**学校编码：10384 分类号 密级**

**学号：23020131153163 UDC**



硕 士 学 位 论 文

**基于异构网络链接预测算法研究**

**The Research of Link Prediction Algorithm**

**on Heterogeneous Information Network**

李友

指导教师姓名：曾湘祥副教授

专 业 名 称：计算机科学与技术

论文提交日期：2016年5月

论文答辩时间：2016年5月

学位授予日期：2016年 月

答辩委员会主席：

评 阅 人：

2016 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

现实中的许多应用，如社交网络、P2P信贷、生物信息、社交投资等许多数据库系统都是构成异构信息网络。因此，针对异构信息网络的数据挖掘任务是当前热门的且充满挑战的研究新方向。本文主要针对异构信息网络中的链接预测问题进行数据挖掘研究。不同于当前已有的许多数据挖掘方法，在同构网络的基础上处理链接预测问题。我们提出的链接预测模型是将实际应用问题模型为异构信息网络，因为异构网络覆盖更全面的实体和关系数据并且蕴含更丰富的语义信息。具体地，我们做了如下三项工作和创新：

(1)针对致病基因网络的链接预测问题，提出了使用异构相似约束的隐因子关系预测模型。隐因子模型假设对象之间在相同特征空间里，距离越近则越相似，将疾病和基因之间的关系预测转化为二分类概率模型：疾病和基因如果在隐含特征空间的距离越近,则它们存在关系的概率越高；反之，若距离越远，则存在关系的概率越低。为了充分利用异构信息网络所覆盖的丰富语义而更准确的挖掘基因和疾病之间可能存在的相关性，我们在基础模型的基础上加入了异构相似约束。最后，为了验证提出的算法的有效性，我们在真实的生物数据上进行实验，对比已有的相关算法，分析算法性能。

(2)我们设计了一个数据驱动的社交投资网络链接预测框架，该框架将投资行为模型为基于meta-path的异构网络链接预测问题，并提供了有效的度量函数来量化meta-path相似度。为了验证提出模型的有效性，我们在CrunchBase上提供的真实数据进行实验。实验结果揭示，我们的模型在一定程度上也可以是一个有用的工具，帮助企业(1)更好的掌握投资者是怎样和何时会投资，(2)更好的提供参考信息，当他们需要寻求外部投资时。

(3)针对P2P信贷网络的链接预测问题，我们提出了一个基于二部图的随机游走投资预测模型。使用来自P2P信贷平台上的真实数据实验，结果说明了我们的预测模型在一定程度上的有效性。多次实验结果也表明，使用我们的预测模型和Logistic模型集成的混合结果，可以取得更有效且稳定的表现。实验结果也在一定程度上说明，我们提出的预测模型和Logistic模型具有良好的互补特性。因此，相对于使用单一模型的预测结果，使用混合模型可以取得更佳的预测表现。

关键词：链接预测；异构信息网络

Abstract

Most real-world applications, including social network, P2P Lending, bioinformatics, social investment network et al, can be structured into heterogeneous networks. Therefore, the data mining works of heterogeneous networks are frontier and challenging research area. In this thesis, we investigate the principles and methodologies of link prediction in heterogeneous information network. Differing from those old mining methodologies, they tackle link prediction problems based on homogeneous. Our link prediction models are based on heterogeneous network, which leverages more extensive entity and relationship data and the rich semantics information. In detail, we finished the following works and innovations:

(1) In view of disease-gene network link prediction problem, we proposed a latent factorization relationship prediction model with heterogeneous similarity regularization. The idea of latent factorization model is that in same feature space between objects with closer distance are more similar. We model disease-gene relationship prediction into binomial classification model: The pair of disease and gene are the closer the distance the higher the probability of related; otherwise, the farther the distance the lower the probability of related. In order to utilize effectively of heterogeneous information conveyed with rich semantic and more accurate mining possible correlation between genes and disease, we joined the heterogeneous regularization in the basic model. Finally, to verify the effectiveness of the algorithm, we test with real-world biological data. Analyzed the performance of algorithm and compared with other algorithms.

(2) We designed a data driven investment behavior prediction framework, which models investment behavior prediction as a meta-path-based Heterogeneous Information Network relationship prediction problem. To validate the proposed model, we perform experiments on real-world data from CrunchBase. Experimental results reveal that our model is also a useful indicator for helping (1) companies better understand how and when investors invest, and (2) be better prepared when they are attempting to seek external investment.

(3) For the link prediction problem of P2P lending network, we propose a random walk investment prediction model based on bipartite graph model. Using the confidence level of investors, we estimate the paid probability of new or current loans. Experimental results on real-world P2P lending data demonstrate that our decision model is effective in a certain extent. In addition, experimental results of the hybrid classification model show that the Logistic classification model and our iteration computation model complement each other. Then, we conclude that the hybrid model is more efficient and stable than the individual model alone.

Key words: Link Prediction; Heterogeneous Information Network

**目 录**

摘 要 I

Abstract III

第一章 绪论 1

1.1 课题概述和意义 1

1.1.1 异构信息网络 2

1.1.2 意义 5

1.2 本文的组织结构 6

第二章 致病基因网络中的链接预测 9

2.1 概述 9

2.1.1 问题简介 9

2.1.1 数据介绍 10

2.2 相关工作 11

2.3 隐因子模型 12

2.4 异构相似约束的隐因子关系预测模型 13

2.4.1 半监督学习 13

2.4.2 基础模型:隐因子关系预测模型 14

2.4.3 模型I:平均异构相似约束的隐因子关系预测模型 15

2.4.4 模型II:个性化异构相似约束的隐因子关系预测模型 16

2.4.5 基因相似性计算 17

2.5 实验分析 19

2.5.1 算法准确性比较 19

2.5.2 参数和的影响 21

2.5.3 隐因子维度D的影响 23

2.6 本章小结 24

第三章 社交投资网络中的链接预测 25

3.1 概述 25

3.1.1 问题简介 25

3.1.2 数据介绍 26

3.2 相关工作 27

3.3 基于Meta-Path的链接预测 28

3.3.1 社交投资异构信息网络 28

3.4 异构社交投资网络的Meta-path 29

3.5 Meta-path相似度量函数 30

3.6 投资网络链接预测 31

3.6.1 链接预测模型 31

3.6.2 链接建立时间预测模型 32

3.7 实验分析 34

3.7.1 综合实验效果比较 34

3.7.2 Meta-path意义探索 34

3.7.3 预测链接建立时间 35

3.8 本章小结 37

第四章 P2P信贷网络中的链接预测 39

4.1 概述 39

4.1.1 问题简介 39

4.1.2 数据介绍 39

4.2 相关工作 40

4.3 投资者和贷款的综合评估 42

4.3.1 二部投资网络 42

4.3.2 初始评估模型 43

4.3.3 第二轮评估 44

4.3.4 多轮迭代至收敛 45

4.3.5 收敛特性 46

4.4 实验分析 46

4.4.1 结果分析 47

4.4.2 多种预测模型比较 48

4.4.3 和Logistic模型的混合 49

4.5 本章小结 51

第五章 总结和展望 53

5.1 总结 53

5.2 展望 54

参考文献 57

攻读硕士学位期间发表论文及科研情况 63

致谢 64

**Contents**

Abstract(CN) I

Abstract(EN) III

Chapter 1 Introduction 1

1.1 Background and Significance 1

1.1.1 Heterogeneous Information Network 2

1.1.2 Significance 5

1.2 Thesis Organization 6

Chapter 2 Link Prediction on Gene-phenotype Hetergeneous Network 9

2.1 Overview 9

2.1.1 Problem Description 9

2.1.1 Experimental Data 10

2.2 Related Works 11

2.3 Latent Factorization Model(LFM) 12

2.4 Latent Factorization Relationship Prediction Model with Heterogeneous Similarity Regularization 13

2.4.1 Semi-Supervised Learning 13

2.4.2 Basic Model:Latent Factorization Relationship Prediction Model 14

2.4.3 Model I:Latent Factorization Relationship Prediction Model with the Average Heterogeneous Regularization 15

2.4.4 Model II: Latent Factorization Relationship Prediction Model with the Personalization Heterogeneous Regularization 16

2.4.5 Genetic Similarity Calculation 17

2.5 [Experimental](javascript:void(0);) [Analysis](javascript:void(0);) 19

2.5.1 Accuracy Comparison 19

2.5.2 Impact of Parameters and 21

2.5.3 Impact of Latent Factorization Dimension **D** 23

2.6 Chapter Conclusion 24

Chapter 3 Link Prediction on Social Investment Network 25

3.1 Overview 25

3.1.1 Problem Description 25

3.1.2 Experimental Data 26

3.2 Related Works 27

3.3 Link Prediction Algorithm Base on Meta-Path 28

3.3.1 Social Investment Heterogeneous Information Network 28

3.4 Meta-path on Heterogeneous Social Investment Network 29

3.5 Measure Function on Meta-path 30

3.6 Link Prediction on Investment Network 31

3.6.1 Link Prediction Model 31

3.6.2 Link Building Time Prediction Model 32

3.7 [Experimental](javascript:void(0);) [Analysis](javascript:void(0);) 34

3.7.1 Aggregate Performance Comparsion 34

3.7.2 Significant of Meta-path Study 34

3.7.3 Link Building Time 35

3.8 Chapter Conclusion 37

Chapter 4 Link Prediction on P2P Lending Network 39

4.1 Overview 39

4.1.1 Problem Description 39

4.1.2 Experimental Data 39

4.2 Related Works 40

4.3 Comprehensive Evaluation of Investor and Loan 42

4.3.1 Bipartite Investment Network 42

4.3.2 Initial Evaluation Model 43

4.3.4 Iteration Computing and Convergence 44

4.3.5 Convergence Properties 45

4.4 [Experimental](javascript:void(0);) [Analysis](javascript:void(0);) 46

4.4.1 Result Analysis 47

4.4.2 Result Comparisons 48

4.4.3 Hybird Model with Logistic 49

4.6 Chapter Conclusion 51

Chapter 5 Conclusion and Future Work 53

5.1 Conclusion 53

5.2 Future Work 54

References 57

Publications 63

Acknowledgement 64

第一章 绪论

我们都生活在一个相互关联在一起的网络世界。我们熟知的信息或数据对象之间，所有个体之间，社区和社区之间，存在要么是直接相连的关系要么是相互之间的交互关系。这些事物构成了一个规模庞大、结构复杂的关联网络。不失一般性，我们可以称这些关联网络为信息网络。我们常见的信息网络有社交网络、互联网、生物信息网[1]、公共交通网、P2P信贷网[2]和电力网络等。显然，信息网络普遍存在于我们日常，并且成为了我们当代信息基础结构的重要部分。对于信息网络的分析，特别的如社交网络和互联网，在计算机科学、社会科学、经济学、生物学等领域的学者中引起了广泛关注和研究。

