浙江大学

本科生毕业论文

题目 P2P 借贷平台中的信用评估模型

姓　　名 朱梦莹

学　　号 3100102796

指导教师 郑小林　副教授

专　　业 计算机科学与技术

学　　院 计算机学院

A Dissertation Submitted to Zhejiang University for the Degree of Bachelor of Engineering

TITLE: The Credit Evaluation Model In P2P Lending

Author: Mengying Zhu

Supervisor: Asso.Prof. Xiaolin Zheng

Major: Computer Science and Technology

College: Computer Science and Technology

Submitted Date: 2014-5-30

浙江大学本科生毕业论文（设计）诚信承诺书

1.本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2.本人在毕业论文（设计）中引用他人的观点和参考资料均加以注释和说明。

3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

毕业论文（设计）作者签名：

_____年_____月_____日

摘要

互联网金融中的 P2P 借贷已成为互联网用户的一种新型融资模式。在其蓬勃发展的同时，也相应产生了许多问题。其中，如何评估 P2P 借贷中的个人信用，建立有效的信用体系至关重要。P2P 借贷的信用问题，本质上就是借贷过程中的信息不对称，而网络交易本身存在匿名性和虚拟性使得这一问题更加严重。因此，分析和识别成功的借贷行为对 P2P 借贷的健康发展是有意义的。

鉴于在 P2P 借贷过程中信用的重要性，本文通过对现有的信用评估模型，结合互联网的特点提出一种基于 PCA 降维支持向量机的信用评估算法。该算法从 P2P 借贷的数据中通过数据预处理等手段抽取出训练样本，使用 PCA 降维去噪和支持多分类的 SVM 来学习样本，并对测试样本进行预测。最后，在分类的基础上通过聚类等手段分析影响 P2P 借贷的因素。

通过实验及分析，我们发现借贷信息、财务信息会影响 P2P 借贷，而社会特征可以对信用风险高的用户起到一个很好的缓冲作用。

关键词 信用评估，P2P 借贷，SVM，聚类

Abstract

P2P lending has become a new type of financing mode for Internet users. Though it is booming, it also produces a lot of problems, among which, how to assess the individual credit in P2P lending is important to establish the effective credit system. The credit problem on P2P lending refers to the problem of asymmetric information in the lending process naturally, and the anonymity and virtual which the trade of the network itself has makes this problem more serious. Therefore, borrowing behavior analysis and recognition is meaningful for the healthy development of the P2P lending.

In view of the importance of credit in the P2P in the lending process, we proposed a PCA dimension reduction based support vector machine algorithm combining with the characteristics of Internet. The algorithm has drawn the training samples from the P2P lending data by data preprocessing; using the PCA to reduce the dimension, then uses multi classification SVM learning samples, and predicts the test sample. Finally, on the basis of the classification, we analyses the factors affecting P2P lending by clustering method.

Through the experiments and analysis, we find that credit information, financial information will affect the P2P lending, and social characteristics can play a good buffer function for high credit risk user.

Keywords Credit Evaluation,P2P Lending, SVM, Clusting

目录

第1章 绪论

1.1 课题背景

1.1.1 互联网金融与 P2P 借贷

互联网金融是现今又一热点话题，在 2013 年 10 月 30 日互联网金融全球峰会 IFC1000 在北京召开，主题为“大金融 大数据 大战略”。并且当今互联网金融的主要热点是电子支付、互联网借贷和网络理财，在这些热点话题的背后都少不了数据的支撑。大数据是互联网金融中不可缺少的一个核心组成部分。因为金融没有类似实物的物理生产、仓储、物流等过程，但其本身是数据的生产、仓储、挖掘、传输、分析和集成。所以大数据对于金融而言，相比其他行业，无疑是有更巨大的影响力。

信息技术的进步导致了电子市场的快速增长(Malone et al., 1987)[30]。电子市场最令人印象深刻的特性之一就是它能够消除或减少对传统中间商角色的依赖，使得产品/服务提供商和最终客户可以直接联系(Patsuris,1998)[32]。尤其这样使得在线点对点(P2P)贷款成为互联网用户一种新型融资模式。

P2P 借贷是指无担保贷款，借贷双方通过没有金融机构中介的网络平台进行借贷行为 (Lin et al., 2009a[25]; Collier &Hampshire, 2010[8]; Bachmann et al., 2011[3])。

作为应用信息技术在金融领域和 Web 2.0 的革命(Iyer et al., 2009[19]; Lin et al., 2009b[26])，P2P 借贷能有效地促进信息发布和搜索，并提供所有必要的功能来完成交易(Brown, 2008[6]; Herzenstein et al., 2008[16]).。

这种创新的借贷模型拥有以下优点：

(1)借款人和贷款人可以轻松在网络平台上发布、搜索信息，并且以较低的交易成本完成交易(Lin, 2009[27]; Lin et al., 2009a[25]; Lin et al., 2009b[26]) ；

(2)低交易成本使得非常小的贷款(例如小额贷款)变得可行；

(3)多笔小额贷款可能汇集在一起，组成一个需要大量资金的大型的基金项目，以此规避风险；

(4)通过查看网上认证信息和在社交网络上搜索信息，贷款人可以收集更多关于借款人的信用记录的信息，来减轻借款人和贷款人之间的信息不对称，减

少贷款风险，并使得贷款范围超出了传统的熟人圈子。

网络 P2P 借贷主要针对小额贷款，这不仅为小企业或个人提供创始资金，还为他们提供短期流动资金(Johnson et al., 2010[20]; Wang et al., 2009[38])。尽管 P2P 借贷平台扮演着金融机构的角色，把借款人和贷款人联系起来，但它的利润不是从借款和贷款中获得，而是在交易过程中抽取一定的佣金(Lin, 2009[27])。这个有效的可持续的贷款方式——在线 P2P 借贷已经收到来自学术界和企业越来越多的关注(Brown, 2008[6]; Galloway, 2009[12]; Lin, 2009[27]; Bachmann et al., 2011[3])。

2005 年以来，在线 P2P 借贷在许多国家，包括美国、加拿大、英国、日本、意大利和中国，以不同的形式经历了快速增长时期。一些在线 P2P 借贷平台是出于慈善目的，旨在收集和提供资金给在贫困中的人，而另一些则有商业目的，有意给借款人和贷款人提供便利。最成功的在线平台是英国的 Zopa，美国的 Prosper 和 Kiva。例如，在 2006 年成立的 Prosper 在 2009 年成功完成 1.7 亿美元的贷款。到 2011 年 4 月为止，总部设在旧金山的非盈利组织 Kiva 通过其平台获得的贷款的总额已经达到 2.05 亿美元 (Kiva,2011[21])。

尽管起步较晚，在线 P2P 借贷在中国也有相当大的发展。例如拍拍贷（PPDai.com）、宜信（CreditEasy.com）、齐放（Qifang.com）等网站也具有一定的影响力。国内最大的 P2P 借贷网站拍拍贷成立于 2007 年，在一年半的时间里就积累了超过 8 万名用户，使得这种金融投资新式迅速在国内发展。在三年内，CreditEasy.com 已经在北京和其他 15 个城市从成千上万的私人投资者中吸收近 1 亿美元，变成全国性的 P2P 借贷平台。

1.1.2 P2P 借贷中信用问题

P2P 借贷在蓬勃发展的同时，也相应地产生了许多问题。例如借贷资金的安全仍会存在隐患，借款人的信息可信度如何保证，一笔贷款产生后缺乏行之有效的对于还款的监管手段。这其中，如何评估 P2P 借贷过程中的个人信用，建立一个良好有效的 P2P 借贷信用体系至关重要。

P2P 借贷的安全问题，本质上就是借贷过程中的信息不对称问题。由于多数 P2P 借贷都是信用贷款，所以具有更大的风险。而网络交易本身就存在的匿名性和虚拟性使得这一问题更加严重。借款人的机会主义行为所带来的信息不对称和不信任使得借款人和贷款人存在匹配效率低下的问题。因此，大多数借款人

只进行一次申请贷款，然后就退出了 P2P 借贷平台(Collier &Hampshire,2010[8])。因此，分析和识别成功先例的借贷行为对在线 P2P 借贷的健康发展是有意义的。

1.1.3 信用评估模型

鉴于在 P2P 借贷过程中信用的重要性以及 P2P 借贷的安全问题，国内外建立了许多传统或现代的信用评估模型。

传统信用评估模型通常指银行采用的信用风险评估模型，主要采用信用评分法，即选取一些相关的财务指标根据事先确定的分值表打分加权。例如国内 P2P 借贷平台——拍拍贷就是采用这种方法。这种方法主观性较强，并且存在如下问题：

- 信用评价指标体系不全。
- 用户提供的财务数据和个人信息数据不准确、不充分。
- 信用评估方法简单单一。

而另一些信用评估模型结合了计算机技术，通过获取第三方社交网络、电子商务平台等信息，将信息进行审核汇总，采用一些数据挖掘手段得到信用情况，例如国外的 Zest Finance、国内的聚信立等。但是这样需要获取到多个第三方平台的支持，从而获取他们平台中的数据。

1.2 本文研究目标和内容

鉴于当前研究的局限性，需要进一步的理解在线 P2P 借贷的信用特点，我准备对以下三个方向做出研究:（1)对比中美不同 P2P 借贷平台的信用模型;（2）基于 P2P 借贷平台产生的数据构建信用模型；（3）分析社会资本在 P2P 借贷中起的作用。

- 对比中美不同 P2P 借贷平台的信用模型

P2P 借贷平台的成功很大程度上取决于其创新的商业模式。深入检查不同的在线 P2P 借贷平台的商业模式不仅有助于我们更好地理解在线借贷的本质，而且还提供了洞察这些平台的改进和新业务模式的设计的机会。

- 基于 P2P 借贷平台产生的数据构建信用模型

通过分析不同 P2P 借贷平台的数据，我们可以获得很多有趣的结论。而基于这些结论，我以一些现有的信用模型为基础，构建出一个 P2P 借贷的信用模型。在这其中，我们将会加入社会资本的因素。

- 分析社会资本在 P2P 借贷中起的作用

在文献阅读的过程中，可以发现社会资本在 P2P 借贷的过程中起着复杂而重要的作用，由于这些文献得出的结论不甚相同，所以我希望通过自己的研究在这方面获得一些新的发现和进展。

1.3 本文结构安排

本文共分六个章节，每个章节的研究内容和主要贡献如下：

第 1 章介绍了本文的研究背景、研究目的和意义、研究内容和主要贡献。

第 2 章主要介绍了 P2P 借贷的相关背景，与 P2P 借贷相关的文献，分析了文献中信用评估的常见模型。并且阐述了现有研究中对影响 P2P 借贷因素的一些结论。在相关技术小节阐述了了本文所用到的分类、聚类算法的原理和相关的数学理论。

第 3 章首先提出了本文的研究方案，然后结合互联网的特点提出一种基于 PCA 降维支持向量机的信任算法。该算法从 P2P 借贷的数据中通过数据预处理等手段抽取出训练样本，使用 PCA 降维去噪和支持多分类的 SVM 来学习样本，并对测试样本进行预测。最后，在分类的基础上通过聚类等手段分析影响 P2P 借贷的因素。

第 4 章主要通过拍拍贷和 Prosper 网站上的数据来验证第 3 章中提出的分类和聚类算法，分析算法结果，得到信用评估模型和聚类结果。

第 5 章通过第 4 章的实验结果，分析影响 P2P 借贷的因素，为高信用风险提出提升个人信用的建议。

第 6 章对本文的研究内容进行总结，并对未来的工作进行展望。

第2章 文献综述

2.1 P2P 借贷市场

个人贷款的概念并不是一个新的商业模式，而一种传统的方式即在没有任何中介的情况下以个人名义借钱的人(Everett, 2008[10]; Herrero- Lopez, 2009[15])。是什么使转移到互联网平台的在线 P2P 借款成为一个新现象。

2.1.1 P2P 借贷平台

在 2005 年第一个借贷平台 Zopa 成立于欧洲(英国)。此后，各种形式的借贷平台开始形成。Garman et al. (2008)[13]阐述了在世界范围内现有的 24 个借贷平台，仅在美国就有 12 个，在 P2P-Banking.com 的博客中称 2010 年全世界 33 个不同的借贷平台。

在美国第一个借贷平台于 2006 年 2 月成立(prosper.com)。如今大多数现有的借贷平台工作在国家层面上，因为不同国家有不同的法律要求(Berger &Gleisner, 2009[4])。下面的表显示了现有的主要借贷平台：

在线 P2P 借贷平台存在不同类型。他们基本上可以分为两种类型：商业和非商业(Ashta &Assadi, 2009)[2]。而商业平台一般仅限于国内市场，非商业性平台通常在全球范围内运作。两种平台之间的主要区别是贷款人的总体意图和他的期望回报。贷方使用商业平台希望在一个合理的风险下获得利益。在非商业平台贷款他们愿意承担的风险但并不一定要获得利益。这里贷款人更希望“捐赠”小额贷款给世界上经济欠发达的地区。

2.1.2 用户角色

在线 P2P 借贷是一个双面的市场，与传统的银行系统并没有太多区别(Klafft,2008)[22]。贷款人和借款人是平台的所有活动的主要目标群体。

因此大部分的研究都集中在这些利益相关者和影响贷款成功的决定因素上(Freedman &G.Z. Jin, 2008[11]; Iyer et al., 2009[19])。贷款人在给定的风险水平下尽可能寻求有利可图的投资机，借款人在一定的违约风险下寻找不同的资金来源。P2P 网站作为中介机构，将这些人组织在一起。他们试图匹配双方的期望。借款人和贷款人有时参与一些可以提现他们共同利益的组织和社区(M. E.

Greiner &Wang, 2009[14]; Herrero-Lopez, 2009[15])。

2.1.3 信用机构、合作银行、监管机构

作为一个(小)金融市场的一部分，P2P 借贷也要受到不同国家的不同的监管限制。根据国家规定，有现有的银行合作伙伴是主要要求。几篇文章提到了银行介入的必要性(Galloway,2009[12])，但这主要是为了促进借贷的过程。这个过程还包括确认借款人信用评级的信用部门或其他外部监测机构的参与。这些系统的确认和识别也因国家而异，导致全球研究在这一领域不适用。

2.1.4 借贷流程

一些平台直接连接贷款人和借款人，而其他平台通过第三方(通常是银行)连接，例如拍拍贷平台，图 2-1。

图 2-1 拍拍贷平台架构图

在线 P2P 借贷平台在借款人的利率设置方式上略有不同。例如 prosper.com 和拍拍贷都是使用拍卖竞价(Galloway,2009[12])，如图 2-2，借款人可以设定一个他们愿意支付的最高利率。在有限的时间内(prosper 上是 14 天)贷款人可以不断提交他们愿意出的金额和他们接受的最低利率。甚至当贷款金额满足借款人所需金额后，贷款人仍然可以提交 bid 来降低其他贷款人提供的最低利率，增加自己愿意出的贷款金额。在这种情况下，bids 的总金额已经比需要的贷款金额更多，那么那些给出最低利率的 bid 将会被采纳。然后所有借款人会收到此时的最高的利率和它的贷款金额，即使最低利率已经变小。

图 2-2 拍拍贷借款流程

如果贷款过程产生一个成功的贷款资助，一些平台像 prosper.com 实现了借款人的支付能力的核查，包括稳定收入的确认，然后才授予借款人的贷款，并且最终启动还款过程(S. Garman, R. Hampshire, et al., 2008[13])。

在线 P2P 借贷平台的收入是通过服务费获得的，他们从借款人和贷款人处获得(Klafft,2008)[22]。大多平台会收取一定比例的贷款资金来让借款人关闭交易，以及后期违约的费用。借款人通常需要支付基于放贷金额一定比例的服务费。

2.2 信用评估模型

鉴于在 P2P 借贷中，信用的重要性，不少平台和研究都利用大量的数据建立了信用评估模型。这里主要分为两类，一类是基于传统银行业的信用评估模型，另一类是基于数据挖掘的信用风险预测。

传统信用评估模型主要指采取银行的信用评级方法，包括以 5C 法为代表的专家判断法和以 5C 法为基础的综合评价法。而这些都需要评估人员根据自己的专业技能、主观判断对影响信用评估决策的某些因素进行权衡。在此基础上，统计模型的引入使得模型的估计和使用相对比较简单，比较容易得到一致的评级结果，如 Logit 模型、Probit 模型。

现代数据挖掘技术被采用来建立信用评分模型(Huang et al., 2007[18])。基于数据挖掘的信用风险预测通常基于借贷的相关属性，例如用户个人信息、财务信息、历史交易信息等，将借贷结果分到“好”和“坏”的债务类别中(Lim and Sohn, 2007[24])。与主观的传统方法相比，自动信用评估模型提供了很多好处(Rosenberg and Gleit, 1994[34];Thomas et al, 2002[35]; Blöchlinger and Leippold, 2006[5])：

