学校编码：10384　　分类号_____密级_____

学号：23020101153052　　UDC_____

硕　士　学　位　论　文

P2P 网络借贷中的数据分析与决策模型研究

Data Analysis and Investment Decision Model in P2P Online Lending

吴小英

指导教师姓名：鞠颖副教授

专　业　名　称：计算机系统结构

论文提交日期：2013 年 4 月

论文答辩时间：2013 年 5 月

学位授予日期：2013 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2013 年　5　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

20　　年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

20　　年　　月　　日

摘要

P2P(Person-to-Person)网络借贷是近年来出现的新兴事物。在网络借贷平台上，借贷双方不需要以银行等传统金融机构为中介，直接通过交易无担保借贷。美国最大 P2P 网络借贷平台 Prosper.com，其超过 120 万的会员及将近 30 亿美元交易量为投资分析提供了充分的数据。本文针对 P2P 网络借贷的特点，从借款人的角度分析借款用途对借款成功率的影响，并建立贝叶斯模型和二部投资网络模型帮助投资者进行投资决策。本文的主要工作及创新点如下：

从借款人的角度出发，研究网络借贷中借款用途对借贷成功率的影响，以单一变量原则建立数学模型，并使用最小二乘法进行参数估计，回归实证研究表明：同等条件下，学生借款比其它种借款成功率低 3.4%，分析其原因发现学生还款率并不比其它的低，因此对学生存在直觉歧视。通过类似的方法，本文还发现用于汽车或者其它方面的债务更容易借到钱。同时，实证研究还得出了其它若干因素，如借款金额、利率等对借款成功率的影响。

从投资者的角度出发，研究借款人的属性(信用评分、是否有住房、债务收入比)及借款项目的属性(借款金额、利率、借款用途、每月还款金额、投标数目)对还款率的影响。首先利用一年前的数据作为模型训练数据集，通过计算两两属性之间的互信息得出它们之间的关联程度，利用 Prim 最大生成树算法建立无向图，最后选任意顶点作为根节点并将目标节点作为每个节点的父节点，建立贝叶斯网络。再利用此贝叶斯模型对一年以后的数据进行预测，将预测结果与实际还款情况进行比较。实验结果表明：贝叶斯模型预测还款概率较高的借贷，实际还款概率确实比较高，特别是，预测概率高于 90%的借款，实际还款率达到了 100%。因此，贝叶斯网络模型通过综合考虑借款人及借款本身的各个属性，能够帮助投资者做出正确的决策。

P2P 网络借贷中，投资人和借款项目形成了一个多对多的二部图网络，投资者和借款项目分别是二部图两边的顶点，每条边的权值为投标金额。本文建立了基于该二部图的评分模型，先利用已知状态的借款项目对投资人进行评分，再用投资人对未知状态的借款项目进行评分，通过这种相互评分，能得到整个二部图

网络各结点的分值，实验结果证明：该方法所得分数对投资决策具有重要的参考价值，分值较高的借款项目具有较好的投资价值。

关键词：P2P 网络借贷，最小二乘法，贝叶斯网络，二部图

Abstract

P2P online lending is an emerging economic lending model, allows individuals to lend to and borrow from each other directly on Internet-based platform without the participation of traditional financial intermediaries. In this study, we focus on Prosper (http://www.prosper.com), the largest online P2P lending market in US, which has helped its 1.26 million members receive over $314 million loans. In this dissertation, we study the influence of the purpose of lending on success rate at online lending marketplace and develop two effective investment models to enhance investment decisions in P2P lending.

The first model is built from the lendee composition perspective, we establish mathematical model, and employ ordinary least square, univariate multiple regression. The experimental results show that, the successful rate of student loan is 3.4% lower than the others, ceteris paribus. The reason is not due to low student repayment, but because of taste-based discrimination. On the other hand, for auto loan or other loan, it is easier to get fund successfully. Meanwhile, we also investigate the influence of other factors, such as loan amount and rate.

It has put forward new challenges to investors about how to make effective investment decisions. Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies. In the paper, we calculate the mutual information of every two variables to measure their mutual dependence and build a Bayesian network model to select loans that would pay back with high confidence. We perform abundant experiments on the data from the world's largest P2P lending platform Prosper.com. Experimental results show that the high Bayesian probability, the high return rate. Especially, those Bayesian probability high than 90%, their return rate is 100%. Therefore, Bayesian network model can significantly help investors make better investment decisions.

Considering investors, loans, and their relation (i.e. the actual amount invested), the P2P lending marketplace can be modeled as a bipartite graph, with investors and

loans on either side, represent lent amounts as the many-to-many investment relationship. Therefore, each loan may be funded by multiple lenders, and each lender invests on different loans. Then, we value the lender score by calculating the return rate of his past investment. With these scores and the bid amount, we can access the loan score. To validate the proposed model, we perform extensive experiments on the real-world data from the world's largest P2P lending marketplace. Experimental results on real-world P2P lending data revealed that an investee that gains investment from good investors who have good past investment performance is a more worthwhile to invest. That is, loans with the high loan score are better choices. The bipartite investor model could effectively indicate the investment value and this model can significantly improve the investment performances.