在本文，我们将现实中的应用问题模型为异构信息网络，并探索其规则和适应的算法来系统地挖掘该异构信息网络。特别地，针对多个不同应用环境下的异构信息网络模型，提出了针对各个具体问题而言相对有效的链接预测算法。不同于我们熟知的网络模型，如同构的图或网络，我们使用的异构信息网络模型利用了节点和链接的丰富语义，同时相对于同构类型的网络取得了更有效的结果。

在这一章节，我们首先在1.1节介绍了本研究的背景和意义，然后再1.2节介绍了本文的组织结构。

1.1 课题概述和意义

当前关于网络科学、社交网络或信息网络的研究，大多是将这些网络视作同构型来研究的。在同构网络中，所有节点都被看作同种对象类型（如社交网络中的用户），并且网络中的链接只存在一种关系类型（如社交网络中的好友关系）。目前，基于同构网络模型的研究成果得到了极具轰动的应用，如著名的PageRank[3]算法和社区检测方法[4]。然而，现实世界中大多数网络是异构的，如生物信息中的致病基因网络，节点可以是基因、疾病、人类或其它动物等。同构信息网络模型存在两个很容易忽视的缺陷：a)同构信息网络将所有节点视作同一种类型，很可能丢失哪些很重要的语义信息；b)如标签图网络中对待每一个节点为唯一类型也很可能丢失掉有价值的模式级别的信息。而如果对致病基因的交互网络，区别所有单个基因是同属于一类称作基因的节点，且每个基因节点和每个疾病或人类节点是属于不同的类型，在挖掘基因和疾病之间的关系时，是非常有意义的。因此，在一些典型的现实应用中，异构信息网络可能能够更好地捕捉该网络中的本质。

异构信息网络在现实生活中是普遍存在的。例如微博社交网络的组成包括用户、照片、博文、职业、兴趣等；而用户-用户之间关注和被关注关系、用户-照片之间的发布关系、用户-职业之间的从属关系，用户-兴趣之间的喜好关系等。类似的例子随处可见，无论是从社交媒体到电子商务等互联网中的各个方面，再到科学、工程或医疗系统。因此不难相信异构信息网络在各个许多领域中，可以有效地模型现实中常见的交互网络。

在本文，通过利用异构信息网络中不同类型的节点和节点之间的关系链接的重要语义，我们研究了已有的相关研究成果，提出了新的模型来预测网络中节点之间可能存在的链接关系，用于解决实际应用问题。异构信息网络通常能够覆盖同构信息网络无法模型的语义结构。在异构网络中，存在多种对象类型，不同类型对象之间的关联通常代表了不同的意义，这种代表不同类型之间的链接在传统同构网络中是很难被表达的，然而在异构网络中，可以很简便的模型这些异构链接。总之，相较于同构信息网络，利用异构信息网络模型，不仅可以处理多种不同的实体对象，也能表达不同对象类型之间的不同关系；这意味着我们既可以综合尽可能多的相关信息，也利于发现更多有价值的模式，这都有利于处理现实中常见的异构问题，尤其在当前数据大爆炸时代，在互联网中产生了大量异构信息，而充分利用和挖掘这些信息正是当前面临的课题。在后面的章节，我们将介绍使用基于异构信息网络模型的建模方法来解决现实应用中常见的关系预测（或链接预测）问题。

1.1.1 异构信息网络

通常信息网络是现实问题中的一个抽象，将一个实体对象抽象为网络中的一个节点，实体之间的交互关系抽象为网络中的一条边。理论和实践都已验证，这样一类抽象不仅可以有力的描述现实问题，而且给出了一系列非常有效的方法来探索和挖掘实际问题，特别的如社交网络中的链接预测问题。下面，我们首先给出信息网络的正式定义:

**定义1** (信息网络) 使用代表一个有向图，映射函数表示节点类型到实体类型的映射，映射函数表示链接类型到关系类型的映射，那么对于任意有，任意有，并且如果两条链接属于同种类型，则这两条链接共享同一个起始节点和相同的结束节点。

不同于定义1中所定义的传统信息网络，异构信息网络显示的区分不同的实体类型和关系类型。特别地，如果网络中A类节点到B类节点存在一条路径，看作，表示A到B的关系，反之表示B到A的关系。在异构网络中，通常和是不相同的，除非A和B是同一种类型，这样便是对称的。如果信息网络中或，也即是实体类型或关系类型的个数不止1个，那么该信息网络被称作异构信息网络，否则即是同构网络。

无论是从社交网络或其他web应用，到生物信息再到工程或商业应用，这其中各个领域中产生存在相互关联的大规模数据集都可以构成异构信息网络。下面我们简单描述本文涉及的三个不同应用领域的异构信息网络模型。

Genes

Non-human

Phenotypes

Phenotypes

**图1.1 致病基因异构信息网络**

(1)致病基因网络：对于致病基因网络，像来自OMIM(0nline Mendelian Inheritance in Man)数据库的关于人类基因和各类表型的信息网络，就是一个典型的异构信息网络。我们可用的基础数据中包括三种不同意义的实体类型：人类基因(G)、人类表型()、和八种和人类基因有关的非人类表型()。对于人类基因类型中的节点，可能包含的链接类型有：人类基因之间的内在关联网络，人类基因和人类表型之间的关联网络，人类基因和非人类表型之间的关联网络。这些实体类型和关系类型之间构成异构网络模式，如图1.1所示。

Location

Category

Investment

Market

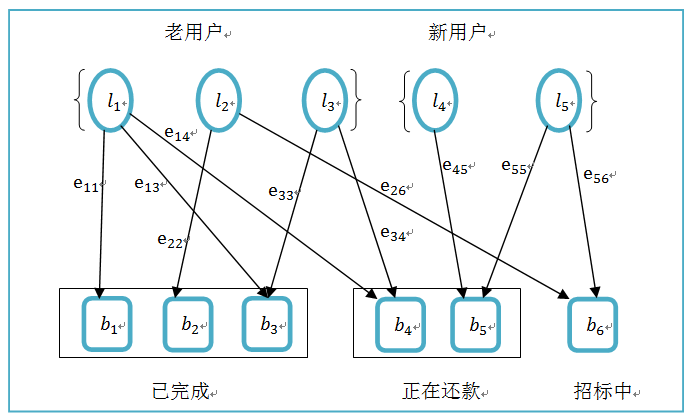
Acquisition

Company

City

**图1.2 CrunchBase数据集异构信息网络**

(2)社交投资网络：已有相当的研究探讨了投资行为，从已有的研究来看，与人的投资行为相关的因素有多种。实际上，由投资者、投资目标和与其相关类型之间构成的投资网络是一个异构信息网络。如本文在第三章，使用CrunchBase(http://www.crunchbase.com)数据集讨论的关于社交投资网络中异构网络链接预测问题。在我们实验的数据集中包括四种不同类型的实体类型：投资者(Investor)、公司或机构(Company)、城市(City)、产品市场(Category)。实体类型之间存在的关系类型有：投资类型(Investment)、市场类型(Market)、地理位置类型(Location)和收购关系(Acquisition)。这些实体和关系类型之间构造的异构网络模式，如图1.2所示(详细说明见第三章)。



**图1.3 P2P信贷网络**

(3)P2P信贷网络：P2P信贷是以互联网为基础而发展的新兴金融市场，目前在国内外取得飞速发展。在本文第四章中，我们论述了使用美国最早的P2P信贷平台Prosper(https://www.prosper.com/)提供的交易数据来分析P2P信贷网络中的异构网络链接预测问题。具体地，我们使用数据集中包含的两种实体类型(投资者和贷款项)和他们之间的竞标关系构成的二部图网络，如图1.3所示(详细说明见第四章)。

在某一应用中的各类信息都可以使用信息网络来抽象。属性值可以被赋予到信息网络中的节点或链接中。例如在微博社交信息网络中，一些位置属性、类别或数值属性通常可以赋予给用户和发布的微博。此外，许多信息网络中的信息都是暂时性的，哪些赋予给节点和链接的信息都是为了动态的描述信息网络。例如，在致病基因网络中，疾病和基因之间的关联，很可能是隔一年或几年就会有新的发现。所以，除了信息网络自身的结构信息外，节点或链接上的属性信息也是很有用信息，在许多应用的挖掘任务中起着关键性的作用。

1.1.2 意义

已有大量的方法被开发出来，用于同构信息网络分析，尤其针对社交网络的研究[5]，例如排名算法、社区检测、链接预测和影响力分析等。然而，大多数这些方法是无法直接应用于挖掘异构信息网络。这是因为在异构信息网络中，跨越不同节点之间的链接代表了不同的语义，并且通常情况下异构信息网络所能覆盖的信息要比同级的同构信息网络丰富很多。一个同构信息网络通常会被视作为一个异构信息网络的投射，相较而言，投射后极大可能导致一些重要信息丢失。就像一个人类疾病和人类基因的交互网络可以通过一个更加完全且复杂的异构异构致病基因网络投射得到。然而，投射过程不可避免的是，人类基因和其他物种基因之间的关系，或人类基因和其他物种的表型关系等非常有价值的网络信息都将丢失。既然异构网络覆盖了实际应用中更全面的信息，那么研究针对异构信息网络数据挖掘任务，提出新型的且更有效数据挖掘方法是当前迫切的需求。

为什么异构网络链接预测会是一项新的研究方向？很显然，信息通过异构节点和链接是完全不同于通过同构节点和链接来广播或传递的。基于我们在异构信息网络中的链接预测研究，我们相信存在许多新的规则可以帮助我们系统性地分析异构信息网络链接预测。规则概括如下：