- 在信用评价过程中降低成本，并且使不良贷款的预期风险降低；

- 基于客观信息的一致性建议，从而消除了人类的偏见和成见；
- 将变化的经济或政策纳入评价系统中；
- 信用评分模型的性能可以监视，跟踪，并随时调整。

下表 2-1 为近期研究中采用的方法：

Reference	Benchmark methods
Hoffmann et al (2007) [17]	Fuzzy rules, ANN, C4.5, Bayes, LDA
Huang et al (2007) [18]	SVM
Liu et al (2008) [28]	LDA
Marinakis et al (2008) [31]	KNN
Wang and Huang (2009) [37]	C4.5, ANN, KNN, SVM, naive Bayes, LDA
Mahmoud et al (2010) [30]	C4.5
Vukovic et al (2012) [36]	KNN
Danenas and Garsva (2012)[9]	SVM
Hanhai Zhou et al(2013)[40]	SVM-KNN
Zhiwang Zhang et al(2014)[39]	SVM, fuzzy SVM

表 2-1 2007-2014 年研究方法的总结

2.3 影响借贷的因素

在传统借贷的背景下，金融机构，如商业银行，充当了交易中介的角色。这些银行以较低的利率吸收存款，然后以更高的利率向客户发放贷款。由于银行使用了复杂的风险评估机制，并且了解更多的借款人信息，他们可以在贷款过程中更有效地缓解信息不对称。相比之下，在网络 P2P 借贷的环境下，贷款人很难对借款人获得全面的信息，导致信息不对称的问题十分严重(Lin et al., 2009a[25])。因此，大多数研究在线 P2P 借贷都集中在以缓解借款人和贷款人之间的信息不对称，从而在贷款过程中减少风险为目的，包括：(1)“硬信用信息”对贷款结果的影响，如个人认证信息、财务信息；(2)“软信用信息”对贷款结果的影响，如社交信息。

2.3.1 借贷流程影响贷款的信贷“硬信用信息”

“硬信用信息”指的是可以准确量化、容易存储、可以有效传播的信用信

息。在 P2P 借贷中，信用信息包括借款人的信用背景，如借款人的负债收入比、信用评级、过去获得的信贷金额和借款人持有信用卡的数量(Lin, 2009[27]; Lin et al., 2009a[25]; Lin et al., 2009b[26])。

国外由于信用制度相对健全，信用信息容易获取，因此国外大多数 P2P 借贷平台都会要求借款人提供自己的财务情况，并且以此作为判断借款人信用的主要指标。典型的财务特征包括：信用评级，每月的详细收入和支出，房子所有权以及债务收入比等。这些信息往往由收集个人信息和财务数据的外部评级机构给出。而国内也会要求用户提供自己的个人认证信息和财产认证信息。

在线 P2P 借贷，因为贷款人无法获得有关借款人的详细信息，贷款人必须依赖于可用来判断借款人的信誉的信号，并相应地做出贷款决定。研究表明，存在两个在贷款的决策中发挥着举足轻重作用的特征：获得信号的成本和信号难度评估(Collier &Hampshire, 2010[8])。在 P2P 借贷中，借款人的个人信息和贷款 listing 上的信息被认为是评估借款人的可信度的重要的信号，用来评估借款人的违约风险和设置利率(Collier &Hampshire, 2010[8]; Lin, 2009[27])。

这里我主要研究的是以下可以获取的“硬信用信息”：

- 借款者的信用评级：就如同传统银行一样，对借款利率影响最大的因素为借款人的信用评级，而借款人的债务收入比的影响虽然显著，但是影响却小得多。通过分析从 Prosper.com 收集的数据，Lin (2009)[27]发现信用评级较低的贷款请求不太可能被资助，并且这样的贷款请求更有可能违约或以很高的利率结束。虽然在中国没有结论性的结果关于贷款信用评级对贷款结果的影响，但陈(2012)[7]表面，信用评级在 Ppdai.com 对获得贷款的概率产生了部分影响，而利率是决定因素。然而，违约率较高的借款人信用水平要低得多。
- 借款人的财务信息：Freedman & Jin (2008)[11]在他们的研究中发现，由用户提供的就业状况和职业对贷款结果也有一定的影响。
- 借贷信息：研究表明，贷款的成功率与利率负相关。在实践过程中，借款人必须权衡这两个因素之间的关系。而交易的进度也会影响贷款人的判断：快要达到贷款目标的借贷有更多机会获得资助，而不是更高的利率。然而，Lin 等人(2009b)[26]和 Puro 等人(2010)[33]指出交易进度在对违约率的影响上并没有显著差异。此外，贷款的目的也会对贷款人的决定产生影响：商业贷款比债务合并贷款获得成功的几率高，并且可以得

到更高的利率(Wang et al., 2009[38])。Collier and Hampshire (2010[13]) 发现，贷款规模、借款人的财务状况(例如负债收入比)和交易进度都对利率产生影响。

总之，在这一领域的研究仍然有限。目前的研究主要是使用从 Prosper.com 收集的数据集(如 Lin et al., 2009a[25])，但这很难将他们的结论普遍化。不同的借贷平台可能采用不同的信贷“硬信用信息”，以及“硬信用信息”对贷款结果的影响需要进一步研究在其他网站的上下文中被研究。

2.3.2 社会特征和“软信用信息”

与“硬信用信息”如信用评分或借款人的财务状况相比，“软信用信息”是指借款人的模糊的无法量化的信息。在 P2P 借贷中，软信用信息可能来自借款人的社交网络，例如网络上的“朋友”，网络上的“团体”，借款列表中添加的照片等(Collier &Hampshire, 2010[8]; Iyer et al., 2009[19]; Krumme &Herrero, 2009[23]; Lopez, 2009[29])。

在线 P2P 借贷平台不仅披露了借款人的个人贷款信息，还提供借款人的社交信息。使用 Web 2.0 技术，涉及 P2P 借贷的贷款人可以很容易地从借款人的社交网络中获取软信用信息(Lin, 2009[27])。小额信贷理论表明，社交网络可以帮助减少贷款过程中的信息不对称，并且可以激励借款人偿还贷款(Krumme &Herrero 2009[23])。社交网络的作用也适用于在线 P2P 借贷的上下文(Lin et al., 2009b[26])。

社会资本对充分获得贷款起着积极的影响，可以减少借款人的可能获得的利率，并且对信用评级较低的借款人产生越来越大的影响(M. E. Greiner &Wang, 2009[14])。根据 Herrero-Lopez(2009) [15]的研究表明，当金融功能并不足以构建一个成功的贷款请求时，培养社交功能可以增加了贷款的机会。

而在这里，我将研究以下两个方面对贷款结果的影响：

- 朋友：“朋友”实际上是通过 P2P 借贷平台联系的人们。他们代表一个一对一的链接从自身到其他借款人或贷款人。这种关系通常是基于家庭、友谊或先前的交易。这种联系公开激励属于借款人第二或更高维度的社交网络中的贷款人基于间接信任给予 bid(Herrero-Lopez,2009[15])。Freedman &Jin (2008)[11]发现有贷款推荐人或是 bid 是由借款人的朋友提交的，很少会拖欠还款，并且有更高的回报率。他们的结论是，借款

人的朋友能更好地识别风险和拥有可信赖性，因为他们拥有更多的额外信息。而且他们认为社交网络的监控为还款提供了更强大的动力。

- 组织与社区：大部分 P2P-Lending 平台都允许用户形成特殊的社区。如果组织存在积极的向导和激励，那么组织可以清除一些信息障碍(Freedman &G.Z. Jin, 2008)[11]。例如成为一个被信任的组织的成员在 Prosper 可以获得更高概率的贷款请求(Herrero-Lopez,2009[15])。但加入一个可信组织并不能完全保证获得贷款，这仍然需要一个合理的报价。Berger (2009)[4]和 Greiner 与 Wang (2009)[14]发现，仅仅加入一个组织就可以显著减少资金的贷款利率。Greiner 与 Wang (2009) [14], Herrero-Lopez (2009)[15] 和 Freedman 与 Jin (2008)[11]声称，拥有社交的贷款比没有朋友关系或不是组织成员的贷款更有可能获得资助。

2.4 相关技术

2.4.1 分类算法

分类是数据挖掘的重要方法之一，它可以从内容丰富、蕴藏大量信息的数据集中提取描述重要数据类的模型，用于做出智能的信用评估决策。分类的目的是学会一个分类函数或分类模型（也称为分类器），该模型能把数据集中的数据项映射到给定类别中的某一个类别。分类可用于预测，分类的输出是离散的类别值。

2.4.1.1 分类的定义

定义 3-1 给定一个数据集$D\{t_1, t_2, \dots, t_n\}$和一组类$C = \{C_1, C_2, \dots, C_m\}$,分类问题就是确定一个映射 $f: D \rightarrow C$,每个元组t_i被分配到一个类中。一个类C_j包含映射到该类中的所有元组，即$C_j = \{t_i | f(t_i) = C_j, 1 \le i \le n, t_i \in D\}$

从上面的定义中，我们把分类看作是从数据集到一组类别的映射，而且这些类是被事先定义的，非交叠的。

2.4.1.2 分类的步骤

分类一般分为如下两个步骤：

第一步，建立模型，描述预定的数据类集或概念集。通过分析由属性描述的数据库元组来构造模型。假定每个元组属于一个预定义的类，由一个称作类

标号属性的属性确定。对于分类，数据元组也称为样本、实例或对象。为建立模型而被分析的数据元组形成训练数据集。训练数据集中的单个元组称为训练样本，并随机的由样本群选取。由于提供了每个训练样本的类标号，该步也称为有指导的学习(即模型的学习在被告知每个训练样本属于哪个类的指导下进行)。它不同于无指导的学习(如聚类)，那里每个训练样本的类标号是未知的，要学习的类集合或数量也可能事先不知道。通常，学习模型用分类规则、决策树或数学公式的形式提供。例如，给定一个借贷信用信息的数据集，可以学习分类规则，根据他们的信用结果优或良来预测信用风险(见图 3-1)。这些规则可以用来为以后的数据样本分类，也能对数据集的内容提供更好的理解。

第二步，使用模型进行分类(见图 3-1)。首先评估模型(或分类方法)的预测准确率。测试样本随机选取，并独立于训练样本。模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比。对于每个测试样本，将已知的类标号与该样本的学习模型类预测比较。如果认为模型的准确率可以接受，就可以用它对类标号未知的数据元组或对象进行分类(这种数据在机器学习文献中也称为未知的或先前未见到的数据)。例如，在图 3-3 中通过分析现有借贷数据学习得到的分类规则可以用来预测新的借贷的信用风险。

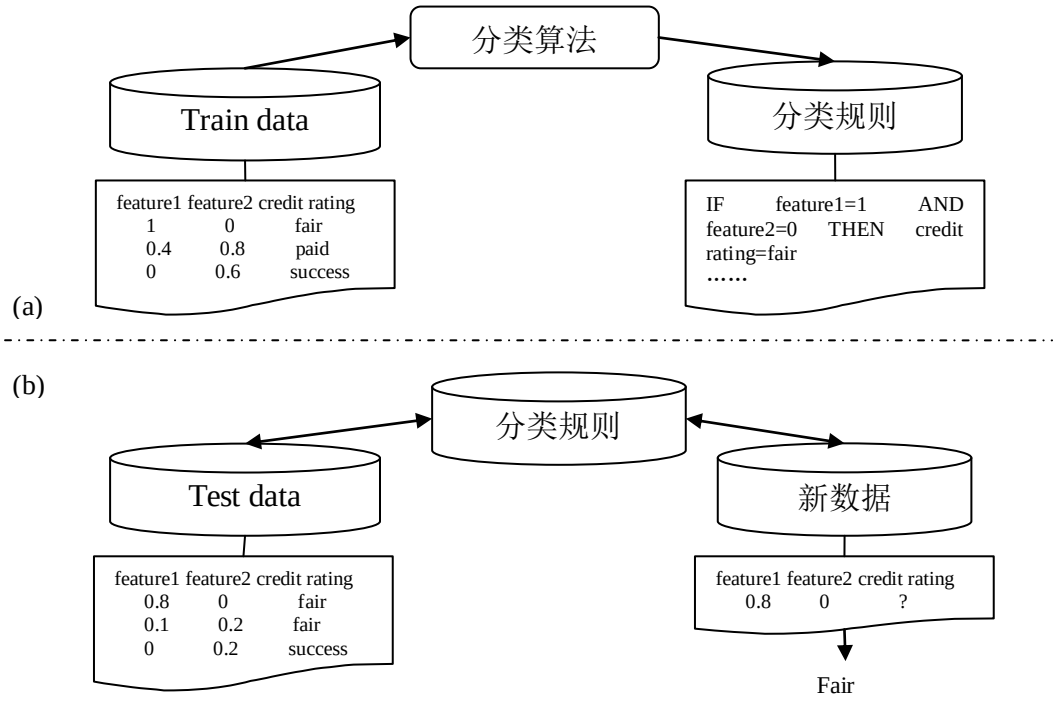

图 3-3 数据分类过程：(a)学习：用分类算法分析训练数据。这里，类标号属性是 credit rating，学习模型或分类法以分类规则形式提供。(b)分类：测试数据用于评估分类规则的准确率。如果准确率是可以接受的，则规则用于新的数据元组分类。

2.4.1.3 分类器的构造方法

分类器的构造方法主要有统计方法、机器学习方法、神经网络方法等。统计方法主要包括贝叶斯法和非参数法；机器学习方法主要包括决策树和逻辑回归法；神经网络方法主要是 BP 算法。本文将着重介绍以下几种分类方法：逻辑回归分类方法、基于距离的分类方法、决策树分类方法、SVM 分类方法。

2.4.1.4 分类方法的评估标准

分类方法有很多种，它们各有所长，也各有所短。通过对它们进行研究和比较评估，一来可以扬长避短，在特定情况下使用最优的分类方法，二来可以对其加以改进，使其性能得到优化。

以二分类为例，一个完美的分类模型就是，如果一个客户实际上(Actual)属于类别 good，也预测成(Predicted)good，处于类别 bad，也就预测成 bad。但从上面我们看到，一些实际上是 good 的客户，根据我们的模型，却预测他为 bad，对一些原本是 bad 的客户，却预测他为 good。我们需要知道，这个模型到底预测对了多少，预测错了多少，分类矩阵就把所有这些信息，都归到一个表里。我们首先可以得到一个分类矩阵如图 3-4，FP 和 FN 就是我们常说的第一类错误与第二类错误，以这四个基本指标可以衍生出多个分类器评价指标。

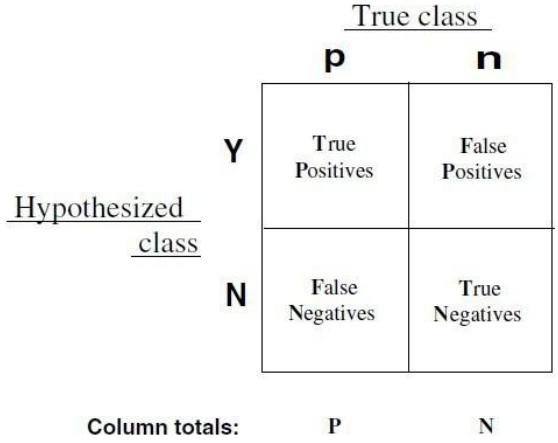

图3-4 分类矩阵

常见的分类指标大致有如下几个：

- $fp\ rate = \frac{FP}{N}$
- $tp\ rate = \frac{TP}{N}$
- $precison = \frac{TP}{TP+FP}$
- $recall = \frac{TP}{P}$
- $accuracy = \frac{TP+TN}{P+N}$
- $F\_measure = \frac{2}{1/precision+1/recall}$

2.4.1.5 逻辑回归分类

Logistic regression 是一种常见的机器学习手段，可以用来回归，也可以用来分类，主要是二分类。而多分类通常有两种类型，一种是有序分类，一种是无序分类。以二分类的逻辑回归模型为例，假设我们的样本是$\{\boldsymbol{x}, y\}$，y 是 0 或者 1，表示正类或者负类，$\boldsymbol{x}$ 是我们的 m 维的样本特征向量。那么这个样本 $\boldsymbol{x}$ 属于正类，也就是 $y=1$ 的“概率”可以通过下面的逻辑函数来表示：

$$p(y = 1|\boldsymbol{x}; \theta) = \sigma(\theta^T \boldsymbol{x}) = \frac{1}{1+exp\ (-\theta^T \boldsymbol{x})} \quad (2.1)$$

这里θ是模型参数，也就是回归系数，σ是 sigmoid 函数。实际上函数(2.1)是由下面的对数几率（也就是 $\boldsymbol{x}$ 属于正类的可能性和负类的可能性的比值的对数）变换得到的：

$$log\ it(x) = ln(\frac{P(y=1|x)}{P(y=0|x)}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m \quad (2.2)$$

