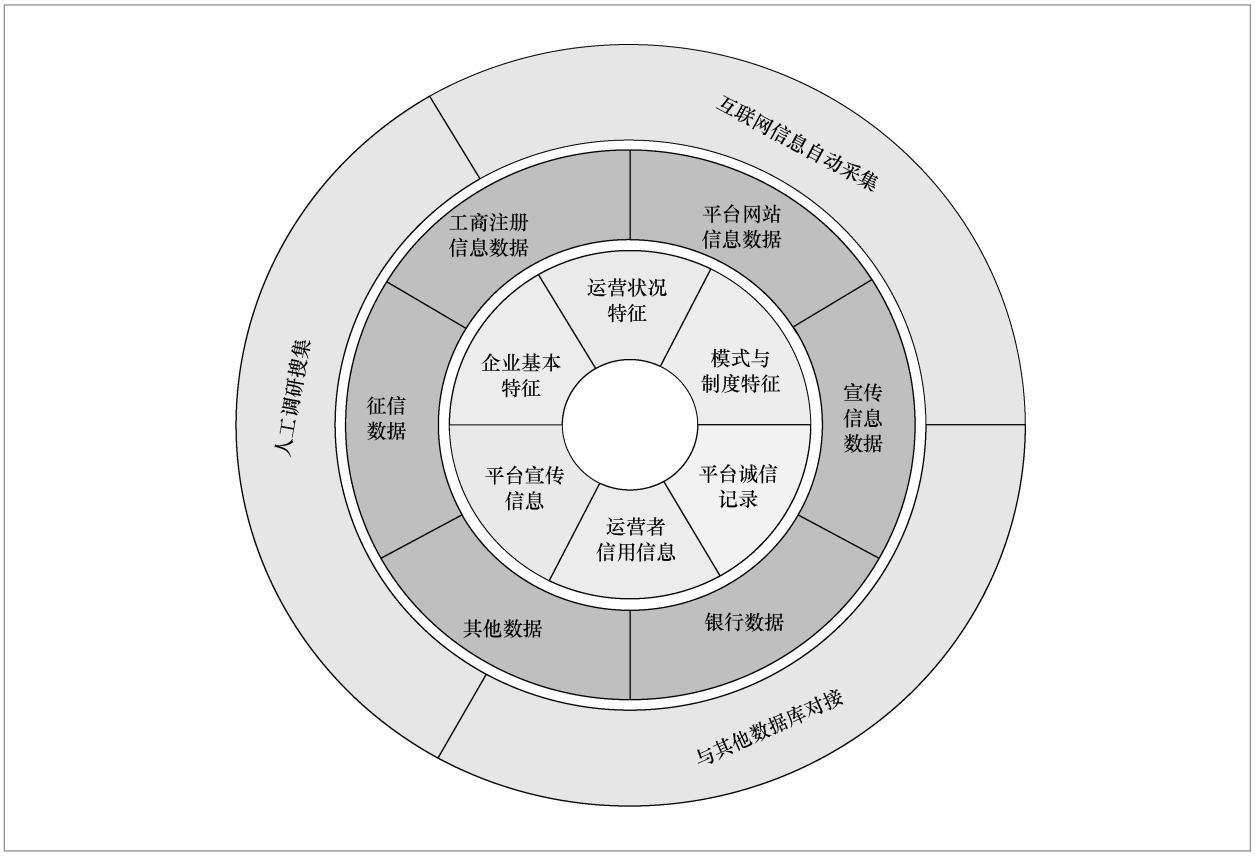


图 2 P2P 网贷平台的特征选取和数据采集

法（如图2外环所示）进行采集，对于网络中的结构化数据及公开的文本数据，可利用相关软件对其进行自动采集，如P2P平台的运营数据、新闻报道及微博、论坛等社交信息；而针对网络中一些特殊的非结构化数据或者非互联网的数据，则选用人工调研搜集的方式，如平台规模及背景实力等；最后对于其他机构已经搜集整理好的数据，则采用数据库导入或对接的方式来直接获取。