(1)信息通过异构型的节点和链接广播或传递。类似于大多数网络分析的研究，链接是网络数据挖掘任务中非常重要的信息传播链路。但是，在异构信息网络挖掘中，信息传播是要通过异构型节点和链接来时实现的。也即是说，如果我们要计算排名或相似度得分都需要通过哪些异构型节点和链接。在异构信息网络数据挖掘中，无论我们使用怎样的计算模型或多么精巧的算法，一条黄金法则我们需要遵守，那就是节点和节点之间是通过异构型关系相互依赖的，我们在设计挖掘算法时必须充分考虑网络中的全局信息。

(2)通过异构网络元结构进行挖掘。不同于同构网络将所有节点视作同一种类型和将所有节点都视作同一种类型，异构信息网络在我们的模型中被看作半结构化和类型化的，也即是说所有节点和链接都形成了一结构类型集合，也被称作网络模式[6]。这种网络模式代表了异构信息网络的元机构。基于这种网络模式，可以方便地对网络进行数据挖掘，并且帮助理解和分析网络中各类对象和对象之间的关系的语义。在本文中基于Meta-Path[7]的链接预测展示通过异构网络元机构进行挖掘的有效性。

(3)使用用户引导的信息网络挖掘。在一个异构信息网络中，不同类型节点之间通常存在多种显式或隐式的关系语义。例如本文中使用的基于Meta-Path的关系路径，则当具体到某一应用下时存在多条meta-path的路径，对于特定的一个实例，不同的meta-path关系路径所占的权重又不一样。我们的异构网络挖掘目的，就是要能确定正确的关系路径并能够根据不同的用户反馈赋予合适的权值。因此有效借用用户引导或反馈会是一个很有用的策略。

1.2 本文的组织结构

本文各章节安排如下：

第一章主要概述了本文研究的课题，包括课题的背景、现状，以及研究的内容和意义。

第二章：致病基因网络中的链接预测。人类基因和疾病之间的关系预测是生物信息学中重要的应用之一。利用异构信息网络，将致病基因网络中各个类型对象和它们之间的交互关系整合在一起，因此可以通过异构类型对象来预测人类基因和疾病之间的关系概率。本章主要描述，对致病基因数据构成的异构信息网络，我们提出了异构相似约束的隐因子关系预测模型。

第三章：社交投资网络中的链接预测。互联网时代，风投和众筹等形式的募资模式得到了越来越普遍的应用。本章主要叙述了由社交投资关系构造的异构信息网络，并预测投资者和集资公司可能发生投资的概率。通过系统地设计基于Meta-Path路径的拓扑特征，我们设计针对该应用环境的有监督学习框架来预测源对象和目标对象之间关系建立的可能性。

第四章：P2PLending网络中的链接预测。近年来，P2PLending在国内外取得爆炸式的发展，吸引了大量投资者和借贷用户。繁荣的资本市场对投资者而言，通过资本借贷获取收益的同时也伴随着未知的坏账风险。在本章，我们从投资者的角度出发，阐述了如何利用P2PLending平台上产生的历史借贷数据所构成异构投资网络，去挖掘投资者们的投资表现，并依据这些优秀的或差劲的投资者的投资行为来预测某项贷款的还款概率，最终依据这项指标来预测用户是否值得投资。

第五章是对全文做了总结和概括，指出了本文的不足之处，并展望了未来的研究方向。

第二章 致病基因网络中的链接预测

本章将介绍人类致病基因和多种生物的表型之间构成的异构信息网络的链接预测方法。2.1节对致病基因网络的链接预测问题和实验数据进行简介。2.2节介绍了和本文相关的前人工作。2.3节首先引出了本文借鉴的，在推荐系统应用中取得突出成效的推荐算法，隐因子模型(LFM)。2.4节详尽地阐述了我们提出的，针对致病基因网络链接预测问题的算法，我们称之为异构相似约束的隐因子关系预测模型。2.5节通过在真实的OMIM数据上进行实验，验证了我们提出的算法的预测效果，并通过和已有的同类算法进行对比分析。和同类算法的对比实验结果说明我们的算法具有一定的借鉴性。

2.1 概述

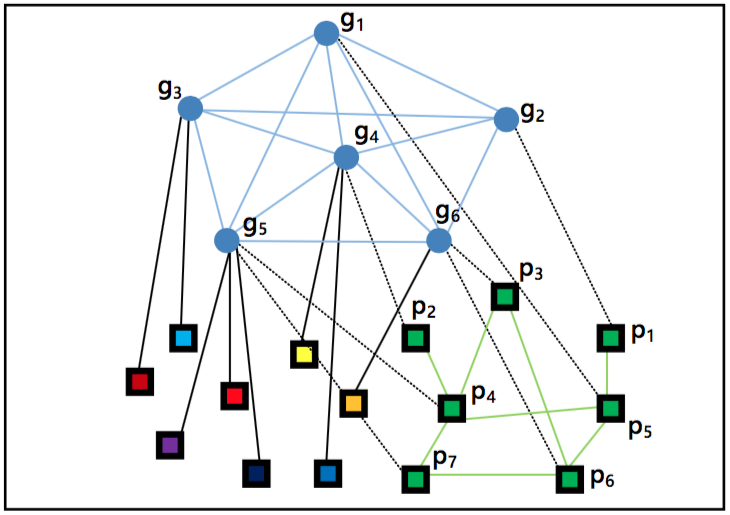
2.1.1 问题简介

在过去很长一段时间里，生物学家们一直在努力通过生物实验来寻找各个疾病与各个人类基因之间的关联，从而能够更深入地探究疾病的病理和产生根源，为疾病的预防和治疗方法的研究提供更有效的处理手段。其困难之一是基因基数庞大，就人类的基因数目就约有2万至2.5万个，要从如此之多的基因中找出疾病的致病基因，要完成大量的生物试验来做确证，这一系列的实验活动也要耗费大量的人力和物力。为了使得生物试验过程中更加的有目的性，生物信息科学家开始利用数据挖掘技术，期望利用已经证实的实验结果信息并借用数据挖掘方法，从许多基因和各类表型相关的信息中挖掘出最可能的致病基因，从而减少大量生物试验工作。

为了挖掘致病基因中隐藏的可能关系，依据生物学家整理的大量且已经验证过的致病基因相关数据，然后构造基因-疾病异构关系信息网络，如图2.1所示。网络中的节点包括人类基因、与人类基因相关的非人类生物的表型以及人类表型；网络中的虚线表示已经证实的人类基因和人类表型之间存在关系，而实线表示通过其他相关实验结果数据计算得到的节点之间存在一定的相关性，且边上的权值即为它们之间的相似值。将非人类物种的表型和人类基因的关系信息加入网络的原因是:随着研究的深入，发现其他的一些生物信息，对致病基因的预测有较大的帮助，所以在人类致病基因网络的基础上，将其他的8类物种的生物数据也加入研究，构造出生物异构信息网络。最后，基于异构信息网络，使用机器学习等方法，分析出在已有的先验信息的基础上最有可能的致病基因。

2.1.1 数据介绍

为了进行验证试验，本文使用如图2.1所示的基因-表型异构信息网络，蓝色圆点表示基因，绿色方点表示人类表型，其他颜色的方点表示非人类表型。该异构网络包括基因功能网络，图2.1中蓝色实线连接在一起的网络；基因表型关联网络，图2.1中虚线相关联的和人类表型网络，以及黑色实线相关联的非人类表型网络；人类表型相似网络，图2.1中绿色实线连接在一起的网络。下面是对异构网络中各个生物网络的具体解释：



**图2.1基因和表型异构信息网络**

(1)基因功能网络。基因功能网络也被称作HumanNet网络[8]。HumanNet网络是一个大型的基因功能网络，该网络集合mRNA表达数据，蛋白质交互数据，蛋白质复合体数据和比较基因组数据。来自21种不同数据源的733,836条记录数据规约在该网络中。

(2)基因-表型链接网络[9]。网络是对基因与表型间关系的描述。网络包括两类节点，基因节点和表型节点，而网络中的每条边表示表型和基因的链接情况。如果基因a已被实验证明是表型b的致病因素，则将会有一条虚线边c对a和b进行链接。考虑到人类与其他物种之间的直系同源关系[10]，我们不仅加入人类基因-表型链接关系(蓝色圆点和绿色方点间使用虚线连接的网络)，还加入8种其他物种的基因-表型链接关系(蓝色圆点和非绿色方点间使用黑色实线链接的网络)。具体的关于人类和8种非人类基因-表型数据的描述可查看论文[11]。

(3)人类表型相似网络[12]。网络反映表型间相似程度。Van[13]研究发现，表型间相似程度反映交互功能相似的基因间相关性，从而推导出相似网络对致病基因的预测起到正相关作用。网络中节点表示表型，而每条边具有一个权值，反映两个表型间的相似程度。可以从论文[14]中下载到该数据集，数据集包含3,209个表型的3,165,225条链接记录。

2.2 相关工作

正确的预测出新的基因和疾病关系已经成为生物信息领域重要目的。GBA是一个非常成功的方法，该方法使用的策略是一种协同策略，新的候选基因被发现是通过已知的致病基因且和该候选基因相关。GBA这种使用相关性策略的方法可以综合多种数据来做预测。Goh等[15]根据多个基因关联同一个疾病构造了一个网络。Tian[16]等人建立的网络融合了蛋白交互信息，基因交互信息，和基因表达相关信息。Ulitsky和Shamir[17]则综合了已发布的网络并依据混合网络做了多项实验。

一个使用最普遍的关联网络是来自HPRD的关于蛋白质和蛋白质的交互网络[9]。在近几年出现了许多方法，如CIPHER[18]，GeneWalker[19]，Prince[20]和RWRH[21]，通过各种不同方式，已经将只是蛋白和蛋白之间的网络拓展到了更多的相关网络。Lee等[22]人在对基于多种数据源的基因和基因之间的关联研究后，证明基因功能交互网络是一种用于预测生物功能特别有效的网络。Lee等人研究的基因功能网络已经被用于模型基因和人类表型[23]及非人类表型[24]之间关联。一个最近被公开的网络，HumanNet，已经在全组基因研究[25]中用于提高预测准确性。相较于单纯的蛋白交互网络，基因功能网络整合了多种类型的信息，拥有更全面的信息覆盖。