Logistic Regression 最基本的学习算法是最大似然。假设我们有 n 个独立的训练样本$\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}, y = \{0,1\}$。那每一个观察到的样本$(\boldsymbol{x_i}, y_i)$出现的概率是：

$$P(y_i, x_i) = P(y_i = 1|x_i)^{y_i}(1 - P(y_i = 1|x_i))^{1-y_i} \quad (2.3)$$

则整个样本集，也就是 n 个独立的样本出现的似然函数为（因为每个样本都是独立的，所以 n 个样本出现的概率就是他们各自出现的概率相乘）：

$$L(\theta) = \prod P(y_i = 1|x_i)^{y_i}(1 - P(y_i = 1|x_i))^{1-y_i} \quad (2.4)$$

最大似然法就是求模型中使得似然函数最大的系数取值 θ。这个最大似然就是代价函数（cost function）。然后使用梯度下降方法可以求得 θ。

在多分类中，有序分类的 logistic 回归可以采用比例优势模型（proportional odds model），又称累积 Logistic 模型或累积比数模型。无序分类的 logistic 回归采用多项 logistic 模型（polynomial Logistic model）。

这两种模型的分析目的是不同的。对于无序分类的多项 logistic 模型，其分

析结果是以其中一类作为参照，其余各类均与参照类比较。例如有“ABC”三类，以“C”作为参照类，则采用多项 Logistic 模型的结果有两个：一是“A”相对“C”的结果，二是“B”相对“C”的结果。

对于有序分类的 logistic 回归，则会体现出“累积”（cumulative）的含义，它也会出现两个结果，但是与多项 Logistic 模型不同，一是“A+B”相对“C”的结果，二是“A”相对“B+C”的结果。

2.4.1.6 基于距离的分类方法及 KNN 算法

K 最邻近结点算法（k Nearest Neighbors,简称 KNN）是实际运用中经常被采用的一种基于距离的分类算法。KNN 算法的基本思想：假定每个类包含多个训练数据，且每个训练数据都有一个唯一的类别标记，计算每个训练数据到待分类元组的距离，取和待分类元组距离最近的 *k* 个训练数据，*k* 个数据中哪个类别的训练数据占多数，则待分类元组就属于哪个类别。

2.4.1.7 决策树分类

一般来说，决策树的构造主要由两个阶段组成:第一阶段，生成树阶段。选取部分受训数据建立决策树，决策树是按广度优先建立直到每个叶节点包括相同的类标记为止。第二阶段，决策树修剪阶段。用剩余数据检验决策树，如果所建立的决策树不能正确回答所研究的问题，我们要对决策树进行修剪直到建立一棵正确的决策树。这样在决策树每个内部节点处进行属性值的比较，在叶节点得到结论。从根节点到叶节点的一条路径就对应着一条规则，整棵决策树就对应着一组表达式规则。

为了对未知的样本分类，样本的属性值在决策树上测试。路径由根到存放该样本预测的叶节点。决策树容易转换成分类规则。

2.4.1.8 SVM 算法分类

SVM 是从线性可分情况下的最优分类面发展而来的，SVM 的基本思想是:通过某种非线性映射,将输入向量 **x** 映射到一个高维的特征空间,在这个高维的特征空间 Z 中,构造最优分离超平面。可用图 3-5 的两维情况说明。图中，实心点和空心点代表两类样本，H 为分类线，B1、B2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔(margin)。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为 0)，而且使分类间

隔最大。

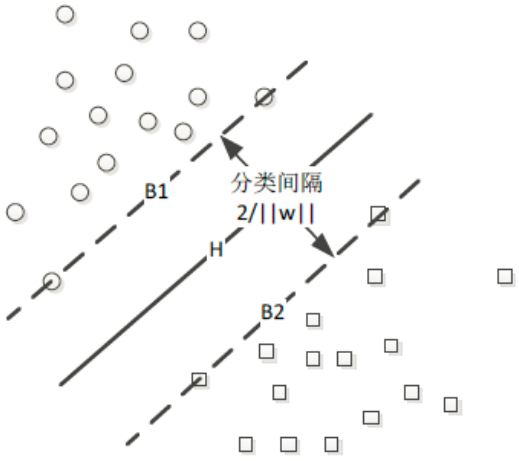

图 3-5 最大间隔超平面

设线性可分样本集为$(x_i, y_i), i = 1, \dots, n, x \in R^d, y \in \{-1, +1\}$是类别符号。d 维空间中线性判别函数的一般形式为$g(x) = w \cdot x + b$，分类线方程$w \cdot x + b = 0$，将判别函数进行归一化，使两类所有样本都满足$|g(x)| \geq 1$，即使离分类面最近的样本的$|g(x)| = 1$，此时分类间隔等于$2/\|w\|$，因此使间隔最大等价于使$\|w\|$(或$\|w\|^2$)最小。要求分类线对所有样本正确分类，就是要求它满足

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1,2, \dots, n \quad (2.5)$$

满足条件(2.5)且使$\|w\|^2$最小的分类面就叫做最优分类面，B1、B2 上的训练样本点就称作支持向量。利用 Lagrange 优化方法可以把上述最优分类面问题转化为其对偶问题，即：在约束条件$\sum_{i=1}^{n} y_i \alpha_i = 0 \ \alpha_i \geq 0, i = 1,2, \dots, n$下对$\alpha_i$求解下列函数的最大值：

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (2.6)$$

(α_i为原问题中与每个约束条件(1)对应的 Lagrange 乘子)。若α^*为最优解，则$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$即最优分类面的权系数向量是训练样本训练的线性组合。这是一个不等式约束下二次函数极值问题，存在唯一解。容易证明，解中将只有一部分(通常是很少一部分) α_i不为零，对应的样本就是支持向量。解上述问题后得到的最优分类函数是：

$$f(x) = sgn\{(w \cdot x) + b\} = sgn\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\} \quad (2.7)$$

式中的求和实际上只对支持向量进行(因为非支持向量对应的α_i均为 0)。b是分类阈值，可以用任一个支持向量(满足式(2.5)中的等号)求得，或通过两类中任意～对支持向量取中值求得。

对非线性问题，可以通过非线性变换转化为某个高维空间中的线性问题，在变换空间求最优分类面。这种变换是利用一种特殊的思路实现的。注意到在上面的对偶问题中，不论是寻优目标函数(2.6)还是分类函数(2.7)都只涉及训练样本之间的内积运算。设有非线性映射Φ将输入空间的样本映射到高维(可能是无穷维)的特征空间 H 中。当在特征空间 H 中构造最优超平面时，训练算法仅使用空间中的点积，即$\Phi(x_i)\Phi(x_j)$，而没有单独的$\Phi(x_i)$出现。因此，如果能够找到一个函数 K 使得$K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$，这样，在高维空间实际上只需进行内积运算，且这种内积运算是可以用原空间中的函数实现的，甚至没有必要知道变换 的形式。根据泛函的有关理论，只要一种核函数$K(x_i,x_j)$满足 Mercer 条件，它就对应某一变换空间中的内积。因此，在最优分类面中采用适当的内积函数$K(x_i,x_j)$就可以实现某一非线性变换后的线性分类，而计算复杂度却没有增加，此时目标函数(6)变为：

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \quad (2.8)$$

而相应的分类函数也变为：

$$f(x)=sgn\{\sum_{i=1}^{n}\alpha_i^{*} y_i K(x_i,x_j)+b^{*}\} \quad (2.9)$$

而算法的其它条件决不变，这就是支持向量机。

2.4.2 降维方法

在数据挖掘过程中，高维数据是非常棘手的研究对象。之所以使用降维后的数据表示是因为：

- 在原始的高维空间中，包含有冗余信息以及噪音信息，降低了准确率；而通过降维,我们希望减少冗余信息所造成的误差,提高识别的精度。
- 希望通过降维算法来寻找数据内部的本质结构特征。
- 通过降维来加速后续计算的速度

降维大致有两大类别，一类是从原始维度中提取新的维度，例如主成分分析或因子分析。另一类是从原始维度中选择一些子集，即称为特征选择。

2.4.2.1 基于特征选择的降维方法

特征选择本质上继承了 Occam's razor 的思想，从一组特征中选出一些最有效的特征，使构造出来的模型更好。并且这样进行过特征选择，可以有效的剔除不必要的特征，降低之后获取数据集的难度。

对于数值自变量而言，可以使用两样本 t-test 考察因变量取一种值时自变量

的均值与因变量取另一种值时自变量的均值是否相等，然后选择那些检验结果显著（不相等）的自变量；并且因变量为分类变量当因变量为分类变量时，可以将其取值两两配对，针对每对取值进行上述 t-test，然后选择那些对因变量的任何一对取值检验结果显著的自变量。

t-test，又称 student 假设检验方法，用来检验每一类抽样样本是否符合同一高斯分布。

样本 t-test 的过程如下：

（1）配对设计：首先将所有特征中的第 w（$w=1...k$）个特征y_w分成 m 份，分别在 m 份中随机抽取 n 份样本，分别为$y_w = \{y_{1w}, y_{2w}, \dots, y_{nw}\}$。我们的目的是检验 n 份样本是否服从同一正太分布。即检测在方差相同的情况下，不同样本的所服从正太分布的均值是否相等。

（2）检验步骤：

(1)建立虚无假设

H0:$\mu 1 = \mu 2$，假定两个总体平均数之间没有显著差异；

H1:$\mu 1 \neq \mu 2$，假定两个总体平均数之间有显著差异；

(2)计算统计量 t 值，要评断两组样本平均数之间的差异程度，其统计量 t 值的计算公式为：

$$t_w = \frac{\overline{Y_{1w}} - \overline{Y_{2w}}}{\sqrt{\frac{\sum y_{1w}^2 + \sum y_{2k}^2}{n_{1w} + n_{2w} - 2} \times \frac{n_{1w} + n_{2w}}{n_{1w} \times n_{2w}}}} \tag{2.10}$$

(3)根据自由度 *degree of freedom(df)=n-1*，查 t 值分布表，找出规定的理论 t 值，理论值差异的显著性水平设为 0.05 级。显著水平理论值记为$t_{df}0.05$。

(4)比较计算得到的t_w值和理论 t 值，推断虚无假设发生的概率。

(5)根据是以上分析，做出最终判断，选择造成分布差异显著的特征。

2.4.2.2 基于特征变换的降维方法

特征变换即从原始维度中提取新的维度，这种方法可以尽可能保留原始数据中特征之间的关系，从而起到提高模型准确度，降低数据存储空间的作用。

在这之中，Principal Component Analysis(PCA)是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较

多的原数据点的特性。

主成分分析是一种较早发展起来的线性维数缩减方法(Hotelling,1933)。

在主成分分析（PCA）中，方差视为变量的信息的度量。原始变量的“总变化”（信息）。在主成分求法中有两种常见的求法：特征值分解和奇异值分解。这里我采用的是奇异值分解，也就是最小二乘的有化解。

2.4.3 聚类

聚类分析是数据挖掘的一项重要功能，而聚类算法是数据挖掘研究领域中一个非常活跃的研究课题。聚类是把一组对象按照相似性归成若干类别，即“物以类聚”。它的目的是使得属于同一类别的对象之间的距离尽可能的小，而不同类别的对象间的距离尽可能的大。聚类的定义可以如下表示：

定义 3-2 聚类分析的输入可以用一组有序对(X, s)或(X, d)表示，这里 X 表示一组样本，s 和 d 分别是度量样本间相似度和相异度（比如距离）的标准。聚类分析的输出是一个分区，若$C = \{C_1, C_2, \dots, C_k\}$,其中$C_1(i = 1,2, \dots, k)$是 X 的子集，如下所示：

$$C_1 \cup C_2 \cup \dots \cup C_k = X \tag{2.11}$$

$$C_1 \cap C_2 = \emptyset, i \neq j \tag{2.12}$$

C 中的成员$C_1, C_2, \dots, C_k$称为类，每一个类都是通过一些特征描述的，通常有如下集中表示方式：

（1） 通过类的中心或类的边界点表示一个类。

（2） 使用聚类树中的结点图形化地表示一个类。

（3） 使用样本属性的逻辑表达式表示类。

聚类分析就是使用聚类算法来发现有意义的聚类，它的主要依据是把相似的样本归为一类，而把差异大的样本区分开来，这样所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，而与其他聚类中的对象彼此相异。在信用评估中可以把一个簇中的数据对象当作一个类型的整体来对待。作为一个数据挖掘的功能，聚类分析能作为一个独立的工具来获得数据分布的情况，观察每个簇的特点，集中对特定的聚类做进一步的分析。

2.5 本章小结

本章主要介绍了 P2P 借贷的相关背景，与 P2P 借贷相关的文献，分析了文

献中信用评估的常见模型。并且阐述了现有研究中对影响 P2P 借贷因素的一些结论。我们发现，虽然在线 P2P 借贷在美国和中国有不同的操作模式，但“硬”和“软”信用信息在这两个国家都可能影响贷款的结果。

在相关技术部分阐述了本文所用到的分类、聚类算法的原理和相关的数学理论。首先提出了本文的研究方案，接着阐述了分类算法的定义、一般步骤、评价指标，然后简要介绍了基本的分类算法的原理和算法步骤，如 KNN、决策树、线性回归、SVM；同时也介绍了特征降维方法，如基于特征选取的 t-test 算法、基于特征变换的 PCA 算法。最后，介绍了如何确定最佳聚类数的算法、高斯混合聚类算法。

第3章 研究方案

3.1 概述

本文主要研究是基于中美两个 P2P Lending 网站 Prosper 和拍拍贷的数据，对其进行数据挖掘，建立一个信用评估模型，然后依据信用评估模型的结果进行聚类分析，提出应对信用风险的建议（图 3-1）。

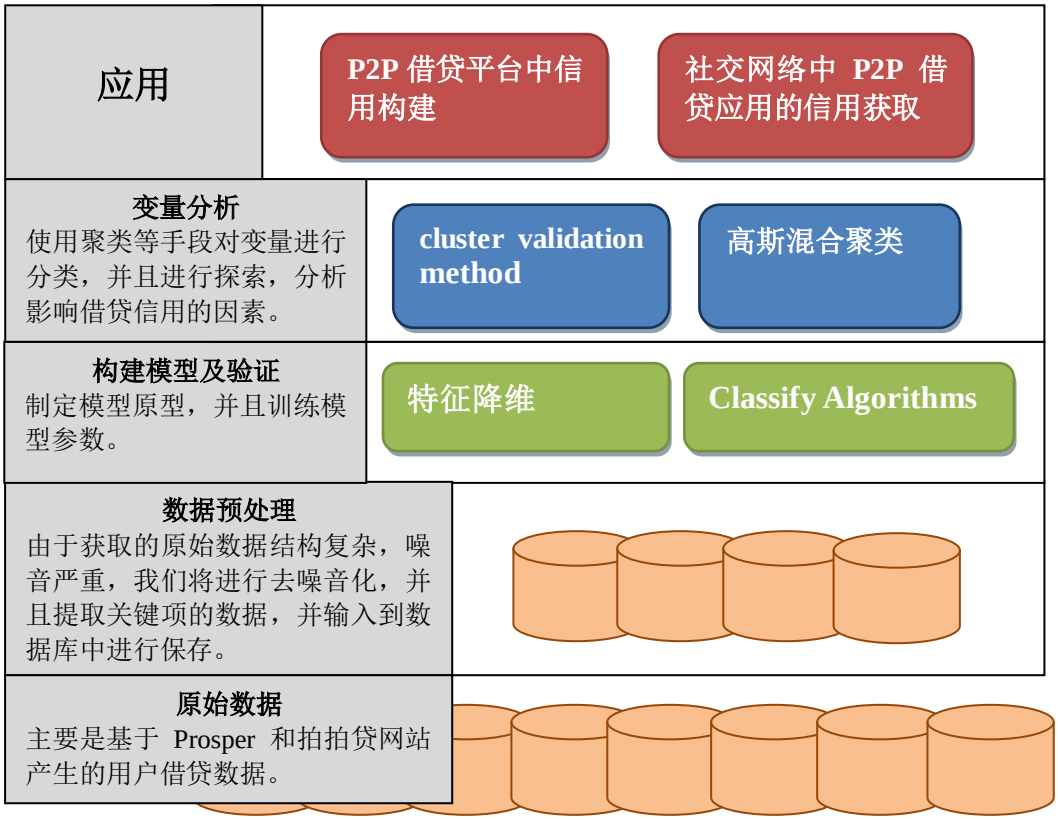

图 3-1 研究框架图

在具体的研究方法方面，首先要查找相关的文献资料，了解在线 P2P 借贷的研究进展和现有的互联网信用模型。这些资料不仅可以帮助了解整个信用模型的组成部分和构成，更能够从中吸取前人的经验教训。然后会选取若干文献，关于要做的项目所要用到的技术都要重点关注。这里最主要使用到的技术是根据 SVM、KNN 等分类方法构建信用模型以及用聚类的方法获得主要影响 P2P 借贷的因素。