本文实际采集了100家正常平台和100家问题平台的上述所有特征集合数据，并结合经过专家评分后的专家库数据共同作为模型训练时的数据源。

4.2 数据预处理

最初采集到的数据结构类型各异，同时存在大量的缺失、异常等问题，因此需要对其进行一系列的预处理才能用于之后的数据挖掘与建模工作之中。本文主要的预处理过程包括以下几步。

4.2.1 文本信息处理

对于原始数据中大量的新闻报道等非结构化数据，需要进行语义分析及文本分类，将其转换为相应数值指标。转换过程为：通过关键词自动提取及人工判断，选出可以区分不同主题（平台非法性、平台投诉类、平台虚假宣传类）的关键词，然后运用这些关键词制定相应的检索表达式，以实现对相关信息的自动分类检索，最后将有关某家平台的各个主题内信息数量除以所有相关信息量，得出舆情投诉率等数值型指标。

4.2.2 缺失值处理

在对平台运营数据采集的过程中出现了一些数据缺失现象，针对此问题，主要采用字段均值及拟合函数的方法来解决，而针对个别存在大量数据缺失的字段，则选择直接弃用。

4.2.3 异常值检测

在运营数据采集过程中还会出现少量的异常值，如果不对其进行有效处理，则会严重影响模型分析效果。本文对于异常值的处理综合采用了以下几种方法：通过距离方法来检测，即设立一个阈值，将数据中与平均值之间的距离（欧式距离）大于这个阈值的点设为异常点；通过聚类分析，相似或相邻近的数据聚合在一起形成了各个聚类集合，而位于这些聚类集合之外的数据对象则被认为是异常数据；利用拟合函数法对数据进行平滑处理以发现异常点。

4.2.4 数据一致性处理

在原始数据中，经常会出现数据单位及类型不一致的现象，例如：有些平台综合利率采用月利率计算，有些则用年利率计算，此时就需要对其单位进行统一，解决方法是在程序里使用正则匹配等方法来统一数据单位和数据类型。

4.2.5 数据转换

最后，由于采集到的数据包含多种结构类型，因此需要各个特征字段计算口径统一才能用于模型的建立。针对不同的字段特征将选用以下几种方法进行处理。

（1）归一化

归一化是一种简化计算的方式，即将有量纲的表达式变换为无量纲的表达式，使数值的绝对值变为某种相对值关系。由于建立的指标取值区间相差较大，因此利用此方法对其进行归一化处理。归一化转换的方式包括线性函数转换、对数函数转换和反正切函数转换3种，本文针对不同字段特征选择不同的函数形式进行转换。

（2）数据泛化

数据泛化是一个从相对低层概念到更高层概念且对数据库中与任务相关的大量数据进行抽象概述的一个分析过程。本文主要是运用主成分分析法对大量的原始特征字段进行降维处理，排除一些相关性较强的无用字段，以提高建模过程的运行速率与最终模型的准确性。

4.3 模型构建与优化

本文将基于Spark分布式计算平台，利用机器学习方法选取多种模型来对训练样本集进行训练，并通过测试样本集对其准确性进行检验，最终通过对原始数据字段及预处理的反复调整以期得出一个最优的P2P平台风险预警模型。在建模过程中，特征字段选取、模型选择及结论解释3部分内容将是本部分研究的主要关注点。

4.3.1 特征字段选取

针对预处理之后大量的可用特征字段，需要通过相关性分析和卡方检验等方法逐一验证这些特征与平台欺诈事件的相关性，将对P2P网贷平台风险影响不显著的无效字段进行有效剔除，以保证分析结果的准确性及模型运算效率。

4.3.2 模型选择

能根据实时的数据集进行学习并不断修正优化自身的判断能力，是对优质模型的基本要求。由于模型输出的风险指标主

要用于判断P2P平台存在欺诈的风险性，因此输出变量是二项分布，且风险指标必须是序数型变量。可用于该种情况的分析模型包括逻辑回归、人工神经网络、贝叶斯网络等。系统为每一种备选模型进行建模，通过对比并最终选择出最佳的风险预警模型。