我们可以把基因和疾病关联问题看作监督学习问题或链接预测问题，其中每一个基因和疾病对通过多个衍生特征来表征(显式或者隐式的使用核函数)，然后使用分类器学习分类正样本和负样本参数，训练数据正样本使用已知被确证的基因和疾病关系对，负样本使用未知但根据先验知识判断和可能为不存在关系的基因和疾病对。ProDiGe[26]、Katz和Catapult[27]方法综合了广泛的异构数据集并且使用机器学习的方法来发现隐含的基因和疾病的关联。

研究内容和创新之处。在本章，我们首先简述了隐因子模型在推荐系统中的应用。接着我们详细地描述了致病基因网络链接预测中的半监督学习问题，并解释了我们如何将半监督学习问题转化为我们熟知的有监督学习问题。致病基因网络链接预测具体来说也是二分类问题，基因和疾病之间可以是由0值和1值构成的关系矩阵，而当前应用于推荐系统的隐因子模型是针对评分矩阵来设计的，所以推荐系统中的隐因子模型不能直接应用于致病基因网络的链接预测。在基础模型的介绍部分，我们详细地描述了如何将隐因子模型转换为适用于网络链接预测的隐因子关系预测模型。在模型I的介绍部分，我们结合致病基因网络的异构特点，考虑异构网络中的基因相似网络和表型相似网络对模型参数学习的影响，最终在基础模型的基础上添加了基因和表型的平均异构相似约束。在模型II的介绍部分，考虑到模型I的方法不能有效地处理哪些基因或表型之间的相似度本身存在较大差异的情形，模型II在基础模型的基础上添加了基因和表型的个性化异构相似约束。

2.3 隐因子模型

基于隐因子模型的推荐算法[28]，当前在用户和电影的评分推荐中取得了显著的推荐效果。考虑一个的评分矩阵*R*表示*m*个用户对*n*个物品的评分矩阵，基于隐因子模型的方法是找到两个在列上相同维度的矩阵和使得它们满足下式：

(2.1)

式中和可以分别表示为用户和物品在维特征空间的特征矩阵，而和的矩阵乘积近似源评分矩阵*R*。

虽然已有很多类似的矩阵分解求算法，但是我们知道，在实际推荐应用中，矩阵*R*的维度大则千万以上，小则十万，并且*R*是个稀疏矩阵，其中大量的值是未知的。如果使用传统的矩阵分解计算方法，不仅计算耗时长甚至不可行，且浪费了大量计算。因此在实际求解时，和的求解是使用最小二乘优化算法来求解的，具体优化公式定义为如式(2.2)所示：

(2.2)

式中；是一个指示函数，如果已知则，否则为0。为了使得目标函数(2)最小化，通常有两种优化方法，一种是交叉最小二乘法，一种是随机梯度下降法。交叉最小二乘法是分别对用户因子向量与项目因子向量分别求偏导，并令偏导等于零，每次迭代计算更新直到满足终止条件。随机梯度下降法的思想是先赋予一个初始点，对每一个样本让目标函数沿着负梯度的方向移动，直至移动到极小值点或者满足终止条件即可。由于交叉最小二乘法在计算上比随机梯度下降算法复杂，运算速度慢，所以使用随机梯度下降算法。

2.4 异构相似约束的隐因子关系预测模型

2.4.1 半监督学习

在我们的实验的应用数据中，给定任意一种表型，实际很难验证一个基因和该表型是完全没有关联的(即分类标签为0)。我们的模型的训练策略因此变成了没有确切证据情况下的非确切推理。我们的实验存在一个难题便是，虽然当前某个生物实验可以肯定的证明一对基因和表型之间存在关联(即分类标签为1)，但是没有足够的证据解释一个基因和表型之间是否存在关联的。因此生物学家们发布在OMIM数据库中的数据都倾向是指明某一基因和某一表型之间存在关联。既然基因和表型之间的关系是不完整的，因此我们的数据有两个关键的特性：1)对于每一种表型，我们仅仅知道很少几个和它相关的基因，因此我们只能知道表型和基因之间存在关联，但却不知道它们是一定没有关联的；2)我们已知的基因和表现之间存在关系的实例，相对于哪些未知关系的实例来说，比例是万级别的，并且从先验知识可知它们中大多数是没有关系的[27]。

传统的有监督学习方法要求训练样本即有正样本也要有负样本，因此不能直接应用于我们的实验，而这样的问题也被称作半监督学习。Liu等人[29]研究了多种不同的负样本选取方法。生物学家确信在大量的基因和表型关系对中，仅存在很少量的正样本实例。如果在所有的基因和表型关系对中随机的选取样本，那么大多数都是负样本。因此，一些简单的操作是在所有的基因和表型关系对中，随机的选取部分样例作为负样本用于训练模型。Mordelet等人提出的ProDiGe方法中，使用了基于PU Learning学习框架的负样本选取方案。

考虑PU Learning存在计算法复杂度高且操作复杂的缺陷，而简单随机抽样存在完全不可控制的因素且没有充分利用已知的先验信息，所以本文采用随机游走的方法进行负样本选取。令表示基因g可能到达人类表型p的概率，的具体定义如下:

(2.3)

式中和分别表示基因相似矩阵和人类表型相似矩阵，表示OMIM数据库中已确定的人类基因和表型之间的关系矩阵。那么我们令负样本集为，指定阈值*θ*，则。由于OMIM数据中的人类基因和表型关系对可以组合成千万级别，所以实际用于训练的是从小于或等于*θ*的实例中随机选取的，是从大于θ的实例中随机选取的，的个数占比极小。那么为什么我们要选取大于*θ*的部分的样例作为负样本呢？原因是，我们知道人类基因和表型之间是千万级别的组合，并且根据生物学家的先验知识可知，其中的绝大多数是没有关联的(即为负样本)。所以为了提高模型的泛化能力，我们加入少量的作为训练负样本。

2.4.2 基础模型:隐因子关系预测模型

在2.3节中介绍了隐因子模型在推荐系统中的应用。我们知道在推荐系统中，评分是从低到高分几个等级，但是在二分类问题中，目标值只有0和1，如果只是简单的套用式(2.2)所定义的模型，显然是不可取得。不同于用户物品评分矩阵，在致病基因网络的关系矩阵中，*R*中已知的值只有0和1，所以我们不能像式(2.2)一样使用最小二乘的优化模型。

同样，我们令矩阵和分别表示人类基因和表型在维特征空间上的特征矩阵，那么令和分别表示基因和表型之间存在关联的概率和不存在关联的概率，并且我们认为基因和表型之间如果存在关联的概率越大，则和在维特征空间上的欧几里得距离越近。因此，和的具体定义如下式所示：

,

(2.4)

式中，是一个大于0的很小数，在我们的实验中设为0.0001；由(2.4)中的定义可知，如果的欧氏距离越远则为0的概率越大，反之，若距离越近则为1的概率越大。特别地，时，；反之时，。

因此，为了学习模型(2.4)中的参数，我们可以应用极大似然估计法来估计模型参数。令和，则似然函数为

，

则对数似然函数为

(2.5)

对求极大值，得到的估计值。这样就变成了求解对数似然为目标的最优化问题，同样我们采用的方法是梯度下降法。和的梯度更新公式如下：

2.4.3 模型I:平均异构相似约束的隐因子关系预测模型

从我们对图4所描述的致病基因异构信息网络可知，基因和表型之间存在多个不同类型网络，而在基础模型中只利用基因-表型链接网络来学习基因和表型的特征矩阵，所以我们在基础模型的基础上设计了模型I。相较于基础模型，模型I更加全面的利用了更多的先验信息，将基因-基因的相似网络和表型-表型的相似网络信息加入学习约束。既然我们在式(2.5)的模型中认为在确定特征空间里，如果基因和表型之间存在关联，那么该基因和表型对之间的空间距离就应当很小；反之，若它们之间不存在关联，则空间距离就会很大。模型I的详细定义如下式所示：

(2.6)

式中，表示基因的相似邻居集合，表示表型的相似邻居集合，表示基因和邻居的相似度，表型和邻居的相似度。

从式(2.6)定义的优化函数可知，模型I是在基础模型上增加了两项关于基因和表型的相似约束：

， (2.7)

由(2.6)式我们可以知道，为了使目标函数最大化，则基因和表型约束项就应当尽可能小，也即是要使基因和表型与它的邻居们和的相似距离的中心靠近。式(278)的约束公式中和相似项说明，如果基因和邻居的相似值越大，则的相似距离中心离越近；同样，如果表型和邻居的相似值越大，则的相似距离中心离越近。反之，如果相似值和越小，则的相似距离中心离越远，的相似距离中心离越远。在2.1节已经说明的人类表型相似网络数据中，的值是可以直接得到的，而的计算我们将在2.4.5部分详细介绍。

2.4.4 模型II:个性化异构相似约束的隐因子关系预测模型

在模型I中我们加入了基因相似网络和人类表现相似网络的约束项，限制基因和表型与它的邻居们和的相似距离的中心靠近。但是，这种模型方法不能有效地处理哪些基因或表型之间的相似度本身存在较大差异的情形。如果忽视了基因或表型自身的特性或差异性，很可能会导致和得到不恰当的结果。因此，为了解决模型I的缺陷，我们提出具有个性化异构相似约束的隐因子关系预测模型，即模型II，它的详细定义如下：

(2.8)

式中，、、、等同于我们在式(2.7)中所用的。从式(2.8)中不难看出，如果和的相似值值越大，则意味着和在特征空间上的距离越近；反之，若值越小，则和在特征空间上的距离越远。类似的，如果和的相似值值越大，则意味着和在特征空间上的距离越近；反之，若和的相似值值越小，则和在特征空间上的距离越远。模型II的另一个特点是，它可以通过相似网络中节点之间的临近关系，间接的调整节点在特征空间的相互距离。更具体地，如果基因有邻居基因，并且基因有邻居基因，而基因和基因不是邻居；那么实际上通过模型II，我们会间接地最小化和在特征空间上的距离，当我们最小化下面约束时：