3.2 基于分类算法的信用评估模型

分类应用在 P2P 借贷的信用评估模型中，可以有效的对信用风险进行分级。在这里，我的信用评估模型基本流程如下图 3-2：

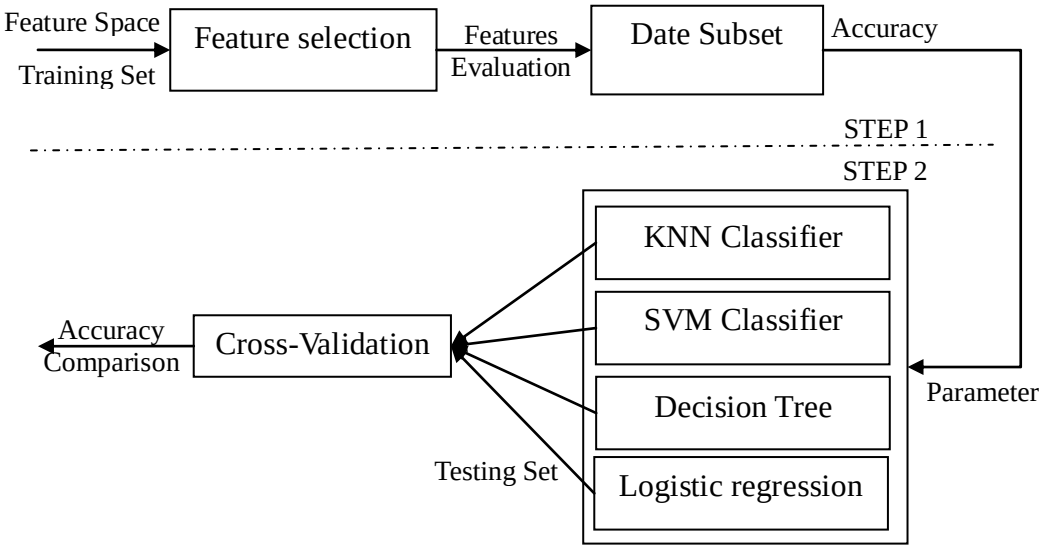

图 3-2 基于降维分类的信用评估模型基本流程图

3.2.1 变量相关性

通过对实验数据集的分析，我可以得到拍拍贷和 Propser 数据集中的特征之间相关性并不强（图 3-6），并且与分类标号的相关性也不强（图 3-3 中第一列），所以我可以假设有些变量并不能很好的反应数据的本质结构，会对模型造成影响。因此，使用降维算法对数据集降维后，可以有效去除或转换相关性不强的特征，从而提高分类算法的准确率。

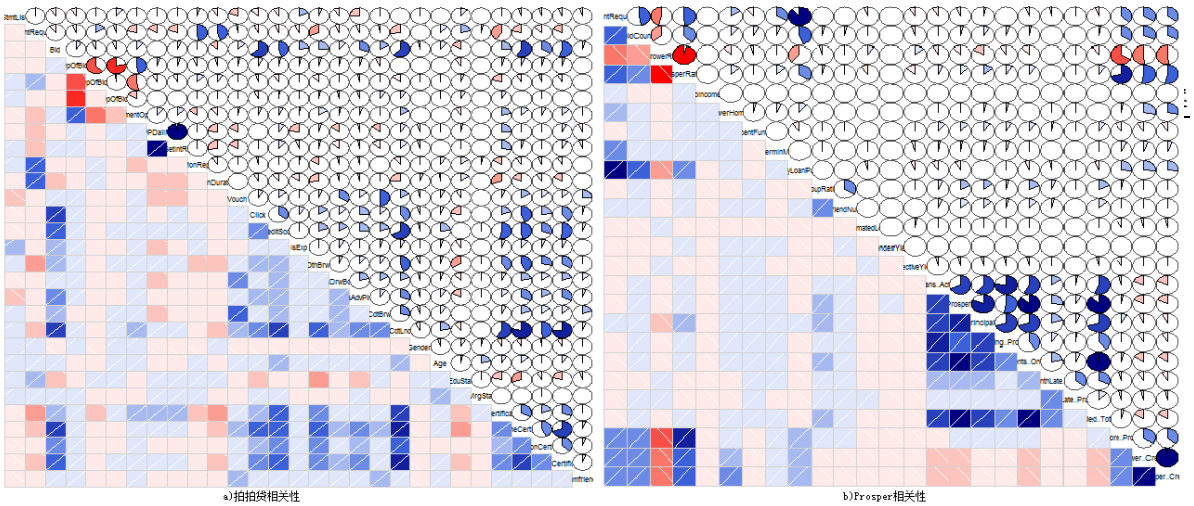

图 3-6 特征相关性分析：(a)为拍拍贷特征相关性；(b)为 Prosper 特征相关性；面板下半部分斜线的方向将相关性分成正相关和负相关两类。同时蓝色代表正相关，粉色代表负相关。颜色越深，涂色面积越大，意味着相关性越强。

3.2.2 基于 PCA 降维的多分类 SVM 分类算法

在这一节中，将介绍基于 PCA 降维的多分类支持向量机 PCAM-SVM。首先简单给出了 PCA 降维的基本步骤，然后介绍了多分类的支持向量机，最终提出 PCAM-SVM 算法。

3.2.2.1 PCA 降维算法

PCA 算法是对一组相关变量$X_1, \dots, X_r$采用按照方差减少排序的正交线性投影来构造降维的数据集。

$$Z_i = b_{i1}X_1 + b_{i2}X_2 + b_{iq}X_q \ (i = 1, \dots, q) \tag{3.1}$$

其中$b_i = (b_{i1}, \dots, b_{iq})'$为对应的投影向量。$Z_i$称为第 i 个主成分。

奇异值分解方法是把主成分看作在 q 维空间上对 p 维空间点的最优线性近似。

设主成分 $\boldsymbol{Z}$ 和原始数据 $\boldsymbol{X}$ 满足 $\boldsymbol{Z}=\boldsymbol{BX}$，$\boldsymbol{B}$ 是$q \times p$阶矩阵。$\boldsymbol{Z}$ 作为 $\boldsymbol{X}$ 的线性近似，两者还满足 $\boldsymbol{X}=\boldsymbol{AZ}$，$\boldsymbol{A}$ 是$p \times q$阶矩阵。寻找这些参数，最小化：

$$\mathrm{E}\{(\mathrm{X} - \mu - \mathrm{ABX})^{\mathrm{T}}(\mathrm{X} - \mu - \mathrm{ABX})\} \tag{3.2}$$

这就是个最小二乘问题。

最小化的结果要求寻找一个正交矩阵 $\boldsymbol{AB}$，记作$\mathrm{H_q}$，$\mathrm{H_q}$为投影矩阵，把每个点$\mathrm{X_i}$投影到 q 维结构$\mathrm{H_qX_i}$ ，这是由$\mathrm{V_q}$的列限定的$\mathrm{X_i}$的正交投影。投影矩阵$\mathrm{H_q}$可以分解为$\mathrm{V_qV_q'}$，$\mathrm{V_q}$为$p \times q$矩阵。求解的过程可以由奇异值分解完成。对$n \times q$矩阵 $\boldsymbol{X}$，有奇异值分解 X=UDV′，对每一个 q，最小二乘的解$\mathrm{V_q}$包含了 $\boldsymbol{V}$ 的前 q 列，$\boldsymbol{UD}$

的列称为$\boldsymbol{X}$的主成分。这种方法是来自减秩回归的观点（reduced rank regression）。

3.2.2.2 多分类的支持向量机

标准的支持向量机一开始是针对二类别分类问题提出来的，怎样将利用它进行多类别的分类是当前支持向量机研究领域的一个热门问题。支持向量机多分类算法的研究思路总体上来说有两种：一个是直接在所有样本数据上求解一个复杂的二次规划问题，使一个支持向量机在学习结束后就能将多类分开。这种方式比较直接，但计算方式很复杂，效率不高。另外一种思路是结合多个标准的二分类支持向量机来区分多类，常见的方法有 one-against-one 和 one-against-all 两种。这种方式的可操作空间比较大，但是要使用多次二分类 SVM 算法，训练时间比较长。

- 一对多法（one-versus-rest,简称 OVR SVMs）。

  训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。分类时将未知样本分类为具有最大分类函数值的那类。

  假如我有四类要划分（也就是 4 个 Label），他们是 A、B、C、D。于是我在抽取训练集的时候，分别抽取 A 所对应的向量作为正集，B,C,D 所对应的向量作为负集；B 所对应的向量作为正集，A,C，D 所对应的向量作为负集；C 所对应的向量作为正集， A,B,D 所对应的向量作为负集；D 所对应的向量作为正集，A,B,C 所对应的向量作为负集，这四个训练集分别进行训练，然后的得到四个训练结果文件，在测试的时候，把对应的测试向量分别利用这四个训练结果文件进行测试，最后每个测试都有一个结果 f1(x),f2(x),f3(x),f4(x).于是最终的结果便是这四个值中最大的一个。然而这种方法有缺陷，因为训练集是 1:M，这种情况下存在 biased，因而不是很实用。

- 一对一法（one-versus-one,简称 OVO SVMs）。

  其做法是在任意两类样本之间设计一个 SVM，因此 k 个类别的样本就需要设计 k(k-1)/2 个 SVM。当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。

  还是假设有四类 A,B,C,D 四类。在训练的时候我选择 A,B; A,C; A,D; B,C; B,D;C,D 所对应的向量作为训练集，然后得到六个训练结果，在测试的时候，把对应的向量分别对六个结果进行测试，然后采取投票形式，最后得到一组

结果。

我们的算法就是基于 OVO SVMs 设计的多分类 SVM 算法。

3.2.2.3 PCAM-SVM 算法

综上所述，本算法训练的基本过程如图 3-7 所示。

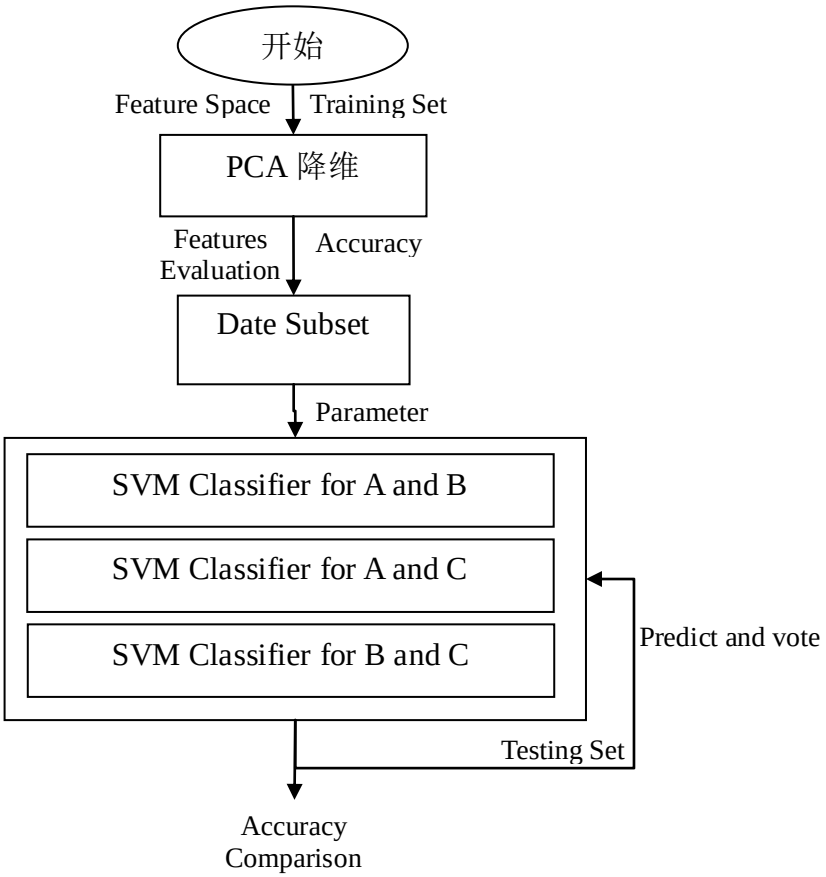

图 3-2 PCAM-SVM 算法流程图

算法的具体描述如下：

- 将特征空间 F 进行 PCA 降维转换成降维后的特征空间F_{pca}。
- 提取训练样本的特征空间$TrainF_{pca}$和测试样本的特征空间$TestF_{pca}$。
- 训练 SVM。分别构建 k(k-1)/2 个 SVM Classifier，使用训练样本进行学习，优化分类器参数。
- 使用 SVM 预测测试样本，获得 accuracy、precision 和 recall

3.3 基于高斯聚类分析影响 P2P 借贷的因素

对于高斯混合模型聚类，首先需要知道聚类的个数。诚然，对于少量数据，

可以根据数据分散的结果大致估计聚类的个数，但是在对大量数据聚类的时候，我们就需要一个可以自动估计聚类个数的指标。这里我们采用 Silhouette(Sil)指标来确定最佳聚类个数。

3.3.1.1 确定最佳聚类数的算法

确定聚类算法最佳聚类数的基本算法思想是：针对具体的数据集，在确定的聚类数搜索范围内，运行聚类算法产生不同聚类数目的聚类结果，选择合适的有效性指标对聚类结果进行评估，根据评估结果确定最佳聚类数。

具体算法如下：

(1)选择聚类数的搜索范围$[k_{min}, k_{max}]$，通常取$k_{min} = 2, k_{max} = int(\sqrt{n})$。

(2)For $k = k_{min}$ to k_{max}

①随机选取k个初始聚类中心Z^k；

②运用 K-means 聚类算法，更新计算成员关系矩阵和聚类中心Z^k；

③检查终止条件，如不满足，则转向② ；

④利用聚类结果计算有效性指标值，转向(2)。

(3)比较各有效性指标值，有效性指标值达到最优所对应的 k 即为最佳聚类数k_{opt}

(4)输出聚类结果：类中心点Z_{opt}，成员关系矩阵U_{opt}，最佳聚类数k_{opt}。

聚类有效性是指评价聚类结果的质量并确定最适合特定数据集的划分。通常采用聚类有效性指标来评价聚类算法产生的哪个聚类结果是最优的，并将最优的聚类结果所对应的聚类数目作为最佳聚类数。目前已经提出了一些检验聚类有效性的函数指标，性能较优的指标主要有 Calinski-Harabasz(CH)指标、Weighted inter-intra(Wint)指标、In-Group Proportion(IGP)指标和 Silhouette(Sil)指标等，其中 Silhouette 指标以其简单易用和良好的评价能力而得到广泛应用。设 a(i)为样本 i 与类内所有其他样本的平均距离，b(i)为样本 i 到其他每个类中样本平均距离的最小值。Silhouette 指标定义为：

$$Sil(i) = \frac{b(i)-a(i)}{max\,(a(i),b(i))} \tag{3.3}$$

Silhouette 指标反映了聚类结构的类内紧密性和类间分离性，既可用于评价聚类质量，也可用于估计最佳聚类数。silhouette 指标的值在[-1,1]范围内变动，所有样本的平均 Silhouette 指标值越大表示聚类质量越好，其最大值对应的类数为最佳聚类数。

3.3.1.2 高斯混合聚类

每个 Gaussian Mixture Model 由 *K* 个 Gaussian 分布组成，每个 Gaussian 称为一个“*Component*”，这些 *Component* 线性加成在一起就组成了 GMM 的概率密度函数：

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sum_k) \tag{3.4}$$

现在假设我们有 *N* 个数据点，并假设它们服从某个分布（记作$p(x)$），现在要确定里面的一些参数的值，例如，在 GMM 中，我们就需要确定π_k、μ_k和$\sum_k$这些参数。我们的想法是，找到这样一组参数，它所确定的概率分布生成这些给定的数据点的概率最大，而这个概率实际上就等于$\prod_{i=1}^{N} p(x_i)$，我们把这个乘积称作似然函数 (Likelihood Function)。通常单个点的概率都很小，许多很小的数字相乘起来在计算机里很容易造成浮点数下溢，因此我们通常会对其取对数，把乘积变成加和$\sum_{i=1}^{N} log\, p(x_i)$，得到 log-likelihood function 。接下来我们只要将这个函数最大化（通常的做法是求导并令导数等于零，然后解方程），亦即找到这样一组参数值，它让似然函数取得最大值，我们就认为这是最合适的参数，这样就完成了参数估计的过程。具体算法如下：

(1) 估计数据由每个 *Component* 生成的概率：对于每个数据x_i来说，它由第 *k* 个 *Component* 生成的概率为

$$\gamma(i,k) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \sum_j)} \tag{3.5}$$

由于式子里的μ_k和$\sum_k$也是需要我们估计的值，我们采用迭代法，在计算$\gamma(i,k)$的时候我们假定μ_k和$\sum_k$均已知，我们将取上一次迭代所得的值（或者初始值）。