4.3.3 结论解释

P2P平台风险评估系统主要用于辅助系统使用者进行决策。而系统使用者进行决策后，往往需要向质疑者提供充分的解释。因此结论解释功能尤为重要。例如当系统面向某个平台的某个风险指标较高，系统使用者或者对象平台直接质疑系统的准确性时，就需要给出合理的解释。而最佳的解释依据应当为原始数据集合中的一般性统计结果。在机器学习的大多模型中，由于模型包含非线性的传递函数，这使得模型通常具有较强的学习能力，但亦将输入和输出的直接联系模糊化，增加了结论解释的难度。在众多的模型中，贝叶斯网络是结论解释能力较强的模型。其利用朴素贝叶斯理论的可逆推性，在输出的结果与原始数据中的一般统计性结果中建立线性联系，使得其结果较容易使用一般统计性结果进行描述。

围绕以上3个核心问题，在整个模型构建与优化的过程中，通过不断地对比分析及交叉验证不同模型各个方面的表现，以最终建立一个最佳的平台风险预警模型。

4.4 预警平台功能展示

整个建模过程最终的目的是搭建出可以面向用户的P2P风险监测预警平台。该平台可以实现两方面的功能：对P2P平台所面临的风险进行实时全面的评分，并针对其风险状况生成详细的风险分析报告，以为其风险的后续应对工作提供必要的建议措施；多维度地展现行业整体风险情况，如将平台按地区、时间、类型等不同内容进行风险分类统计，以清晰直观的方式满足不同用户的多样化需求。

5 结束语

由于P2P行业在我国发展时间比较短，因此相比传统金融机构，其在不断的摸索创新过程中会面临更加多样的风险。而本文的创新之处正是在于将模型的建立与大数据相结合，借助于先进的自动文本采集、Spark分布式计算、文本挖掘等技术来建立更加全面的指标体系，最终利用机器学习的方法对采集到的多维度历史数据进行反复的训练与改进，以构建出一个准确、有效的P2P网贷平台风险预警模型。基于以上模型搭建的预警平台通过数据每日自动更新，便可实现对网贷企业的实时监测预警，并从多种角度展现其风险状况。该平台不但可以用来协助政府监管机构开展相关工作以有效地预防平台跑路、诈骗等问题事件的发生，还可以为广大的平台投资者提供投资风险警示以保障其资金安全。

致谢

本研究得到首都经贸大学金融学院周晔老师、余颖丰老师以及北京大学常国珍博士的帮助，谨致谢意！

参考文献

[1] 陈文等. P2P中国式高收益债券投资指南. 北京: 机械工业出版社, 2015

Chen W, *et al*. P2P Chinese High-Yield Bond Investing for Dummies. Beijing: China Machine Press, 2015

[2] 黄叶苊, 齐晓雯. 网络借贷中的风险控制. 金融理论与实践, 2012(4): 101~105
Huang Y N, Qi X W. Risk control of the P2P lending. Financial Theory & Practice, 2012(4): 101~105

[3] 胡旻昱, 孟庆军. P2P网贷平台发展中的风险及其系统分析. 武汉金融, 2014(6): 45~48
Hu M Y, Meng Q J. Risk of the developing P2P lending and its system analysis. Wuhan Finance, 2014(6): 45~48

[4] 余及尧. 互联网金融财务困境预警与监管对策——基于2013-2014年P2P网贷平台样本数据分析. 福建金融, 2015(2): 42~47
Yu J Y. Internet financial early-warning and regulatory measures--based on P2P lending platform in 2013-2014 sample data analysis. Fujian Finance, 2015(2): 42~47

[5] 马玉娟. 互联网金融风险预警研究——以P2P网络借贷模式为例(硕士学位论文). 锦州: 辽宁工业大学, 2015
Ma Y J. The warning research on internet financial risks--the study of P2P lending (master dissertation). Jinzhou: Liaoning University of Technology, 2015

[6] 黄旭, 王素珍, 赵洋. P2P 平台: 发展与监管. 中国金融, 2014(5): 90~93
Huang X, Wang S Z, Zhao Y. P2P platform: the development and regulation. China Finance, 2014(5): 90~93