和

模型II的这种传递过程最终会到达一种稳定状态，即目标函数收敛。

模型II使用梯度下降法求解的局部最优化更新梯度如下：

2.4.5 基因相似性计算

在2.4.3和2.4.4节中，我们提出的模型I和II加入的基因网络相似约束项要求我们知道基因-基因的相似值。从图4所示的网络可知，除了知道基因和人类表型之间的关联网络外，我们还知道基因和其他8种非人类物种的表型之间的关联网络。因此在实验中，我们将利用基因在8种非人类物种的表型的关联特性来评估基因-基因的相似度。我们可以使用在文本挖掘中两个非常常用的相似度度量方法来计算基因之间的相似度，它们分别是余弦相似度(VSS)和皮尔森相关系数(PCC)[30]。VSS的具体定义如下：

(2.9)

式中表示8个非人类物种中和基因相关联的表型集合，表示8个非人类物种中和基因相关联的表型集合，表示非人类表型集合和公共集合中的元素，则表示基因和非人类表型之间的关系值。从式(2.9)的定义可知，VSS的相似值范围为[0,1]，并且值越大意味着相似度越高。由于在我们使用的数据中，的元素个数存在差异，因此为了保证的有效性，在实际计算的时候，我们令最终的。

不难注意，VSS的计算并没有考虑基因之间的独特性。不可否认，在和的公共表型集合中存在某些基因对所有表型关系度都很大，而某些基因却和这些表现的关系度却很小。因此，使用PCC可以用于解决这样的问题:

(2.10)

式中表示基因和非人类表型相似值的平均值。从式(2.10)的定义可知，的范围为[-1,1]，并且值越大则意味着基因和基因越相似。为了使PCC和VSS的值域范围保持一致，我们使用映射函数将PCC值映射至[0,1]范围内。同VSS的计算结果，在实际计算的时候，我们令最终的。

最后，我们通过人类基因和其它八种非人类物种的表现型之间的关系网络计算得到了基因之间的相似值，以及我们已知的通过HumanNet得到的基因之间的相似值。由于我们一直得到的不是在[0,1]范围内，所以需要先进行归一化处理，最终在实验中使用的。我们根据和的可靠性设置了倍率，并且根据我们的先验知识认为，的更可靠，所以在实验中取值将大于1。

2.5 实验分析

本章所述致病基因网络的链接预测问题，如前所述为一个半监督学习问题。由于本章所提出的算法模型是一类基于模型的协同过滤算法，所以不可避免的存在协同过滤算法中的“冷启动”问题。也即是说，当我们要预测某个疾病和那些疾病有关时，如果已知的疾病和基因关系矩阵中，该疾病和所有基因不存在有关系的元素，那么我们的基础模型就会失去效。但是不同于传统的隐因子模型，式(2.6)和(2.8)所定义的模型可以通过模型中第二项和第三项的异构约束信息，来减弱对“冷启动”问题的影响。

2.5.1 算法准确性比较

为验证我们的算法在致病基因网络中，相对已有的预测算法有效，我们和下述所列的当前已发表的使用相同数据集且相对权威的算法进行比较:

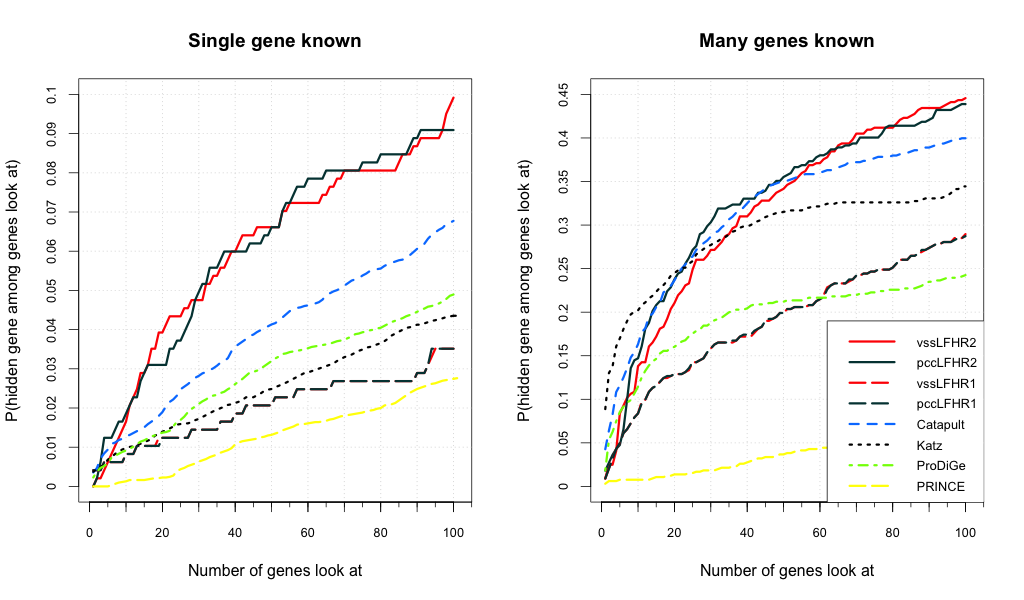
(1)Katz[27]：该方法使用路径数的度量方法，寻找网络中与指定节点较为相似的节点。该方法已经被成功用于社交网络的朋友关系预测。为了计算网络中节点i和j间的相似性，方法通过计算两节点间的不同长度的路径条数进行判断。

(2)Catapult[27]：该方法在Katz的方法上进行改进，通过机器学习方法学习不同路径的权重。此方法结合PU学习的方法，将已知的基因-表型链接看成是正例，将随机选出的未标签样例作为“反例”，通过机器学习的方法，学习分类模型。最后将学习到的模型用来进行基因-表型样例分类，从而挖掘出“致病基因”。

(3)Prince[20]：为了能够基于全局的网络信息对致病基因进行预测，Vanunu[20]提出PRINCE算法。算法只允许在基因功能网络内进行随机游走，而表型相似关系则被集合在初始向量中，所以被认为是RWRH的特殊形式。

(4)ProDiGe[26]：该方法使用多核学习（multi-kernel learning）联合框架进行多种生物网络整合，同时结合positive-unlabeled学习（PU learning）进行致病基因预测。

由于已有的实验都是使用三折交叉验证进行效果验证，所以本文也使用同样的方式进行实验效果对比。已有的算法使用的预测准确性评价标准是，对于需要作预测的基因g和表型p的关系对，运行预测算法后，计算基因g和所有表型存在关系的概率得分，并进行排序。最后根据预测的排名分布，对预测算法的预测准确性进行分析，也是表现型的召回率。为了保证验证策略的有效性，只对排名前100的预测结果进行分析。



**图2.2 算法预测准确率对比**

如图2.2所示，我们将模型I和II的预测的准确性和上述四种算法的准确性进行对比，对比结果分为单链类型和多链类型的准确性对比。所谓单链类型也即是说，我们在预测表型pi和基因gi是否存在关系时，训练数据中没有任何基因和表型pi存在关系；反之，多链类型即表示在预测表型pi和基因gi是否存在关系时，我们已发现至少一个基因和表型pi存在关系。在使用模型II和II实验单链类型和多链类型的表现和基因对时，我们设置了不同的参数。在预测多链类型时，我们令α1=α2=0.5，β1=β2=0.001，隐因子维度D=60；预测单链类型时，令α1=α2=0.05，β1=β2=0.5，隐因子维度D=10。参数，和D对模型预测结果的影响分析将在2.5.2和2.5.3部分作详细说明。

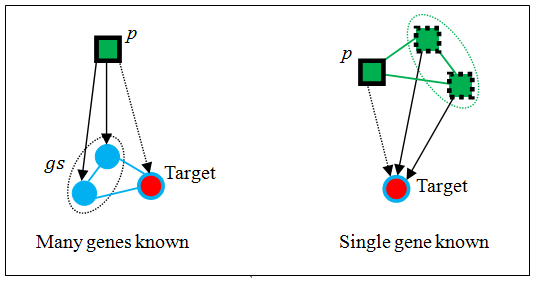
从图2.2所示的实验对比结果可以看出，模型II的表现要明显优于模型I。根据我们对模型I和II的定义和描述可以说明，基因或表型之间很可能不存在明显的群组关系；相反地，它们之间很可能是存在较高的特异性，因此个性化异构约束的隐因子模型(模型II)会表现得更优。从图2.2的对比结果显然也可以看出，我们的方法无论是在单链类型还是在多链接类型上的表现，在一定程度上都是要优于已有的方法。在取top100时，单链类型提高了超过4%，在多链接类型提高了5%(并且这不是我们的预测模型的最佳效果，2.5.2和2.5.3部分作详细分析)。此外，我们也可以看出使用VSS或PCC作为相似度量方法，在我们的方法中对总体效果的影响很小。但在多链类型图中可以看出，使用PCC的模型在top值更小时表现是要更优。

**表2.1 参数和对模型的影响比较**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **β α** | | **0.0001** | **0.001** | **0.005** | **0.01** | **0.05** | **0.1** | **0.5** | **1** |
| **0.0001** | S | 0.043 | 0.048 | 0.029 | 0.035 | 0.035 | 0.023 | 0.027 | 0.027 |
| M | 0.149 | 0.172 | 0.195 | 0.190 | 0.265 | 0.265 | 0.310 | 0.328 |
| **0.001** | S | 0.048 | 0.037 | 0.025 | 0.027 | 0.029 | 0.021 | 0.017 | 0.039 |
| M | 0.186 | 0.154 | 0.197 | 0.183 | 0.308 | 0.344 | **0.376** | 0.369 |
| **0.005** | S | 0.037 | 0.033 | 0.023 | 0.029 | 0.014 | 0.014 | 0.010 | 0.027 |
| M | 0.106 | 0.102 | 0.147 | 0.147 | 0.235 | 0.260 | 0.276 | 0.278 |
| **0.01** | S | 0.052 | 0.070 | 0.052 | 0.041 | 0.037 | 0.029 | 0.012 | 0.017 |
| M | 0.093 | 0.077 | 0.070 | 0.054 | 0.136 | 0.197 | 0.133 | 0.163 |
| **0.05** | S | 0.089 | 0.118 | 0.107 | 0.083 | 0.066 | 0.052 | 0.068 | 0.045 |
| M | 0.023 | 0.011 | 0.023 | 0.036 | 0.048 | 0.075 | 0.079 | 0.061 |
| **0.1** | S | 0.110 | 0.099 | 0.118 | 0.114 | 0.081 | 0.072 | 0.054 | 0.064 |
| M | 0.063 | 0.045 | 0.023 | 0.059 | 0.029 | 0.045 | 0.027 | 0.032 |
| **0.5** | S | **0.145** | 0.107 | 0.128 | 0.076 | 0.099 | 0.091 | 0.072 | 0.066 |
| M | 0.050 | 0.054 | 0.048 | 0.043 | 0.038 | 0.027 | 0.043 | 0.018 |
| **1** | S | 0.107 | 0.107 | 0.110 | 0.083 | 0.081 | 0.089 | 0.066 | 0.050 |
| M | 0.068 | 0.068 | 0.059 | 0.045 | 0.045 | 0.052 | 0.063 | 0.048 |