(2) 估计每个Component的参数：现在我们假设上一步中得到的$\gamma(\mathrm{i},\mathrm{k})$就是正确的“数据$x_\mathrm{i}$由$\mathrm{Component_k}$生成的概率”，亦可以当做该 Component在生成这个数据上所做的贡献，或者说，我们可以看作x_i这个值其中有$\gamma(\mathrm{i},\mathrm{k})x_\mathrm{i}$这部分是由$\mathrm{Component_k}$所生成的。集中考虑所有的数据点，现在实际上可以看作 Component 生成了$\gamma(1,\mathrm{k})x_1, \dots, \gamma(\mathrm{N},\mathrm{k})x_\mathrm{N}$这些点。由于每个Component都是一个标准的 Gaussian 分布，可以很容易分布求出最大似然所对应的参数值：

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i \tag{3.6}$$

$$\sum_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T \tag{3.7}$$

其中 $N_k = \sum_{i=1}^{N} \gamma(i,k)$，并且$\pi_k$也顺理成章地可以估计为$N_k/N$。

(3) 重复迭代前面两步，直到似然函数的值收敛为止。

我们可以看到 GMM 和 K-means 的迭代求解法其实非常相似，因此也有和 K-means 同样的问题——并不能保证总是能取到全局最优，如果运气比较差，取到不好的初始值，就有可能得到很差的结果。对于 K-means 的情况，我们通常是重复一定次数然后取最好的结果，不过 GMM 每一次迭代的计算量比 K-means 要大许多，我们的做法是先用 K-means （已经重复并取最优值了）得到一个粗略的结果，然后将其作为初值（只要将 K-means 所得的 centroids 传入 GMM 函数即可），再用 GMM 进行细致迭代。

3.3.1.3 使用高斯聚类算法分析影响 P2P 借贷的因素

聚类算法可以把一组对象按照相似性归成若干类别，即“物以类聚”。这样的分类可以分析出数据集的内在结构以及特征之间的关系。

我们采用高斯混合聚类，这种方法假设每一组都满足一个高斯分布，这样可以得到每个组的估计密度，并且在聚类上采取的是一个“软” 分类。

在聚类算法完成后，我们采用将每一组特征的均值都可视化，将影响 P2P 借贷的积极因素和消极因素用不同颜色表示，这样可以直观清晰的看出每一组的特征。从而分析各种因素对 P2P 借贷的影响。

3.4 本章小结

本章首先提出了本文的研究方案，然后结合互联网的特点提出一种基于 PCA 降维支持向量机的信任算法。该算法从 P2P 借贷的数据中通过数据预处理等手段抽取出训练样本，使用 PCA 降维去噪和支持多分类的 SVM 来学习样本，并对测试样本进行预测。最后，在分类的基础上通过聚类等手段分析影响 P2P 借贷的因素。

第4章 实验结果

4.1 数据描述

P2P 借贷网站的数据集提供给研究者，用来促进研究者对其商业模式的理解。尽管不同的 P2P 借贷社交网络可能使用不同的命名约定，但是相关细节在概念上是相似的。

4.1.1 拍拍贷数据

4.1.1.1 数据结构

拍拍贷数据主要分为 5 种表：PPDai List Info, PPDai User Info, PPDai Frd Info, PPDai Bid Info 和 PPDai List Hostiry Info。在本文中，List Info，User Info，Frd Info 表将会被结合使用，重新构造一个新的数据集结构。

首先简要介绍一下这几张表的用途：

- List Info：借方创建 Listing 来募集贷方提供的资金。Listing 中会说明借方的情况，以及募集资金的原因。当足够多的资金被募集，Listing 就会变成一笔借贷。
- User Info：在拍拍贷网站上注册的用户。User 可能拥有一个或多个身份，并且可能拥有一个或多个朋友。
- Frd Info：用户的朋友信息，包含用户和他的朋友 ID、创建时间等。

4.1.1.2 数据处理

拍拍贷数据集包含 24,125 Members，52,901 Listings。

同样为了分析 P2P 借贷数据，本文剔除了部分不相关的特征变量并且数值化了非数值的特征变量。有关时间、地理信息、个人描述等特征在本文中被忽略。4 种验证情况（照片、手机、文凭、视频）被数值化成 0 和 1（No 和 Yes），数值越大验证情况越多。

最终，经过预处理的拍拍贷数据集的特征如下描述：

- AmountRequested：投标金额。
- BidN：投标者数量。

- TypOfBid：投标模式：其中 1 为竞标投标；2 为线下投标；3 为友情投标。
- PaymentOption：还款方式：其中 0 为每月还款；1 为到期还款
- PPDaiIR：拍拍贷利率。
- SetIntRt：借款人设定的投标利率，投标人可以通过竞标压低该利率。
- MonRep：每月归还额度。
- LoanDuration：归还期限，单位是月。
- Vouch：担保状态，其中 0 为无担保；1 为个人担保；2 为系统担保。
- Click#：点击次数。
- CreditGrade：信用等级，1-A；2-B；3-C； 4-D； 5-E； 6-HR。
- IsExp：是否体验，0-no；1-yes。
- OthBrw：是否他处借款，1-yes； 2- no。
- IsDrwBck：是否取现，0-no；1-yes。
- IsAdvPln：是否优先计划，0-no； 1-yes。
- CdtBrw：借入者信用。
- CdtLnd：借出者信用。
- Gender：性别，1-男；2-女
- Age：年龄。
- EduStat：学历，1-小学；2-中学；3-中专；4-高中；5-大专；6-本科；7-硕士研究生；8-博士研究生；9-博士后。
- MrgStat：婚姻状况，0-未知；1-未婚； 2-已婚。
- IDVertification：身份证验证，0-无；1-有。
- CellphoneCertification：电话验证，0-无；1-有。
- EducationCertification：学历验证，0-无；1-有。
- VedioCertification：视频验证，0-无；1-有。
- NumFriend：朋友数量。

4.1.2 Prosper 数据

4.1.2.1 数据结构

PROSPER 数据主要分为 8 种表：Category, Credit Profile, Listing, Loan, Group, Loan，Performance, Marketplace 和 Member（图 4-1 所示）。在本文中，Group，

Member，Listing，Loan 表将会被结合使用，重新构造一个新的数据集结构。

图4-1 Prosper的P2P借贷关系模型

首先简要介绍一下这几张表的用途：

- Group：一种拥有共同兴趣的 Member 组成的社交团队。Group 由创建者建立，经由创建者同意可以加入。每个 Group 的评分是由它的 Member 根据借贷情况评价得来，而不是根据 Member 的信用等级评分的。所以，一个信用等级低的 Member 也可以根据他的表现进入一个评分高的 Group。评分范围是 1-5 分。
- Member：在 P2P 借贷网站上注册的用户。Member 可能拥有一个或多个身份。一系列 Members 拥有共同兴趣或者友好关系共同组建 Group。
- Listing：借方创建 Listing 来募集贷方提供的资金。Listing 中会说明借方的情况，以及募集资金的原因。当足够多的资金被募集，Listing 就会变成一笔借贷。
- Loan：当 Listing 募集到足够的资金后，Listing 将会转变成 Loan。Loan 中说明了借方的还款情况。

4.1.2.2 数据处理

Prosper 数据集包含了从 2005 年 11 月开始的所有交易记录和用户信息。由于 Prosper 平台自创建起经过多次调整改变，所以我截取了 2012 年 1 月至 2013 年 10 月的所有数据，包含 41,628 Members，46,033 Listings，135 Groups。

为了分析 P2P 借贷数据，本文剔除了部分不相关的特征变量并且数值化了

非数值的特征变量。有关时间、地理信息、个人描述等特征在本文中被忽略。而信用等级（AA-E）被转换成数值（7-1），同样地，Group Rating（0-5 Stars）是否拥有房产（No，Yes）被转换成数值（0,1）、Prosper Rating 等特征也被转换成了相应的数值。Listing 的描述信息虽然和借贷结果有一定的关联，但是由于提取文本不便，这里本文也忽略了。

最终，经过前期处理的 PROSPER 数据集的结构如下描述：

- GroupRating：借方所在 Group 的评分。
- FriendNum：借方拥有的朋友数量。
- AmountRequested：借款金额。
- BidCount：贷方提供贷款的总个数。
- BorrowerRate：如果 listing 变成贷款的话，借方希望的贷率。
- ProsperRating：借方的信用等级。范围在 7（best）-1（worst）。
- DebtToIncomeRatio：借方的贷款收入比。
- IsBorrowerHomeowner：借方是否拥有房产。
- PercentFunded：借方募集到的贷款的比例。
- LoanTermInMonths： 还款时间，单位为月。
- MonthlyLoanPayment：每个月归还的金额。
- EstimatedLoss：估计损失
- LenderYield：贷款人收益率
- EffectiveYield：实际收益率
- ActiveProsperLoans：所有活跃借贷记录的个数。
- TotalProsperLoans：所有借贷记录的个数。
- ProsperPrincipalBorrowed：Prosper 未偿借贷。
- ProsperPrincipalOutstanding：Prosper 未偿本金。
- OnTimeProsperPayments：按时还款的记录数量。
- ProsperPaymentsLessThanOneMonthLate：还款中拖欠 1 个月以内的记录数量。
- ProsperPaymentsOneMonthPlusLate：还款中拖欠还款 1 个月以上的记录数量。
- TotalProsperPaymentsBilled：所有历史还款记录数量。
- CreditScoreRangeLower：信用分数最低范围。

- CreditScoreRangeUpper：信用分数最高范围。
- ProsperScore：Prosper 内部的信用分数。

4.1.3 数据集比较

表 4-1 对比了拍拍贷和 Prosper 两个数据集的特征分别属于哪些影响 P2P 借贷的因素。

对比项目	拍拍贷	Prosper
投标情况	AmountRequested	AmountRequested
	Bid#	BidCount
	TypOfBid	PercentFunded
	PaymentOption	LoanTermInMonths
	MonRep	MonthlyLoanPayment
	LoanDuration	
	Vouch	
	Click#	
	IsExp	
	OthBrw	
	IsDrwBck	
	IsAdvPln	
利率	PPDaiIR	BorrowerRate
	SetIntRt	
借款人信用	CdtBrw	ProsperRating
	CdtLnd	ProsperScore
	CreditGrade	CreditScoreRangeLower
		CreditScoreRangeUpper
借款者个人信息	Gender	IsBorrowerHomeowner
	Age	DebtToIncome
	EduStat	
	MrgStat	
	IDVertification	
	CellphoneCertification	
	EducationCertification	
	VedioCertification	

社会信息	NumFriend	FriendNum
		GroupRating
历史借贷		EstimatedLoss
		LenderYield
		EffectiveYield
		ActiveProsperLoans
		TotalProsperLoans
		ProsperPrincipalBorrowed
		ProsperPrincipalOutstanding
		OnTimeProsperPayments
		ProsperPayments LessThanOneMonthLate
		ProsperPaymentsOneMonthPlusLate
		TotalProsperPaymentsBilled

表 4-1 拍拍贷和 Prosper 数据集中特征的分类对比

4.2 信用评估模型

4.2.1 分类算法

首先，我们将数据集分为训练样本和测试样本，这里训练样本取整个数据集的 70%，而 30%为测试样本。接着我们分别使用了 KNN、SVM、决策树、逻辑回归等方法对训练样本进行学习，并对测试样本进行预测。

4.2.1.1 参数训练

首先，我们以所有 Prosper 和拍拍贷数据集随机取 70%作为训练样本，30%作为测试样本。

接着，我们训练了 KNN 算法的 K 值。在图 4-3 中，我们发现对于拍拍贷数据集 K=3 时，KNN 算法的准确率最高，而在图 4-4 中 Prosper 数据集 K=15 时，KNN 算法的准确率最高。

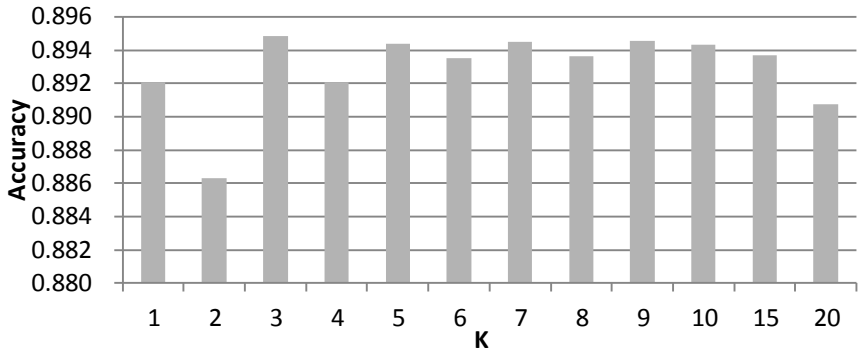

图 4-3 在拍拍贷数据集上 KNN 算法随 k 不同准确率变化图

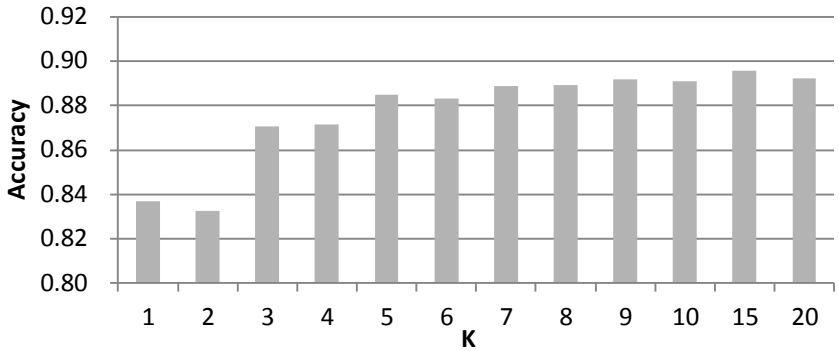

图 4-4 在 Prosper 数据集上 KNN 算法随 k 不同准确率变化图

然后使用了 SVM、决策树、逻辑回归对训练样本进行学习。

4.2.1.2 Accuracy、Precision 和 Recall

当我们得到所有模型的最优参数后，我们可以比较每一个模型的预测表现。由于两个数据集的分类结果都有三类：未成功募集资金（流标）、成功募集资金（成功）、成功还清债务（还清），所以这里我们假设分类矩阵如表 4-2，而 accuracy、precision 和 recall 的求法如下：

- $precison = \frac{TP}{TP+TN+TM} * 2 + \frac{MM}{MM+MN}$
- $recall = \frac{TP}{P} * 2 + \frac{MM}{MM+FM}$
- $accuracy = \frac{TP+MM+FN}{P+N+M}$
- $F\_measure = \frac{2}{1/precision+1/recall}$

	流标 N	成功 M	还清 P
流标 F	FN	FM	FP
成功 M	MN	MM	MP
还清 T	TN	TM	TP

表4-2 分类矩阵

这里表 4-3 显示了在拍拍贷数据集各个分类模型的 accuracy、precision 和 recall。这其中，SVM 算法无论从 accuracy 还是从 precision、F_measure 来看，都是最优的；而决策树算法在 recall 是最优，比 SVM 提高了 1.5%。在 accuracy 方面，SVM 算法比逻辑回归、决策树、KNN 分别提高了 11.6%，3.0%，0.8%；在 precision 方面，SVM 算法比逻辑回归、决策树、KNN 分别提高了 13.6%，9.2%，2.7%；而在 recall 方面，SVM 算法比 KNN、逻辑回归分别提高了 2.0%，1.1%；在 F_measure 方面，SVM 算法比逻辑回归、决策树、KNN 分别提高了 7.4%，3.9%，2.5%

算法	特征数	Accuracy%	Precision%	Recall%	F_measure%
Logistic regression	28	80.8	80.7	89.6	84.9
Decision tree	28	87.6	84.0	92.0	87.8
KNN	28	89.5	89.3	88.8	89.0
SVM	28	90.2	91.7	90.6	91.2

表 4-3 在拍拍贷数据集上各分类算法的 Accuracy、Precision 和 Recall

而表 4-4 则说明了在 Prosper 数据集上各个分类模型的 accuracy、precision 和 recall。同样，SVM 算法无论从 accuracy 还是从 recall 来看，都是最优的；而决策树算法在 precision 是最优，比 SVM 提高了 2.1%。在 accuracy 方面，SVM 算法比逻辑回归、KNN、决策树分别提高了 1.5%，0.1%，0.02%；在 precision 方面，SVM 算法比逻辑回归、KNN 分别提高了 5.9%，0.8%；而在 recall 方面，SVM 算法比逻辑回归、KNN、决策树分别提高了 18.1%，2.2%，0.2%。

算法	特征数	Accuracy%	Precision%	Recall%	F_measure%
Logistic regression	25	88.4	74.7	33.6	46.3

Decision tree	25	89.7	80.8	41.2	54.6
KNN	25	89.6	78.5	40.6	53.5
SVM	25	89.7	79.1	41.8	54.7