[7] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012(6): 647~657
Li G J, Cheng X Q. Big data research: the future of science and technology and economic and social development of major strategic areas--research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012(6): 647~657

[8] 胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述(上). 标准科学, 2013(9): 29~34
Hu X W, Zhang B L, Li D F. Overview of big data research and application (part A). Standard Science, 2013(9): 29~34

[9] 黎连业, 王安, 李龙. 云计算基础与实用技术. 北京: 清华大学出版社, 2013
Li L Y, Wang A, Li L. Cloud Foundations and Practical Technology. Beijing: Tsinghua University Press, 2013

[10] Tom Mitchell. Machine Learning. New York: McGraw Hill Higher Education, 1997

[11] 陈康, 向勇, 喻超. 大数据时代机器学习的新趋势. 电信科学, 2012, 28(12): 88~95
Chen K, Xiang Y, Yu C. The new trend of big data era of machine learning. Telecommunications Science, 2012, 28(12): 88~95

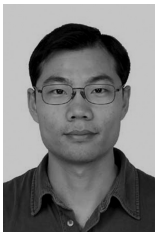
[12] 高承实, 陈越, 荣星等. 网络舆情几个基本问题的探讨. 情报杂志, 2011(30): 52~56
Gao C S, Chen Y, Rong X, *et al*. Some basic problems on network opinion research. Journal of Intelligence, 2011(30): 52~56

作者简介

林春雨, 男, 现任北京拓尔思信息技术股份有限公司高级副总裁、助理研究员, 负责公司大数据中心建设和云服务运营工作, 在社会化媒体的技术运营和管理上有丰富的实战经验, 其同时兼任北京金信网银金融信息服务有限公司总经理, 为各地金融监管机构提供非法集资监管服务。另外, 作为国家信息安全专项舆情云服务项目组长、中关村大数据产业联盟副秘书长, 为多个国家部委、省级客户、大型企事业单位提供过相关高端舆情服务, 并通过联盟和产业对接, 积极推动大数据的发展。

李崇纲，男，北京金信网银金融信息服务有限公司常务副总经理，拓尔思信息技术股份有限公司高级顾问，中国计算机学会大数据专家委员会委员，中关村互联网金融协会副秘书长，中关村大数据产业联盟专家组成员。专注于大数据在政府、金融等行业领域的应用，拥有10多年网络数据挖掘分析、互联网大数据分析经验，担任多家政府企业舆情管理咨询顾问，是国内首款舆情监测系统的设计者，长期跟踪互联网大数据行业变化。目前主持开发国内首个大数据防控金融信用风险与智能决策支持系统。

许方圆，男，国网能源研究院能源决策支持技术研发中心中级工程师，主要从事智能电网技术和政策的分析研究，近年来主要研究方向为需求侧响应实施与应用、电力系统中的数据挖掘应用、全球能源互联网，发表论文10余篇。



许会泉，男，北京金信网银金融信息服务有限公司研发总监，负责公司互联网金融、机器学习、大数据产品研发、管理工作，在计算机系统架构设计、大数据应用、舆情产品应用等方面具有丰富的实战经验，近年负责主持研发了公司金融大数据打非监测预警云平台、互联网金融风险模型等多个大数据产品。

石磊，男，北京金信网银金融信息服务有限公司互联网金融行业数据分析师，主要负责研究行业目前所具有的非法集资风险特征，并基于大数据对相关企业风险进行监测与评判，拥有丰富的理论及实战经验。

卢祥虎，男，北京金信网银金融信息服务有限公司机器学习算法工程师，目前从事P2P风险预警建模相关的算法设计工作，在机器学习算法领域具有一定的理论与实战经验，擅长机器学习中数学算法的优化与改进等。

收稿日期：2015-09-30

论文引用格式：林春雨，李崇纲，许方圆等. 基于大数据技术的P2P网贷平台风险预警模型. 大数据, 2015037

Lin C Y, Li C G, Xu F Y, ***et al***. A model of pre-warning based on the big data technology for P2P lending platform. Big Data Research, 2015037