2.5.2 参数和的影响

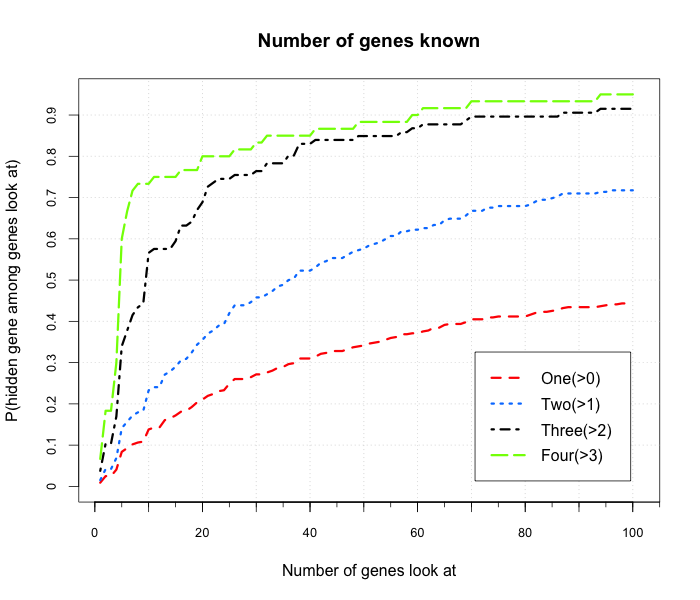
在我们提出的方法中参数和对预测结果的好坏起了很重要的作用，控制基因信息对我们的模型的影响，而则控制了表型对我们的模型的影响。本节我们将讨论参数和对模型I和II的影响，并且由于参数和在模型I和II中的作用是类似的，所以我们只实验分析模型II。此外，相似度计算函数也是类似的，所以只讨论使用其中一种的情况；再由于我们主要是分析参数和对模型的影响，所以为了减少计算量，我们令多链时隐因子维度D=30，单链时D=5(详细地分析见2.5.3)。



**图2.3 单链和多链类型学习过程**

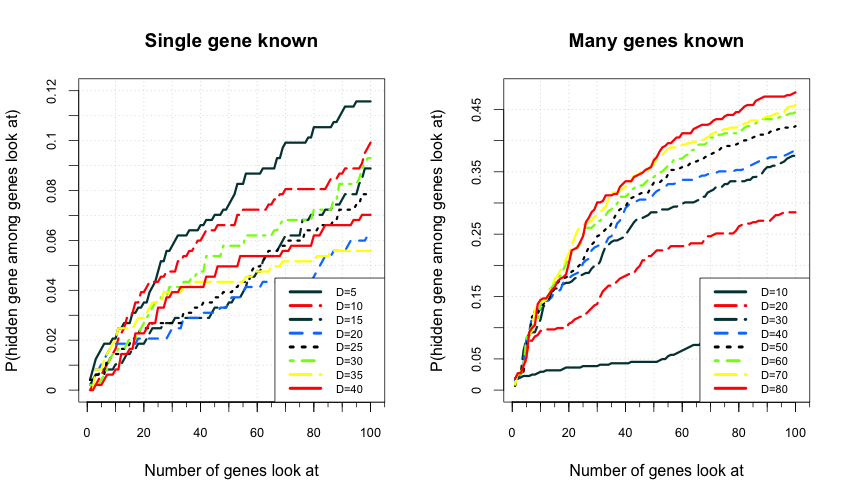
如表2.1中所示的top100时的预测准确率，我们可以很明显地注意到，当相对的权重增大时(表2.1的右上角多链部分)，算法对多链接类型的基因和表型对的预测准确得到显著提升；相反，在表2.2中，当相对的权重增大时(表的右上角单链部分)，算法对单链接类型的基因和表型对的预测准确显著低于相对的权重较小的部分(表2.1左下角单链部分)。出现上述结果的原因也很符合实际，因为对于如图2.3所示多链接类型，意味着表型和一些基因()已经存在关系，则哪些和基因相似的邻居基因(Target)就更有可能和表型存在关系，所以这时基因信息显然就要比表型信息重要许多。另一方面，对于单链接类型，因为不知道表型和哪些基因存在关系，这时和表型相似的邻居表型有关系的基因(Target)则有可能和表型存在关系，于是这时表型信息就显得重要了许多。

从上述我们对异构相似约束的隐因子关系预测模型的分析，按照我们实际中直观的理解，如果我们知道和表型存在相关的基因越多，则对表型的链接的预测准确性就会越高。为了解释这条结论，我们令α1=α2=0.01，β1=β2=0.5，隐因子维度D=60，使用VSS的模型II预测多链接时，在知道不同条数链接条件下的预测准确性，如图2.4所示。图中四条线分别表示知道至少1条、2条、3条和4条时，模型II的预测准确性对比。特别地，我们可以注意到，当知道至少3条和4条时，预测准确率在top10时上升得非常快。



**图2.4 不同链接数的预测准确率对比**

2.5.3 隐因子维度D的影响



**图2.5 不同特征维度的预测准确率对比**

隐因子维度D的选取在很大程度上会影响模型I和II的预测准确性。如果维度太小，虽然计算复杂度低了，但模型可能欠拟合预测准确性低；反之，维度太大了，计算复杂度将增加，而准确性可能得不到理想的提升，并且也有可能由于过拟合而使准确性降低。实验时使用VSS作为相似度量函数，在预测多链类型时，我们令α1=α2=0.5，β1=β2=0.001；预测单链类型时，令α1=α2=0.05，β1=β2=0.5。实验结果如图2.5所示，在预测单链类型的疾病-基因对时，我们看到，低维的特征模型相对高维的表现要更好，造成这种结果的原因很可能是，由于缺乏有效的训练数据，导致模型过拟合而降低了泛化能力；相反地，在预测多链类型时，模型的表现随着特征维度的增大而提高，并且在D=80、取top100时比Catapult提高了超过7%，但是计算复杂度增长的同时提升速度也在减缓。所以，根据我们多次的实验结果分析，得出的经验建议是:如果有效数据充足且没有严格的时间限定的情况下，可以适度增大D值；反之，应当设置较小的D值。

2.6 本章小结

在本章，我们针对致病基因网络的链接预测问题，提出了使用异构相似约束的隐因子关系预测模型。具体地，我们首先将预测疾病和基因之间关系的问题转化为半监督学习问题。接着，根据隐因子模型的关于对象之间在相同特征空间里，距离越近则越相似的算法思想，我们将疾病和基因之间的关系预测转化为二分类概率模型：疾病和基因如果在隐含特征空间的距离越近它们存在关系的概率越高；反之，若距离越远，则存在关系的概率越低。为了充分利用异构信息网络所覆盖的丰富语义而更准确的挖掘基因和疾病之间可能存在的相关性，我们在基础模型的基础上加入了异构相似约束，即平均异构相似约束和个性化异构相似约束的隐因子模型。最后，为了验证提出的算法的有效性，我们在真实的生物数据上进行实验，并和已有的相关算法的预测准确性进行了比较。

第三章 社交投资网络中的链接预测

3.1 概述

3.1.1 问题简介

以互联网为媒介的新型投资模式众筹(crowdfunding)，近年来在国内外飞速发展。通过互联网以众筹的形式，从大量网络用户中为一个工程项目或创业活动等筹集资金，国内常见的互联网众筹渠道有微信、支付宝等。因此，在互联网众筹领域，预测投资者投资行为已然成为另一个新的行为预测问题。

已有大量的学术研究探讨了投资行为。已有的研究发现影响人做出投资行为的因素有许多种，心理因素[31]、地理因素[32]、投资经验[33]甚至基因[34]等因素都已被相关学者解释为影响投资决策。此外，一些研究员在基于投资者和公司之间构造的同构关系网络的基础上，萃取多种社交网络特征，然后依据从社交网络中萃取的同构特征来建立同构网络链接预测模型[35]。同构(社交)网络是将网络中的所有节点都看作一种类型(如人)，并且网络中的所有链接也都被视作一种类型(如友谊)。然而实际中，大多数网络是异构的，应该存在多种对象和链接类型。就同构网络而言，一方面，同构网络把所有网络节点看作同一种类型很可能丢失重要的语义信息；另一方面，如果将所有节点都看作同一种类型，那么重要的模式信息也很可能无法获取。因此在投资网络问题中，相较于投资者等其他实体类型而言，将公司单独看作一种实体类型会比把网络所有的实体都看作一种类型更有意义。因此，异构信息网络将可能更有效地捕获投资网络中的本质语义特征。

目前已有一些方法[36]用于预测在一个投资网络中，投资者节点和公司节点之间在未来是否会出现链接(投资者投资某公司)。这些已有的方法都是基于当前或历史数据构造的信息网络，并将问题模型为类似于社交网络的链接预测问题。此外，这些链接预测方法都针对同构信息网络而设计的。在本章，我们将同构信息网络拓展为异构信息网络，把同构网络链接预测问题转为异构网络连接预测问题。

3.1.2 数据介绍

本章实验所使用的CrunchBase数据是一个开放的数据集。CrunchBase数据集包含创业公司信息、投资者信息、创始人信息、并购信息、上市公司等其他相关信息。CrunchBase依靠社区人员提供信息以及自发的填写和修改上面的信息。截至到2014年5月，CrunchBase中包含了46,015条公司(或机构信息)信息，106,075条投资记录，以及12,068条并购记录。CrunchBase主要包含以下实体类型和关系类型。