表 4-4 在 Prosper 数据集上各分类算法的 Accuracy、Precision 和 Recall

4.2.2 降维算法

首先，我们通过 t-test 方法对数据在不同特征上的分布进行一致性校验，并根据拒绝虚无假设所犯错误概率的 p 值对上下文对用户的评分产生影响的程度进行排序，当 p 值越小时，拒绝虚无假设“$\mu_1 = \mu_2$，在同一特征中抽取的样本符合同一分布”所犯错误越小，则特征对分类结果产生的影响越大。为便于观察我们画出 P 值分布图如下图 4-5 和图 4-6：

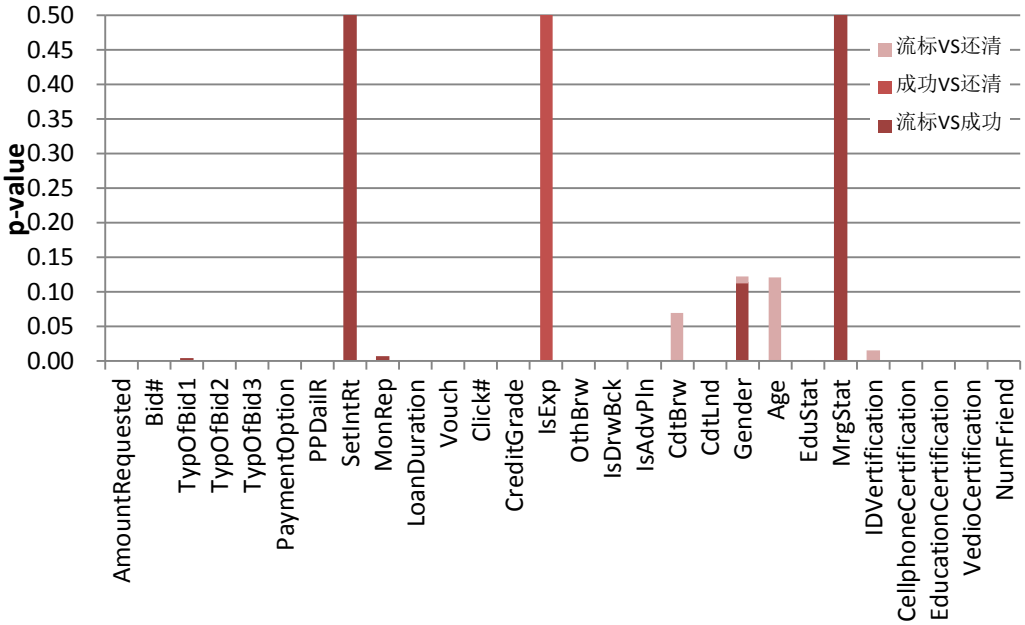

图 4-5 在拍拍贷数据集上 t-test 方法的特征的 p-value 值

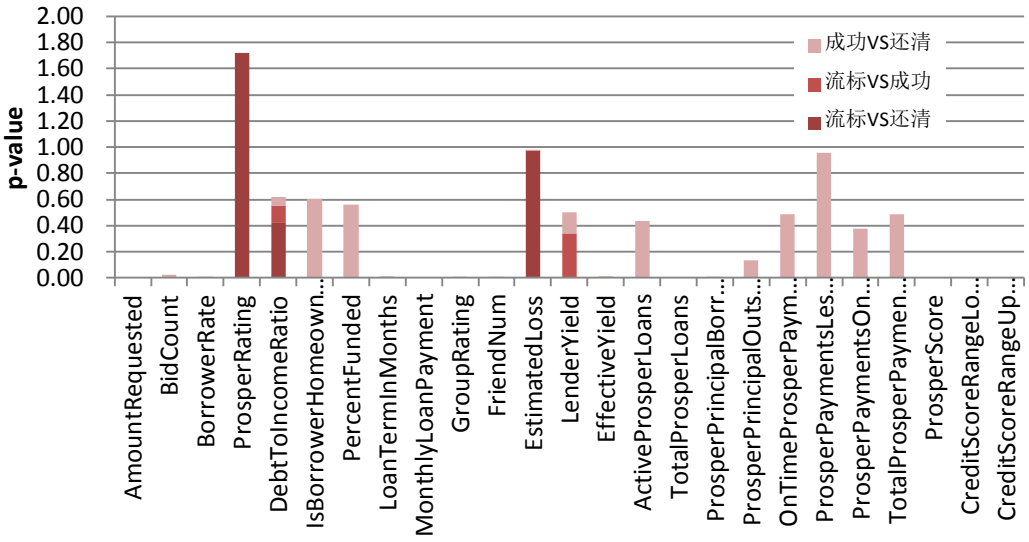

图 4-6 在 Prosper 数据集上 t-test 方法的特征的 p-value 值

从图中，我们分别为拍拍贷和 Prosper 选取了 p-value 值最小的 13 和 13 个特征项，作为分类算法的子数据集。

然后，我们也是使用了 PCA 对数据进行降维，图 4-7 表示了对于拍拍贷数据，PCA 降维以后各个新特征值的 Standard deviations。当我们选取其中 Standard deviations 值最大的 16 个特征值时，占据了 85%以上，如图 4-9(a)；图 4-8 表示了对于 Prosper 数据，PCA 降维以后各个新特征值的 Standard deviations。当我们选取其中 Standard deviations 值最大的 16 个特征值时，占据了 85%以上，如图 4-9(b)。

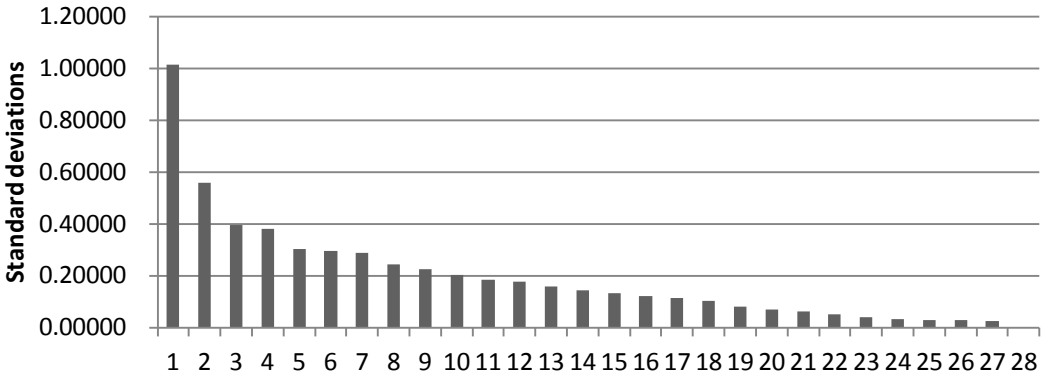

图 4-7 在拍拍贷数据集上 PCA 算法的特征的 Standard deviations 值

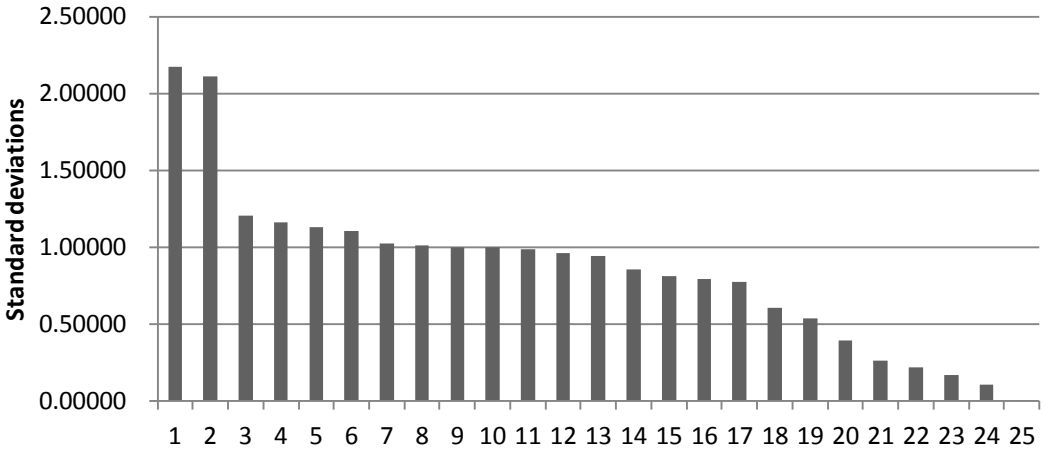

图 4-8 在 Prosper 数据集上 PCA 算法的特征的 Standard deviations 值

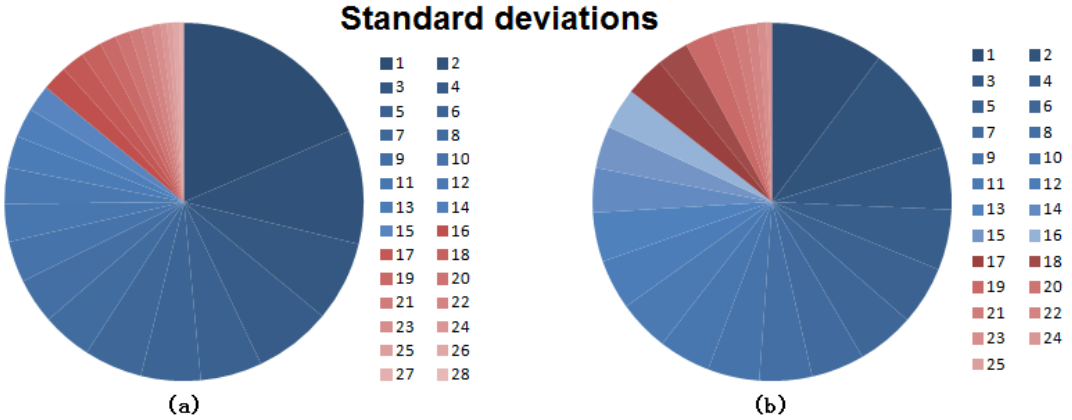

图 4-9 (a)在拍拍贷数据集上各特征的 Standard deviations 值占据的比例(b)在 Prosper 数据集上各特征的 Standard deviations 值占据的比例

4.2.3 降维分类算法

4.2.3.1 参数训练

我们训练了经过 t-test 和 PCA 降维后使用 KNN 算法的 K 值。在这里，我们发现对于拍拍贷数据集，t-test 降维后，K=20 时，KNN 算法的准确率最高；PCA 降维后，K=6 时，KNN 算法的准确率最高（图 4-8）；而对于 Prosper 数据集，t-test 降维后，K=15 时，KNN 算法的准确率最高；PCA 降维后，K=15 时，KNN 算法的准确率最高（图 4-9）。

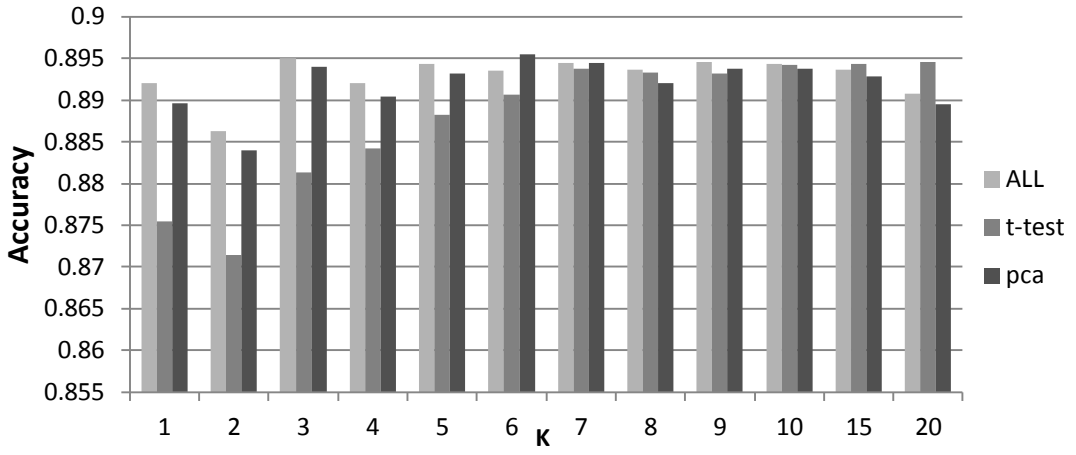

图 4-8 在拍拍贷数据集上经过降维后，KNN 算法随 k 不同准确率变化图

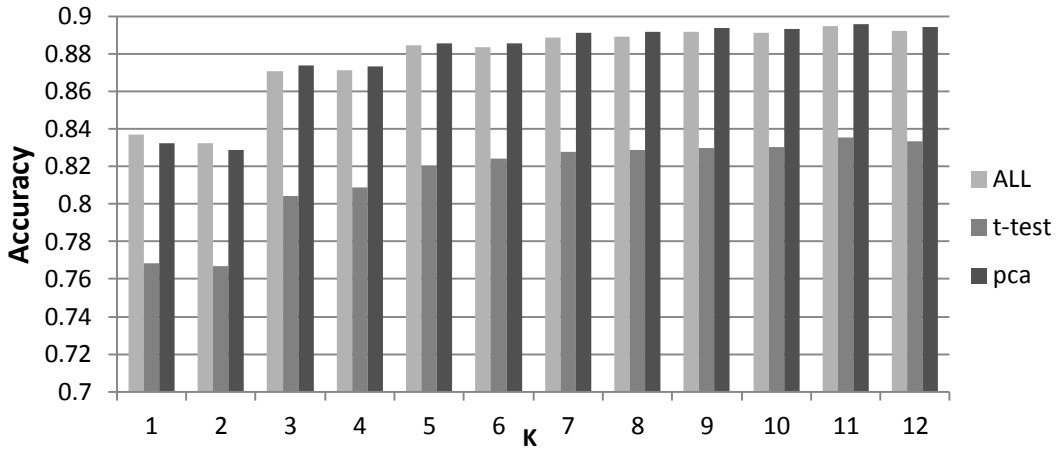

图 4-9 在 Prosper 数据集上经过降维后，KNN 算法随 k 不同准确率变化图

然后同样使用了 SVM、决策树、逻辑回归对降维后的训练样本进行学习。

4.2.3.2 Accuracy、Precision 和 Recall

当我们得到所有模型的最优参数后，我们可以比较每一个模型的预测表现。表 4-5 显示了在拍拍贷数据集各个降维分类模型的 accuracy、precision 和 recall。这其中，PCAM-SVM 算法无论从 accuracy 还是从 precision、recall 来看，都是最优的。

算法	特征数	Accuracy	Precision	Recall	F_measure%
Logistic regression+ALL	28	80.8	80.7	89.6	84.9
Logistic regression+ t-test	13	78.6	80.0	88.6	84.0

Logistic regression+ PCA	15	77.9	80.6	89.6	84.9
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
KNN+ ALL	28	89.5	89.3	88.8	89.0
KNN+ t-test	13	89.5	87.4	88.4	88.0
KNN+ PCA	15	89.6	90.3	87.8	89.0
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
Decision tree+ ALL	28	87.6	84.0	92.0	87.8
Decision tree+ t-test	13	87.0	82.0	**94.5**	87.8
Decision tree+ PCA	15	82.8	80.7	92.2	86.1
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
SVM+ALL	28	90.2	91.7	90.6	91.2
SVM+ t-test	13	90.8	89.7	90.0	90.0
PCAM-SVM	**15**	**91.6**	**91.8**	90.8	**91.3**

表 4-5 在拍拍贷数据集上各分类算法的 Accuracy、Precision 和 Recall

而表 4-6 则说明了在 Prosper 数据集上各个分类模型的 accuracy、precision 和 recall。同样，PCA+SVM 算法无论从 accuracy 还是从 recall 来看，都是最优的，但是从 precision 上，决策树略优于 PCA+SVM 算法。

算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
Logistic regression+ALL	25	88.4	74.7	33.6	46.3
Logistic regression+ t-test	13	81.9	30.6	33.3	31.9
Logistic regression+ PCA	16	81.2	73.2	34.3	46.7
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**

KNN+ ALL	25	89.5	78.5	40.6	53.5
KNN+ t-test	13	83.5	77.8	40.7	53.4
KNN+ PCA	16	89.5	**80.9**	40.2	53.7
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
Decision tree+ ALL	25	89.7	80.8	41.2	54.6
Decision tree+ t-test	13	83.4	80.5	41.5	54.7
Decision tree+ PCA	16	87.6	79.2	41.3	54.3
算法	**特征数**	**Accuracy**	**Precision**	**Recall**	**F_measure%**
SVM+ALL	25	89.7	79.1	41.8	54.7
SVM+ t-test	13	82.8	75.6	43.0	54.8
PCAM-SVM	**16**	**89.8**	**78.7**	**42.0**	**54.8**

表 4-6 在 Prosper 数据集上各分类算法的 Accuracy、Precision 和 Recall

4.3 影响 P2P 借贷的因素

通过分类算法，我们可以将数据分为三类，分别是流标的高信用风险 listing、募得资金但未还款的中信用风险 listing、募得资金并且还款成功的低信用风险 listing。