Key words: P2P online lending; Ordinary Least Square; Bayesian Model; Bipartite Investment Model

目录

Contents

第一章 绪论

P2P（Personal to Personal）网络借贷来源于个人对个人的小额借贷，P2P 小额借贷是一种将比较小额度的资金聚集起来借给有资金需求的人的一种商业模式。它的社会价值主要体现在满足个人资金需求、发展个人信用体系以及提高社会闲散资金利用率三个方面，由 2006 年孟加拉国诺贝尔和平奖得主尤努斯教授首次提出。随着互联网技术的迅速发展和普及，P2P 小额借贷逐渐由单一的线下模式转变为线下线上并行模式，并随之产生了 P2P 网络借贷平台。P2P 网络借贷平台是个人通过第三方在收取一定利息的前提下，向其他个人提供小额借贷的金融模式，其客户对象主要有两方面，一是将资金借出的放款人，另一个是需要贷款的借款人。P2P 网络借贷平台通过这种借贷方式来缓解人们由于在不同阶段收入差异而导致的消费力不平衡问题，从而使更多人群享受到了 P2P 小额信贷服务。

1.1 课题的背景和意义

P2P 借贷是指不需要以银行等传统金融机构为中介，借贷双方直接通过网络平台交易的无担保借贷[1-9]。借款人可以以低于银行贷款利息快捷方便借到钱而放款人可以获得高于银行存款利息，P2P 借贷模式很快就在网上流行开。目前，美国最大的网络借贷平台是 Prosper[10]，欧洲最大的网络借贷平台是 Zopa。国内正处于初步发展阶段，出现了宜信、拍拍贷、人人贷等网络借贷平台。

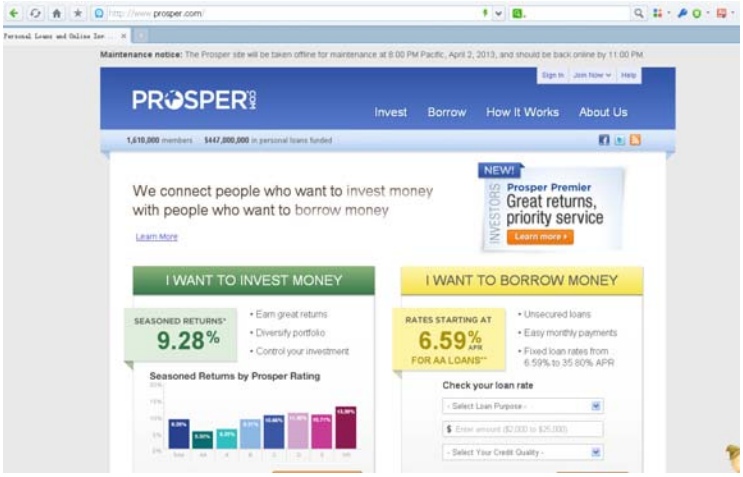

图 1-1 Prosper 网络借贷平台

近几年，P2P借贷大量的交易数据促使了新的研究，例如美国的Prosper.com，超过120万的成员及将近30亿美元交易量为投资分析提供了充分的数据。数据挖掘[11-13]是通过分析每个数据，从海量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示三个步骤。数据准备是从相关的数据源中选取所需要的数据并整合成可用于数据挖掘的数据集；规律寻找是指用某种方法将数据集中所包含的规律找出来；规律表示是尽可能以用户可理解的方式将规律表示出来。数据挖掘的主要任务有关联分析、聚类分析、分类分析、异常分析、演变分析等等。近年来，数据挖掘引起了信息产业界的很大关注，其主要由于是存在大量数据，可以广泛被使用，并且急切需要将这些数据转换成有用的信息和知识。

本课题的研究目标是通过数据挖掘的各种算法分析美国 Prosper 平台上数据，从而得出帮助借款者及放款人的策略。使用统计及各种数据挖掘方法从多个角度发现数据背后的规律，有助于更好地了解网络借贷这一新兴事物，并帮助投资人寻找风险小回报高的投资机会，避免不良借款人运用网络借贷进行欺诈。本文的研究有利于更好的了解网络借款这一新兴金融工具，对国内网络借贷的发展有借鉴意义。

1.2 国内外研究现状

近几年，P2P借贷大量的交易数据促使了新的研究，在这个新的借贷模式下，网络借贷中的社交网络有了广泛的研究。Berger & Gleisne分析了点对点借贷金融中介的角色，提出了借款如果有放款人的推荐可以提高信用，是否有中介对最终的利率起重要作用[14]。Freedman & Jin研究了社交关系是否解决了P2P借贷的信息不对称问题，他们发现由于放款人的知识有限以及由于Prosper消除了组领导者(leader)的回报导致了对组成员投资的预计收益明显低于非组成员[15]。Lin等从结构网络和关系网络两方面研究网络借贷中的社会关系，发现关系网络方面能够比较准确表示借款成功率、利率等[16]。Collier & Hampshire 建立了一个评价和设计社区名誉系统的理论框架[17]。Sergio也建立了一个基于模型的聚类方法，用于测量借款金额在分析评估中的社会交互影响[18]。

为了帮助借款人更好地做投资决策，Luo & Xiong对投资构成进行分析，发现