*实体类型*

投资者(Investor)。投资者在CrunchBase数据集中的构成有个人和公司(或者机构)，例如Carret Camp，Google等。在接下来的计算式中，为了表达方便，我们使用表示投资者集合和表示投资者*i*。

公司(Company)。公司在CrunchBase数据集中的标签就是公司(或机构)，例如Google，Uber，AOL等等。在接下来的计算式中，为了表达方便，我们使用表示公司集合和表示公司*i*。

产品市场(Category)。在当前Crunchbase收集的数据集中，包含741种类型的产品市场。例如谷歌公司的产品市场有软件、搜索、视频等。同样为了便于表达，在接下来的计算式中，我们使用作为产品市场种类集合和表示类别*i*。

城市(City)。许多产业都有地理特性，例如北京和杭州互联网公司活跃，深圳硬件设备制造公司相对活跃，上海则是金融行业相对活跃等。一项研究[32]揭示地理因素不同也会影响投资者的投资决策。在接下来的计算式中，为了表达方便，我们使用表示城市集合和代表城市*i*。

*关系类型*

投资关系(Investment)。一条投资关系记录的创建是在一个投资者(Investor)和一个公司（Company）之间发生了投资行为后被创建。例如，google公司在2005年12月投资了AOL公司。在接下来的计算式中，为了表达方便，我们使用表示投资关系集合，并且代表投资者*i*投资了公司*j*。

市场关系(Market)。市场关系表征一个公司的产品属于哪个产品类别(Category)。例如Uber的业务主要包括汽车和交通业务。在接下来的计算式中，为了表达方便，我们使用表示市场关系集合，并且表示公司*i*的业务包括产品类别*j*。

地理位置关系(Location)。地理位置关系指示某一公司的总部坐落位置。例如，Uber的中部位于San Francisco。在接下来的计算式中，为了表达方便，我们使用表示地理位置关系集合，并且代指公司*i*坐落于城市*j*。

收购关系(Acquisition)。如果一个大公司收购了一个小公司或者两个大公司强强联合，则表示一条收购关系记录。例如，google公司在2014年5月收购Adometry公司。在接下来的计算式中，为了表达方便，我们使用表示收购关系集合，并且指示公司*i*收购公司*j*。

3.2 相关工作

Eugene和Daphne[37]基于投资网络中投资者和公司之间的社交关系相似或者不相似来评价投资者投资某一公司的概率(被预测为正类或负类的概率)。他们设计的方法是基于同构网络的链接预测模型，但是实际中的投资网络都是异构的，网络应该存在多种节点类型和链接类型。我们在本章提出的方法即是基于异构信息网络的。Guang等人[38]的研究通过使用来自CrunchBase数据集以及来自TechCrunch上的栏目信息和新闻内容，并利用从这些信息中获取的主题和概要特征来预测公司并购活动。Guang等人的工作主要是针对不同领域的收购或者并购活动，他们的工作也没有考虑影响投资行为的相关因素。

Sun和Han[39]提出了基于meta-path拓扑特征的异构信息网络链接预测方法。在异构信息网络中，为了测量跨过不同链接类型的节点之间的对象类型的相似度，PathCount，NormalizedPathCount[40]，RandomWalk[41]等，针对基于meta-path异构信息网络相似度度量函数被提出。此外，Shi和Kong[42]提出了一个种新型的基于meta-path的异构信息网络相似度量函数，称作HeteSim。使用HeteSim的相似度量计算同时也被应用于致病基因异构信息网络。

在本章中描述的方法，是我们首次提出使用基于异构信息网络的meta-path作为主要特征方法，去预测一个投资行为是否会发生。例如，给定一个投资者和一个公司，我们会去预测是否投资这个公司。具体地，我们会去计算影响投资者和公司之间发生投资关系的多种异构因素所构成的meta-path相似特征，并使用一套基于时间序列的机器学习框架来预测未来某个投资者和某个公司是否会发生投资关系。实验结果揭示，我们的模型可以作为一种有效的方法为公司寻找投资目标，因为这些公司和潜在的投资者们在各个方面都表现了很高的相似性。

3.3 基于Meta-Path的链接预测

在这一部分，我们将介绍如何使用CrunchBase数据集构造的异构投资信息网络来预测投资行为。具体地，在下文中详细解释，我们怎样将投资者和公司之间的投资行为模型为异构网络链接预测模型，并且说明如何测量异构投资网络中投资者和公司之间的meta-path相似度。

3.3.1 社交投资异构信息网络

异构信息网络模式是信息网络中的一种特殊类型，不同于传统的同构信息网络将数据抽象为一个单一的有向或无向的图。我们给出异构信息网络的定义如下：

**定义2** (异构信息网络模式) 给定一个模式是一个异构信息网络的一个元模板，其中网络节点类型到实体类型的映射函数是，函数表明对于网络中任意类型节点满足；同样，网络链接类型到关系类型的映射函数是，函数意味着任意类型链接满足；此外，是一个由节点类型和关系类型构成的有向图。特别地，在异构信息网络中，节点类型或关系类型的值，至少有一个大于1。

在CrunchBase数据集构成的异构信息网络中，实体类型和关系类型分别是和。代表和之间的投资关系，记作。表示源类型，表示目标类型，它们之间产生的关系是。从到的路径可以认为是的逆关系，记作。

使用CrunchBase数据集，我们使用3.1部分介绍的实体和关系类型构造了一个异构信息网络，如图1.2所示。图1.2中一个节点代表一种实体类型，边指示一种关系类型。图1.2即是本章用于验证所提出的方法有效性的异构信息网络模式。为了构造图2所示的异构信息网络模式，我们首先需要选择作为节点的实体类型，然后依据这些实体类型之间的关系将这些节点连接在一起。图1.2所示的网络几乎覆盖了CrunchBase数据集中可用的所有和投资相关的信息，除了一些URL、序列码等信息。网络建立起来之后，我们即可依据网络萃取meta-path，并依据meta-path来模型投资者和一个公司之间的相似度。在接下来的部分，我们将详细地解释社交投资网络的meta-path和它的相似度量函数。

3.4 异构社交投资网络的Meta-path

在这一部分，我们将介绍meta-path相关概念以及如何将它应用到异构社交投资网络中。Meta-path是一种链接异构信息网络中两种类型节点的特殊路径，它的具体定义如下：

**定义3** (Meta-path) 给定一个meta-path 是一条定义在图模式中的一条路径，记作形式；令混合关系表示节点类型到的关系，表示关系运算符。

**表3.1 路径长度不大于4的Meta-path**

|  |  |
| --- | --- |
| Meta-path | Semantic Meaning of the Relation |
| *Inv-Com-Inv-Com* | 通过的邻居来计算和的相似度 |
| *Inv-Com-Cat-Com* | 通过感兴趣的市场领域来计算和的相似度 |
| *Inv-Com-Cit-Com* | 通过和相关的来计算和的相似度 |
| *Inv-Com-Com* | 通过的邻居来计算和的相似度 |
| *Inv-Com-Com-Inv-Com* | 通过的邻居和的邻居来计算和的相似度 |
| *Inv-Com-Cat-Com-Com* | 通过感兴趣的市场领域和的邻居来计算和的相似度 |
| *Inv-Com-Com-Cat-Com* | 通过的邻居和的产品市场来计算和的相似度 |

从定义3可知，一个meta-path是一条基于异构信息网络的包含多种关系的路径。哪些定义在同构信息网络中，基于邻居和基于路径来度量两个节点之间相似性的特征，都可以通过异构信息网络中不同的meta-path来一般化的表达[43]。从图1.2中所示的网络，我们萃取了所有路径长度不大于4的重要meta-path路径。这些路径的起点都是*Inv*节点并且终点都是*Com*节点。在实验部分，我们会更详细地解释meta-path长度越长，在计算投资者和公司之间的相似值时的意义越弱。所有从*Inv*节点到*Com*节点的meta-path都可以通过遍历图1.2所示的网络模式来获取。因此，我们我们在表3.1中概述了所有长度不大于4的重要meta-path。

3.5 Meta-path相似度量函数

一旦我们获取了基于meta-path的拓扑特征(见表3.1)，那么用于定量的计算meta-path相似度的有效度量函数应当被提出。

公式(3.1)被称作HeteSim[42]，计算在meta-path 上和之间的相似度。HeteSim可以解释为分别从源对象和目标对象出发，并在路径的中间位置相遇的cosine概率分布。HeteSim的取值范围为0到1。

(3.1)

式中是meta-path 下的可达路径矩阵，并且表示在meta-path 模式下对象到达对象；表示矩阵的第a行；并且是通过式(3.2)来计算的。我们令表示的矩阵转置。和分别表示meta-path 的左右部分。

(3.2)

在式中表示异构信息网络中邻接关系矩阵的归一化形式，是邻接关系矩阵集合中的一个元素。表示关系矩阵转置，或者说两个对象之间的反向关系矩阵。特别地，在我们的工作中，我们使用图1.2所示信息网络模式，则必须是。

PathCount表示的是依据meta-path ，从节点出发到达节点的所有路径条数，它的具体定义为等式(3.3)：

(3.3)

式中表示从节点出发到达节点的一条路径。

RandomWalk表示在meta-path 下，从节点出发到达节点的随机游走概率，并且它也表示从出发到达的所有路径的概率之和。RandomWalk的具体定义为等式(3.4)：

(3.4)

Features

Labels

**Training**

Features

Labels

**Test**

**图3.1 投资者行为预测框架**

3.6 投资网络链接预测

如图3.1所示，预测框架由两部分构成，训练和测试部分。给定投资者和公司之间构成的关系对作为训练实例，我们首先在时间间隔内，从聚合网络中提取相关的异构信息网络特征。接着，我们记录投资者和公司关系对在未来时间间隔的实际发生情况，并作类别标签。然后，我们就可以构造预测模型来学习这些相关异构特征的权值。

3.6.1 链接预测模型

我们的目标变量是预测一个投资者在未来一段时间内是否会投资一个特定的公司，记作。在我们的工作中，使用logistic regression模型作为预测模型。对于每一个投资者和公司构成的关系训练对，记作，我们认为它们服从概率值为()的Bernoulli分布:(1)令表示维向量，其中包含常数1和个投资者和公司之间异构特征值；(2)是分类标签值，表示投资者*i*在未来指定的某段时间内是否投资了公司*j*(如果(如果它们之间发生了投资关系，则；否则)。发生概率的定义具体为等式(3.5)。