接下去，我将通过高斯混合聚类方法分析为何会产生这三类信用风险的原因，也就是影响 P2P 借贷的因素。

4.3.1 确定最佳聚类数

表 4 代表每个子数据集的最佳 k 值的取值情况。

Prosper	低信用风险	*k=3*	拍拍贷 低信用风险 *k=7*

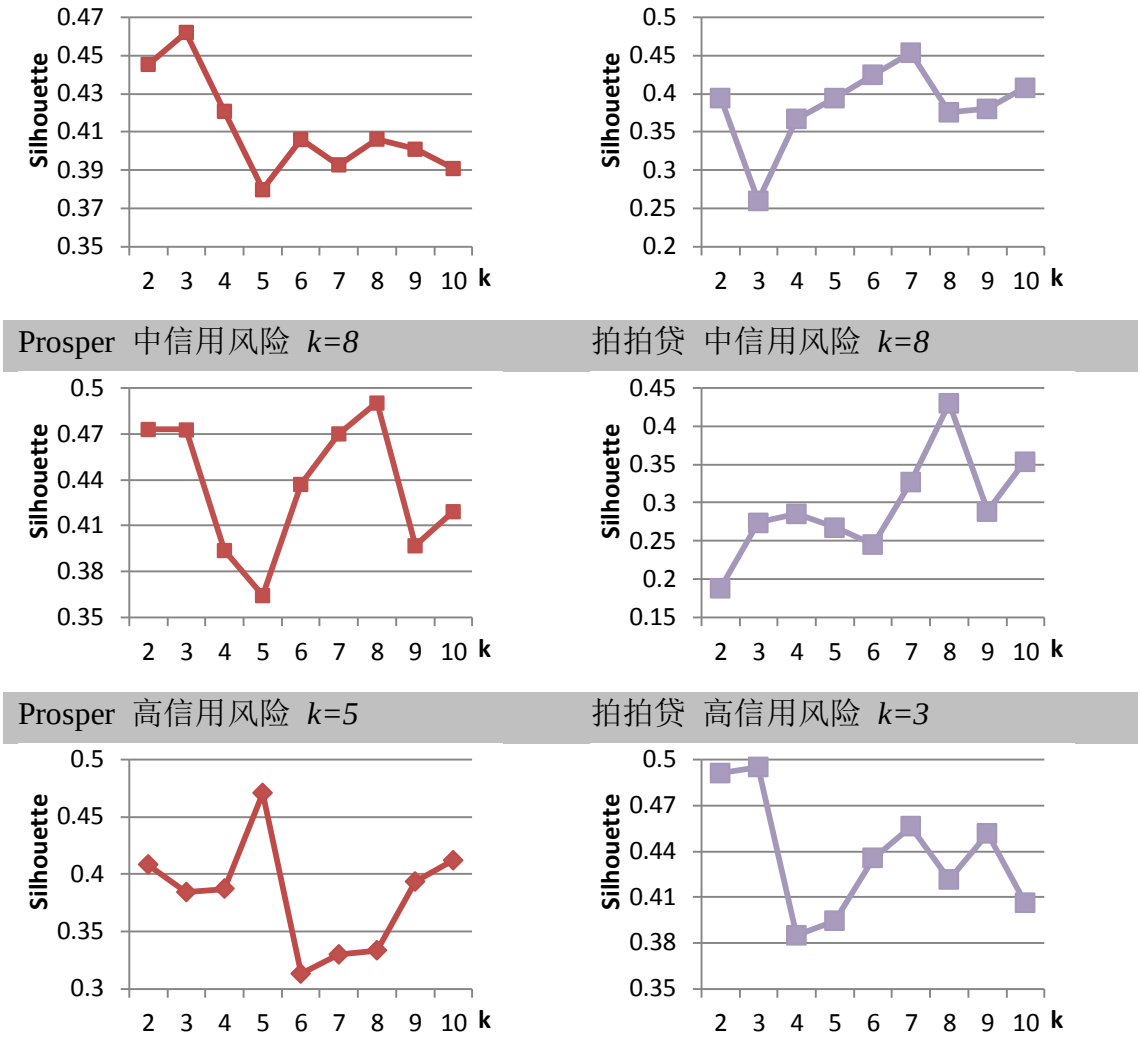

表4-7 所有数据集求得的最佳*k*值

4.3.2 Prosper 数据

在表 4-8 至表 4-13 中，每一列都代表着一个高斯分布。每一个聚类的比例在第一行，其余行为每个高斯分布的特征向量。每个特征向量中的值都被颜色标记来表示对贷款者的吸引程度。其中，绿色表示具有吸引力，红色表示贷款者投资存在风险。颜色的深度是根据图表中的最大值和最小值，如图 4-10。

图 4-10 颜色与对贷款者吸引程度关系图

4.3.2.1 低信用风险

首先，我们对低信用风险的 listing 进行聚类，可以得到表 4-8 的结果：

dataset	total			color set		
P	0.27	0.53	0.20	max	min	avg
Amount Requested	5757.94	8016.48	7876.45	35000	0	9300.38
BidCount	52.10	84.00	94.11	779	1	73.23
Borrower Rate	0.26	0.21	0.14	0.3304	0.0565	0.24
Credit Rating	1.56	3.28	5.08	7	0	3.34
DebtTo IncomeRatio	0.25	0.22	0.19	9.44	0	2.31
IsBorrower Homeowner	0.00	1.00	0.00	1	0	0.53
LoanTermIn Months	40.69	39.56	33.53	60	6	48.53
MonthlyLoan Payment	214.29	286.30	303.48	2259.76	0	296.82
Group Rating	1.48	1.96	2.53	5	0	0.03
GroupParticipa tionRate	0.01	0.01	0.01	0.02		
FriendNum	0.01	0.04	0.03	21	0	0.01
Percent Funded	0.98	0.99	0.99	1	0	0.94

表4-8 Prosper数据集中低信用风险的聚类结果

聚类算法找到 3 个聚类结果在表 4-8 中。从聚类 1 可以清楚的看出较高利率可以吸引更多的资金，而聚类 3 告诉我们如果信用评级很高，即使利率偏低也依然会有较大吸引力。聚类 2 则说明了拥有住房等象征财务实力的证明，也可以吸引贷款人。

除此以外，所有三个聚类都表明 group 评分高的和有较低收入比这两个因素是主要因素来吸引更多的资金，并且最终能成功的几率更高。

4.3.2.2 中信用风险

聚类算法找到 8 个聚类结果在中信用风险的 listing 中，如表 4-9。可以清楚的看出 group 评分高的是主要因素来吸引更多的资金，并且最终能成功的几率更高。有趣的是，通过这张表也可以发现 Credit Grade 高并不是吸引资金的绝对因素，一些 Credit Grade 低的 listing 或通过提高自己的利率或通过增加

自己的社会特征资本来吸引资金。然而，一些 listing 的借款方由于没有加入 group，缺乏还款的监督作用，所以按时还款率不高，变成了中信用风险的用户。

dataset	total								color set		
P	0.12	0.16	0.16	0.16	3.19E-05	0.21	0.20	9.58E-05	max	min	avg
Amount Requested	8357.85	11720.08	6588.89	5096.87	7997.50	14390.38	11246.83	8866.67	35000	0	9300.38
BidCount	60.42	123.94	45.74	38.29	485.87	102.48	94.10	44.94	779	1	73.23
Borrower Rate	0.21	0.12	0.25	0.29	0.13	0.17	0.14	0.15	0.3304	0.0565	0.24
Credit Rating	2.57	5.82	1.80	0.93	2.00	4.68	5.34	4.67	7	0	3.34
DebtTo Income Ratio	0.52	0.21	0.26	0.29	7.00	0.24	0.22	2.90	9.44	0	2.31
Is Borrower Homeowner	1.00	1.00	0.00	1.00	0.00	1.00	0.00	0.67	1	0	0.53
Loan TermIn Months	35.51	35.67	43.54	45.34	36.00	60.00	44.92	44.00	60	6	48.53
Monthly Loan Payment	319.30	394.90	231.41	182.56	385.65	353.43	338.75	259.95	2259.76	0	296.82
Group Rating	1.32	2.26	1.62	1.56	-	1.63	1.88	-	5	0	0.03
Group ParticipationRate	0.001	0.002	0.004	0.002	0	0.002	0.004	0	0.02		
Friend Num	0.01	0.02	0.01	0.01	0.00	0.01	0.02	0.00	21	0	0.01
Percent Funded	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	0	0.94

表4-9 Prosper数据集中的中信用风险的聚类结果

4.3.2.3 高信用风险

如表 4-10 所示，高信用风险的借贷 listing 中总共产生 5 个聚类。从这张表可以清楚的看出低收入比、Credit Grade 低、加入 group 的参与率低都会造成 listing 流标。尤其是聚类 3 和 4 说明了即使利率很高，缺乏社会资本的用

户依然无法吸引贷款方投标。

dataset	total					color set		
P	0.20	0.23	0.24	0.16	0.17	**max**	**min**	**avg**
Amount Requested	11298.51	17542.69	8036.13	3682.81	14895.0	35000	0	9300.38
BidCount	51.26	78.05	43.21	20.78	63.67	779	1	73.23
Borrower Rate	0.23	0.13	0.27	0.32	0.14	0.3304	0.0565	0.24
Prosper Rating	2.59	5.72	1.42	0.04	5.59	7	0	3.34
DebtTo Income Ratio	0.25	0.20	0.27	0.32	0.16	9.44	0	2.31
IsBorrower Homeowner	0.00	1.00	1.00	0.00	0.00	1	0	0.53
LoanTerm InMonths	47.89	47.09	42.23	36.09	46.45	60	6	48.53
MonthlyLoanPayment	380.94	510.88	290.92	160.40	446.63	2259.76	0	296.82
Group Rating	5.00	-	-	-	1.00	5	0	0.03
GroupParticipationRate	0.0003	0	0	0	0.0015	0.02		
Friend Num	0.00	0.00	0.00	0.00	0.00	21	0	0.01
Percent Funded	0.31	0.29	0.33	0.29	0.29	1	0	0.94

表4-10 Prosper数据集中高信用风险的聚类结果

4.3.3 拍拍贷数据

4.3.3.1 低信用风险

聚类算法找到 7 个聚类结果在表 4-11 中。可以清楚的看出身份验证如 ID 验证、电话验证、视频验证、学历验证是吸引更多资金的主要因素之一，并且最终能成功的几率更高。

而由曾经的借款情况产生的拍拍贷借方和贷方的信用分数也是成功的一个重要衡量指标，尤其该用户历史贷款成功次数越多，越容易取得贷款，并且还款几率更高。而朋友数量多的用户的 listing 可以设置更低的利率，承担的还款风险减少了。

另外是否加入拍拍贷的优先计划、保障计划等也都可以一定程度上吸引贷款方。

dataset	total							color set		
P	0.080	0.205	0.076	0.108	0.066	0.130	0.169	max	min	avg
Amount Requested	5469.99	6115.33	7625.78	6762.62	8708.18	6799.90	5397.96	222000	11	13730.99
Bid#	25.41	27.22	27.56	30.50	23.29	27.95	27.69	323	0	8.53
Payment Option	0.98	1.00	0.99	0.98	0.92	0.99	1.00	1	0	0.96
PPDaiIR	14.02	16.31	17.93	12.73	14.15	18.70	16.45	27	0	17.13
SetIntRt	15.27	16.95	18.07	18.26	17.49	18.85	16.79	27	0	17.31
MonRep	1851.35	1788.19	1948.52	2738.55	2908.09	1784.58	1399.02	102250	0	2013.73
Loan Duration	4.39	4.52	5.33	3.44	4.91	5.24	5.12	36	1	8.01
Vouch	0.13	0.05	0.02	0.02	0.02	0.01	0.04	1	0	0.02
Credit Grade	1.53	1.90	2.35	1.62	1.40	2.18	2.06	6	0	1.04
OthBrw	1.00	1.00	1.00	0.00	0.08	1.00	1.00	1	0	0.74
IsDrwBck	0.03	0.09	0.05	0.14	0.01	0.03	0.09	1	0	0.02
IsAdvPln	0.05	0.12	0.11	0.10	0.01	0.14	0.10	1	0	0.05
CdtBrw	1906.96	730.44	734.03	917.80	791.59	392.05	480.25	24308	-8	163.21
CdtLnd	62.10	73.33	88.58	72.55	50.86	76.82	78.95	131	0	40.12
IDVertification	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1	0	0.80
Cellphone Certification	0.76	0.95	0.93	1.00	0.00	0.94	0.96	1	0	0.30

Education Certification	0.35	0.00	1.00	0.35	0.09	0.00	1.00	1	0	0.14
VedioCertification	0.00	1.00	0.97	0.75	0.35	0.91	1.00	1	0	0.29
Num Friend	1.00	3.08	1.00	8.92	3.16	1.00	8.56	694	1	2.25

表4-11 拍拍贷数据集中低信用风险的聚类结果

4.3.3.2 中信用风险

聚类算法找到 8 个聚类结果在表 4-12 中。可以看出朋友数量多，验证途径越多的是主要因素来吸引更多的资金，并且最终能成功的几率更高，而且普遍这样的用户的 listing 并不会设置过高的利率。可见，借款方更看重的是贷款方的社会信息，而不是单纯的利率高低。

dataset	**total**								**color set**		
P	0.254	0.078	0.097	0.145	0.059	0.047	0.043	0.276	**max**	**min**	**avg**
Amount Requested	9150.96	7858.56	10565.68	9488.11	9744.39	7131.19	8306.52	8795.00	222000	11	13730.99
Bid#	31.96	24.52	41.81	35.32	33.35	23.21	21.54	34.20	323	0	8.53
Payment Option	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	0	0.96
PPDaiIR	18.93	9.66	17.35	18.13	16.10	11.39	8.45	18.10	27	0	17.13
SetIntRt	19.06	9.66	17.35	18.17	16.10	11.39	8.45	18.10	27	0	17.31
MonRep	1457.56	1143.41	1544.52	1591.41	1252.64	946.88	1138.44	1265.17	102250	0	2013.73
Loan Duration	7.56	8.18	8.26	7.56	8.31	8.71	8.05	8.14	36	1	8.01
Vouch	0.03	0.77	0.01	0.02	0.00	0.45	0.98	0.01	1	0	0.02
Credit Grade	2.12	2.00	2.42	2.21	2.13	2.40	2.05	2.27	6	0	1.04
OthBrw	0.40	0.00	0.49	0.45	1.00	0.69	0.96	0.78	1	0	0.74
IsDrwBck	0.01	0.73	0.04	0.07	0.13	1.00	0.02	0.00	1	0	0.02
IsAdvPln	0.00	0.31	1.00	1.00	1.00	0.28	0.30	0.00	1	0	0.05
CdtBrw	120.22	2232.63	301.26	292.01	434.65	2309.01	4428.26	278.01	24308	-8	163.21
CdtLnd	65.67	64.64	79.08	70.66	71.72	76.87	71.66	71.64	131	0	40.12
IDVertification	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	0	0.80

CellphoneCertification	0.89	0.91	0.98	0.98	0.97	0.98	0.83	0.95	1	0	0.30
EducationCertification	0.23	0.60	1.00	0.22	0.00	0.67	0.51	0.58	1	0	0.14
VedioCertification	0.85	0.70	0.96	0.94	0.94	0.97	0.68	0.91	1	0	0.29
NumFriend	1.00	46.26	1.00	2.62	1.00	56.32	76.91	1.00	694	1	2.25

表4-12 拍拍贷数据集中的中信用风险的聚类结果

4.3.3.3 高信用风险

如表 4-13 所示，对于高信用风险的 listing 总共产生 3 个聚类。从这张表可以看出所有 listing 的朋友数量都很少，说明如果借款方想要获得成功的借贷，需要更关注自己的社交关系。

并且这些用户也需要增加自己的验证程度，只有当信息认证比较完善的时候，才有可能获得贷款，而非提高贷款的利率。

dataset	**total**			**color set**		
P	0.432	0.322	0.247	**max**	**min**	**avg**
AmountRequested	13885.32	4062.60	31701.24	222000	11	13730.99
Bid#	3.14	7.55	0.04	323	0	8.53
PaymentOption	0.96	0.99	0.87	1	0	0.96
PPDaiIR	18.75	17.84	14.10	27	0	17.13
SetIntRt	18.83	17.88	14.10	27	0	17.31
MonRep	2124.05	1093.14	3309.98	102250	0	2013.73
LoanDuration	8.08	5.48	11.98	36	1	8.01
Vouch	0.00	0.00	0.00	1	0	0.02
CreditGrade	1.15	0.95	0.04	6	0	1.04
OthBrw	0.02	0.93	0.00	1	0	0.74
IsDrwBck	0.00	0.01	0.00	1	0	0.02
IsAdvPln	0.01	0.04	0.00	1	0	0.05

CdtBrw	31.10	68.41	0.11	24308	-8	163.21
CdtLnd	35.55	44.24	13.54	131	0	40.12
IDVertification	1.00	1.00	0.00	1	0	0.80
Cellphone Certification	0.09	0.40	0.00	1	0	0.30
Education Certification	0.04	0.20	0.00	1	0	0.14
Vedio Certification	0.14	0.33	0.00	1	0	0.29
NumFriend	1.00	1.28	1.54	694	1	2.25

表4-13 拍拍贷数据集中高信用风险的聚类结果

4.4 本章小结

本章主要通过拍拍贷和 Prosper 网站上的数据来验证第 3 章中提出的逻辑回归、KNN、决策树、SVM 分类和高斯混合聚类算法，分析了分类算法和 t-test、PCA 降维后的分类算法的结果，得到信用评估模型，然后利用聚类结果简要分析了影响 P2P 借贷的因素。

第5章 分析与讨论

5.1 分类算法的结果分析

通过几个分类算法和降维后分类算法的结果对比，我们可以看出 SVM 算法在 accuracy、precision 和 recall 方面都表现最优。

经过 PCA 降维后，数据集可以有效压缩，并且由于剔除了噪音，在算法准确率上有了一定的提升。而经过 t-test 进行特征选择后，我们发现分类算法并没有很好的提升，这可能是因为去除的特征依然对分类结果有一定的作用，这种方式损失了数据内在结构的完整性；但是降维后的分类算法在效率上得到提高，而且特征选取的降维方法可以使获取特征降低了难度。

5.2 影响 P2P 借贷的因素分析

5.2.1 个人信息对借贷的影响

个人信息包含个人身份信息、学历信息、财务信息，这些都是传统借贷中十分重要的参考依据。所以在网络 P2P 借贷中，贷款方也会着重重视这方面的信息。

在 Prosper 平台上，一个拥有不错的收入比、拥有住房、银行信用等级良好的用户往往可以成功获得贷款；并且他们并不需要设置过高的利率来吸引贷款方，这样可以降低借贷的成本。

而拍拍贷平台上，身份、学历、电话、视频的认证几乎是判断用户可信程度的极重要的指标，如果一个新用户想要在拍拍贷平台上获得贷款，那么尽可能的完成拍拍贷平台的各项认证是十分有必要的。

5.2.2 借贷信息对借贷的影响

P2P 借贷本身的一些信息也会对借贷的结果产生影响，最直观的想法就是利率高的容易获得贷款。然而从实验结果分析，我们发现，高利率并不能完全吸引贷款人冒着信用风险来对借款人进行放贷。

相反，如拍拍贷平台这样，尽量多争取加入一些拍拍贷的优先计划、保障计划，在一定程度上比设置高利率更有效。

5.2.3 社会特征对借贷的影响

5.2.3.1 减少信息不对称

信息不对称理论是指在市场经济活动中，各类人员对有关信息的了解是有差异的；掌握信息比较充分的人员，往往处于比较有利的地位，而信息贫乏的人员，则处于比较不利的地位。该理论认为：市场中卖方比买方更了解有关商品的各种信息；掌握更多信息的一方可以通过向信息贫乏的一方传递可靠信息而在市场中获益；买卖双方中拥有信息较少的一方会努力从另一方获取信息。

社交网络的关系维度可以降低交易过程中的信息不对称。使用从 Prosper.com 收集的数据，Lin 等人(2009a[34];2009b[35];2009[36])研究了社交网络在提高贷款成功率和降低贷款利率中起的作用，他们发现社交网络可以有效地降低交易过程中的信息不对称。在线社交网络在减少信息不对称，提高信用评级方面起了重要作用(Everett, 2008[10])。研究进一步表明，成为一个受信任组织的成员能提高贷款的成功率，同时也能帮助信用评分较低的人们以负担得起的利率获得贷款资助(Lopez et al., 2009[37])。在信息技术的帮助下，社交网络可以向贷款人发送有价值的信号，但信号的有效性取决于他们的可靠性和可验证性(Krumme &Herrero 2009[32])。

5.2.3.2 降低借贷风险

社会资本一般指与社交网络相关联的资源。个体的社会资本可以从他的社交网络，包括朋友和同事(Burt, 1992[11])，或社交网络中的组员(Portes, 2000[43])来评估。在 P2P 借贷中，社会资本作为软信用信息的主要来源可能会影响贷款的成功率和利率。实验发现，社会资本越多的借款人，拥有更大的机会可以获得他们的贷款资助，并且可以得到更低的利率。当借款人的信用评分较低时，贷款人需要更多的信息来进一步评估借款人的信誉以降低贷款风险。

在这种情况下，贷款人依赖于借款人的社交网络来获取信息产生自己的决策。

而建立一个自组织的团体可以帮助在线 P2P 借贷市场有效运作，一些论文的研究揭示了这种混合结果：允许该组织领导人在完成贷款后获得奖励(如成功贷款费用)是有害的(Hildebrand et al., 2010[26])。而我们的实验也证明了组织贷款的投资回报率明显低于非组织贷款。当借款人属于某个组织中，贷款违约率也会相对较低。

5.3 本章小结

本章通过分析和总结第 4 章的实验结果，分析了分类算法的结果，确定了适合的信用评估模型；然后分析了个人信息、借贷信息、社会特征对 P2P 借贷的影响，为高信用风险借贷提出了增加借贷成功率、提示个人信用的建议。

第6章 本文总结

6.1 论文主要工作

本文调查了在线 P2P 借贷平台中信用风险缺乏的现状，对拍拍贷和 Prosper 两个平台的数据进行分析。通过对现有的信用评估模型结合互联网的特点提出一种基于 PCA 降维支持向量机的信任算法。该算法从 P2P 借贷的数据中通过数据预处理等手段抽取出训练样本，使用 PCA 降维去噪和支持多分类的 SVM 来学习样本，并对测试样本进行预测。

最后，在分类的基础上通过高斯混合聚类分析了影响 P2P 借贷的因素。并且深入分析了个人信息、借贷信息和社会特征对 P2P 借贷的具体影响，为借贷用户提出如下建议：

- 在 P2P 借贷过程中，受信任的群体可以更容易获得贷款，并且不需要设立过高的贷款利率来吸引贷方。
- 由于在线 P2P 借贷的互联网特点，尽可能多的验证个人信息可以增加信任程度，降低信用风险。
- 信任缺乏的人群一般通过增加财务吸引力（如高利率）获得贷款，但是这样无疑加重了借方的债务负担。
- 一些信任缺乏的人群通过增加更多的社会特征（如加入高评分组织，增加朋友数量）来获得贷款，这样不仅可以更容易获得贷款，还可以利用朋友和组织的监督作用，提高还款率。

6.2 将来的工作

在未来，P2P 借贷的发展将围绕一个问题：P2P 借贷能否为人们提供一个有益的金融服务，从而对经济困难的家庭产生积极的影响。同样，我们相信传统的金融机构对在线 P2P 借贷平台的发展也持有不小的兴趣。这个开创性的对真实数据的分析，揭示了社会交往对增加信贷的成功的影响。然而，有几个问题仍然没有答案，所以我们提出了进一步研究：

- 文本分析：事实上，无论 prosper 还是拍拍贷提供的数据结构都远远不止我们精简后的这么简单，有许多表单和项目在本文中都忽略考虑了，而这些部分也许会让我们发现更有趣的结论。
- Loan 情况的分析：在本实验中，我们着重研究了借款人如何成功的申请到一笔成功的借贷，但是对还款结果，我们并没有分析。在 P2P 借贷的过程中，如何确保贷款最终能按时还清这也是一个需要研究的重要问题。
- 隐藏社交信息：在社交信息的分析中，我们暂时只使用了最基本的一些情况，如 group 的评分、朋友的数量等。但是，我们还可以利用更多手段来综合分析在借贷过程中产生的社交联系。
- scale-free networks：用户的 Bids 往往会对其他用户产生吸引力，从而导致最终产生大量的 Bids 而使得 Listing 成功。这一现象，符合典型的无标度网络原则。所以，一些 small world networks 的分析也可以应用到这一领域。

参考文献

[1] ALBORZI MAHMOUD, M. P. M., & KHAN, B. M. (2010). Using genetic algorithm in optimizing decision trees for credit scoring of banks customers.Journal of Information Technology Management.

[2] Ashta, A., & Assadi, D. (2009). An analysis of European online micro-lending websites. Cahiers du CEREN, 29, 147-160.

[3] Bachmann, Alexander, et al. (2011). Online Peer-to-Peer Lending--A Literature Review.Journal of Internet Banking & Commerce, 16(2).

[4] Berger, S., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. BuR Business Research Journal, 2(1).

[5] Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. Journal of Banking & Finance, 30(3), 851-873.

[6] Brown, C.M. (2008). Is peer-to-peer lending right for you? Black Enterprise (39:2), pp. 146-146.

[7] Chen, D.Y. (2012). Is online peer-to-peer lending market effective? A study on herding behavior in China,Working Paper (School of Management, Fuzhou University).

[8] Collier, B. C., & Hampshire, R. (2010, February). Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (pp. 197-206). ACM.

[9] Danenas, P., & Garsva, G. (2012). Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach. Procedia Computer Science, 9, 1324-1333.

[10] Everett, C. (2008). Group membership, relationship banking and loan default risk: the case of online social lending. Relationship Banking and Loan Default Risk: The Case of Online Social Lending (March 15, 2010).

[11] Freedman, S., & Jin, G. Z. (2008). Do social networks solve information

problems for peer-to-peer lending? evidence from prosper. com (No. 08-43). com .NET Institute Working Paper.

[12] Galloway, I. (2009). Peer-to-peer lending and community development finance (No. 2009-06). Federal Reserve Bank of San Francisco.

[13] Garman, S. R., Hampshire, R. C., & Krishnan, R. (2008). Person-to-person lending: The pursuit of (more) competitive credit markets.

[14] Greiner, M. E., & Wang, H. (2009). The role of social capital in people-to-people lending marketplaces.

[15] Herrero-Lopez, S. (2009, June). Social interactions in P2P lending. InProceedings of the 3rd Workshop on Social Network Mining and Analysis (p. 3). ACM.

[16] Herzenstein, M., Andrews, R. L., & Dholakia, U. M. (2008). The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. papers. ssrn. com.

[17] Hoffmann, F., Baesens, B., Mues, C., Van Gestel, T., & Vanthienen, J. (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. European Journal of Operational Research, 177(1), 540-555.

[18] Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847-856.

[19] Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?.

[20] Johnson, S., Ashta, A., & Assadi, D. (2010). Online or Offline?: The Rise of "Peer-to-Peer" Lending in Microfinance. Journal of Electronic Commerce in Organizations (JECO), 8(3), 26-37.

[21] Kiva. (2011). Latest statistics. Available at http://www.kiva.org/about/facts (Accessed on 8/15/2011).

[22] Klafft, M. (2008, March). Peer to peer lending: auctioning microcredits over the internet. In Proceedings of the International Conference on Information Systems, Technology and Management, A. Agarwal, R. Khurana, eds., IMT, Dubai.

[23] Krumme, K. A., & Herrero, S. (2009, August). Lending behavior and community structure in an online peer-to-peer economic network. InComputational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 613-618). IEEE.

[24] Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. Expert Systems with Applications, 32(2), 427-431.

[25] Lin, M., Prabhala, N., & Viswanathan, S. (2009a). Social networks as signaling mechanisms: Evidence from online peer-to-peer lending. WISE 2009.

[26] Lin, M., Prabhala, N. R., & Viswanathan, S. (2009b). Judging borrowers by the company they keep: Social networks and adverse selection in online peer-to-peer lending. SSRN eLibrary.

[27] Lin, M. (2009). Peer-to-peer lending: An empirical study.

[28] Liu, S. A., Wang, Q., & Lv, S. (2008, July). Application of genetic programming in credit scoring. In Control and Decision Conference, 2008. CCDC 2008. Chinese (pp. 1106-1110). IEEE.

[29] Lopez, S. H., Pao, A. S. Y., & Bhattacharrya, R. (2009). The effects of social interactions on P2P lending. MAS Final Project, 1-24.

[30] Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. Communications of the ACM, 30(6), 484-497.

[31] Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N., & Zopounidis, C. (2008). Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. Journal of Global Optimization, 42(2), 279-293.

[32] Patsuris, P. (1998). Cut out the middleman, Forbes

[33] Puro, L., Teich, J. E., Wallenius, H., & Wallenius, J. (2010). Borrower decision aid for people-to-people lending. Decision Support Systems, 49(1), 52-60.

[34] Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. Operations research, 42(4), 589-613.

[35] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications. Siam.

[36] Vukovic, S., Delibasic, B., Uzelac, A., & Suknovic, M. (2012). A case-based reasoning model that uses preference theory functions for credit scoring.

Expert Systems with Applications, 39(9), 8389-8395.

[37] Wang, C. M., & Huang, Y. F. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. Expert Systems with Applications, 36(3), 5900-5908.

[38] Wang, H., Greiner, M., & Aronson, J. E. (2009). People-to-people lending: the emerging e-commerce transformation of a financial market. In Value Creation in E-Business Management (pp. 182-195). Springer Berlin Heidelberg.

[39] Zhang, Zhiwang, Guangxia Gao, and Yong Shi. "Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors." European Journal of Operational Research (2014).

[40] Zhou, Hanhai, et al. "Application of the Hybrid SVM-KNN Model for Credit Scoring." Computational Intelligence and Security (CIS), 2013 9th International Conference on. IEEE, 2013.

致谢

首先，我衷心感谢我的导师郑小林副教授在学习和生活中给予我的谆谆教诲和悉心关怀。本论文从选题、构思、修改到成文，每个环节都凝结了实验室老师大量的心血，充斥了实验室学长学姐的耐心指导。郑小林教授专业知识渊博，治学态度严谨，在学术上精益求精、积极进取，在工作中求真务实，在生活中平易近人，给我留下了深刻的印象。在这短短一年的相处时间里，我深深感受到了郑老师出色的人格魅力，这在以后的求学道路上无疑给我树立了旗帜鲜明的方向标，是我一生取之不尽的宝贵财富。

此外，我要感谢浙江大学电子服务研究中心的所有师兄师姐们，他们是陈超超博士、魏守贤博士、林臻博士、扈中凯博士、叶夏菁、许欢、邓志豪、倪泽明、许凌之、洪福兴、马国芳、肖力涛、赵家骏除此以外，还有已经毕业的曾晶学姐，以及同一届的王梦晗同学。在学习和工作中，我总能从他们身上学到很多东西，我的每一点进步都离不开他们对我的支持和帮助

感谢计算机学院的各位老师，谢谢你们在我本科生阶段在学习上带领我进入丰富多彩的计算机科学与技术的世界。

感谢对我本科论文进行评审的各位专家教授，感谢你们对论文的指导和提出的宝贵意见。

我要特别感谢我的父母，他们在我的成长中起到了不可磨灭的影响，是我人生中最重要的依靠和指引。

最后，感谢所有给予我指导、帮助、关心和支持的老师、亲人和朋友们。

朱梦莹

2014 年 5 月

本科生毕业论文（设计）任务书

一、题目： **P2P 借贷平台中的信用评估模型**

二、指导教师对毕业论文（设计）的进度安排及任务要求：

该毕业论文要求对互联网金融和 P2P 借贷相关国内外研究现状进行深入分析的基础上，提出信用评估预测、信用风险分析等方法，并重点围绕信用评估预测等展开研究。实验数据要求采用国际知名的 Prosper 数据集和拍拍贷数据集，实验结果要清晰可信。进度安排如下：

3.1-3.15 研究方案确定

3.16-4.15 模型的建立与算法实现

4.15-5.15 实验与分析

5.16-5.30 论文纂写

起讫日期 2014 年 3 月 1 日 至 2014 年 5 月 30 日

指导教师（签名） 郑小林 **职称** 副教授

三、系或研究所审核意见：

负责人（签名）

年 月 日

毕 业 论 文（设计） 考 核

一、指导教师对毕业论文（设计）的评语：

互联网金融中 P2P 借贷的信用评估是当前的研究热点，作者以该方向为研究目标，选题具有较高的应用价值。《P2P 借贷中的信用评估》在对 P2P 借贷的信用评估等国内外研究现状进行深入分析的基础上，提出了基于 PCA 降维支持向量机的信用评估算法。通过在 Prosper 数据集和拍拍贷数据集上开展的实验分析表明，该方法具有一定的先进性和较高的实用性。论文工作表明作者具有扎实的基础理论知识，以及一定的科研工作的能力。论文条理清楚，层次分明，达到了本科毕业论文的要求。

指导教师(签名）郑小林

2014 年 **5** 月 **31** 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述占（10%）	开题报告占（20%）	外文翻译占（10%）	毕业论文（设计）质量及答辩占（60%）	总 评成绩
分值					

答辩小组负责人（签名）

年 月 日