(3.5)

式中是一个为回归系数向量，对应的自变量分别是常数项和各个异构拓扑特征。这样回归系数就可以通过常规的优化框架来学习，如最大对数似然估计法。优化公式的具体描述如下：

(3.6)

对于式(3.6)，我们可以使用随机梯度下降法来求解回归系数。

3.6.2 链接建立时间预测模型

已有的一些研究都是将投资行为预测看作链接预测，并且关注的都是预测链接在未来某段固定的时间是否会建立。例如，“是否一名投资会投资某一个公司？”然而，我们可能更感兴趣的是，未来的什么时间链接会被建立？

在这部分，我们提出了基于GLM[44]的预测模型。我们希望可以直接将投资行为发生时间模型为关于meta-path的函数，并且提供优化算法来学习模型参数。同时，我们也依据不同的假设来模型投资行为发生时间服从不同的概率分布。

GLM的主要思想是将响应变量模型为，并且是关于各个meta-path的线性关系表达的函数(链接函数)。各个meta-path的线性关系式为，其中是维的向量，包括常数项1和个关于投资者和公司之间的meta-path的相似特征值；回归系数向量。那么我们的目标即转为通过训练数据来学习参数。依据对服从不同的概率分布假设(通常都是指数家族函数)，有不同形式的参数集，并且链接函数的形式也不一样。特别地，logistic regression模型是GLM的一种，其中响应变量服从Bernoulli分布。

我们首先考虑的是指数分布[45]，因为指数分布是在模型事件发生时间中使用最频繁的分布。指数分布函数的概率密度函数定义为等式(3.7)。

(3.7)

式中，，其中表示事件发生的平均等待时间。那么，指数分布的分布函数定义为式(3.8)。

(3.8)

然后，我们再考虑Weibull分布[26]，该分布是指数分布的一般化。Weibull分布的概率密度函数定义为：

(3.9)

式中，，；其中表示时间发生的平均时间，表示事件在随着时间的推移发生的概率大小。具体地，当时，意味着事件发生的概率随着时间推移事件概率是递增的(如果一个事件，现在没有发生，那么在之后的时间发生的概率将会增大)；当时，表明事件发生的概率是随着时间的推移是递减的(如果一个事件，现在没有发生，那么在之后发生的可能性将变小)；特别地，当时，Weibull分布将退化为指数分布，并且平均等待时间为，时间发生的概率不随时间先后改变。Weibull分布的分布函数定义如下：

(3.10)

*模型为Weibull分布*。因为指数分布是Weibull分布的特殊情形，那么我们只需要考虑基于Weibull分布假设的链接预测模型；此外，我们认为分布模型中，即如果一个投资者当前一段时间内没有投资一家公司，那么在未来的一段时间发生的概率将会增大。

在基于Weibull分布的情形下，我们假设投资者和公司之间构成的训练样本对的投资行为发生时间是相互独立的，并且都服从同一的Weibull分布，但是它们的平均等待时间是不一样的。依据上述假设，我们即可评估各个随机变量的期望值，即为。然后我们可以使用链接函数，则，其中代表常数项。那么为了估计参数值，我们可以使用最大对数似然估计，优化公式定义如下：

(3.11)

式中和都是指示函数，函数表示如果链接建立时间满足指定条件则函数值为1，否则为0。等式(3.11)的意义是：如果链接建立时间是在时间间隔内发生，则使用密度函数；否则使用概率值，在我们的实验中，我们置为0。通过将等式代入(3.11)，我们可以得到关于和对数似然函数为：

(3.12)

式即表示基于Weibull分布假设的指数似然函数。

这样我们的学习模型就是一个优化问题，目标是找到和使得指数似然函数最大化。我们使用常用的Newton-Raphson [27]法来求解目标优化问题。

3.7 实验分析

为了验证我们提出的模型的表现，我们在CrunchBase提供的真实数据上进行实验，和已有方法的对比实验结果表明，我们在一定程度上取得了更好的表现。和先前的研究的评价指标一样，我们使用AUC指标来对比已有的实验效果和我们的实验效果。特别地，该实验的样本数据的正负样本比例严重失衡，负样本数远远大于正样本数。因此，在试验中，我们使用等概率分布来随机选取负样本，使得正负样本比例达到正常水平。

3.7.1 综合实验效果比较

我们比较了同构网络和异构网络meta-path的实验表现。总体上，所有算法在AUC指标的表现都超过了基准的表现(0.6)。在实验中，对于我们从异构投资网络中萃取的meta-path，我们分别使用了*HeteSim*，*PathCoun*t和*RandomWalk*来测量列在表3.1中的七条meta-path相似值。最终的对比实验结果如图3.2所示。

**图3.2 基于同构和异构网络的投资行为预测表现比较**

图3.2中HomoDT， HomoSVM和HomoNB将投资行为预测模型为同构网络链接预测问题，然后基于同构网络提取链接预测相关的网络特征，并且分别是使用决策树、SVM和朴素贝叶斯作为学习算法。*HeteSim*，*PathCount*和*RandomWalk*是分别用于计算meta-path相似的，并用3.6.1部分讨论的链接预测模型来作预测。实验结果表明，使用*HeteSim*和*PathCount*作为meta-path相似度测量函数所取得的预测效果要比基于同构信息网络的预测效果好。

3.7.2 Meta-path意义探索

在这一部分，我们将分析在CrunchBase数据集中，各个meta-path拓扑特征在学习模型中对于预测投资者和公司之间投资关系的显著性意义。在表3.2中，我们列出了所有7条meta-path在不同相似测量函数计算下相似特征的回归系数。

**表3.2 基于不同相似函数计算的 meta-path特征显著性意义**

|  |  |  |  |
| --- | --- | --- | --- |
| Meta-Paths | 𝓅-value | | |
| **HeteSim** | **PathCount** | **RandomWalk** |
| *Inv-Com-Inv-Com* | < 2e-16(\*\*\*) | <2e-16(\*\*\*) | <2e-16(\*\*\*) |
| *Inv-Com-Cat-Com* | < 2e-16(\*\*\*) | <2e-16(\*\*\*) | <2e-16(\*\*\*) |
| *Inv-Com-Cit-Com* | < 2e-16(\*\*\*) | <2e-16 (\*\*\*) | <2e-16(\*\*\*) |
| *Inv-Com-Com* | 0.0218(\*) | 0.9964( ) | 0.0556(.) |
| *Inv-Com-Com-Inv-Com* | 8.79e-7(\*\*\*) | 0.0335(\*) | 0.8813( ) |
| *Inv-Com-Cat-Com-Com* | 0.0968(.) | 0.3801( ) | 0.2682( ) |
| *Inv-Com-Com-Cat-Com* | 0.0807(.) | 0.0180(\*) | 0.2239( ) |

0 : '\*\*\*'; 0.001 : '\*\*'; 0.01 : '\*'; 0.05 : '.' ; 0.1 : ' '.

根据表3.2所展示的结果，我们概括了如下结论：（1）相对于使用RandomWalk计算meta-path相似性，使用HeteSim和PathCount计算的各个meta-path相似值，在显著性排名上更相近；（2）总体上，在投资行为预测时，HeteSim是最好的相似测量函数，因为Inv-Com-Inv-Com、Inv-Com-Cat-Com、Inv-Com-Cit-Com、Inv-Com-Com和Inv-Com-Com-Inv-Com的𝓅-value说明它们和目标变量都有显著的相关性；因此我们也推测公司产品类型、投资经历和地理因素[32]对于投资者的投资决策都有重要影响，而由于Inv-Com-Com显著性很弱则认为并购因素对投资决策影响很小；(3)最后，根据六组使用不同meta-path相似度测量函数的实验结果，我们发现meta-path的重要性会随着路径加长而变弱。

**图3.3预测未来不同时间段内的F-score表现**

%

months

3.7.3 预测链接建立时间

我们在3.6部分提出了基于GLM的时间预测模型。我们的预测模型是学习一个关于meta-path拓扑特征的函数，并预测链接建立的时间。其中模型是根据链接建立时间服从不同分布的假设，通过使用最大对数似然估计法来学习各个meta-path拓扑特征的回归系数。

**表3.1 具体的时间间隔表**

|  |  |  |  |
| --- | --- | --- | --- |
| Group |  |  |  |
|  | 2007/11-2010/11 | 2010/11-2013/11 | 2013/11-2014/05 |
|  | 2007/05-2010/05 | 2010/05-2013/05 | 2013/05-2014/05 |
|  | 2006/11-2009/11 | 2009/11-2012/11 | 2012/11-2014/05 |
|  | 2006/05-2009/05 | 2009/05-2012/05 | 2012/05-2014/05 |
|  | 2005/11-2008/11 | 2008/11-2011/11 | 2011/11-2014/05 |
|  | 2005/05-2008/05 | 2008/05-2011/05 | 2011/05-2014/05 |

为了验证在异构信息网络中基于meta-path拓扑特征预测链接建立时间的模型，我们设置了六组不同的时间间隔来进行实验结果对比。(具体可参见表3.3)

**表3.2 模型泛化能力和预测准确性对比**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| months | HeteSim (%) | | | PathCount (%) | | | RandomWalk (%) | | |
| Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| 6 | 99.4 | 45.8 | **62.7** | 23.3 | 14.9 | 18.2 | 99.5 | 3.1 | 6.0 |
| 12 | 99.5 | 40.6 | **57.6** | 21.6 | 14.0 | 17.0 | 98.3 | 3.5 | 6.7 |
| 18 | 99.6 | 35.8 | **52.7** | 20.9 | 14.1 | 16.8 | 99.7 | 3.7 | 7.1 |
| 24 | 99.4 | 38.4 | **55.5** | 20.0 | 10.0 | 13.3 | 99.5 | 4.3 | 8.2 |
| 30 | 99.6 | 35.3 | **52.1** | 24.4 | 8.7 | 12.8 | 84.5 | 5.3 | 10.0 |
| 36 | 99.5 | 32.1 | **48.5** | 23.9 | 5.3 | 8.7 | 98.3 | 5.1 | 9.7 |

在推荐系统中，F1-score[18]是在用于评价推荐质量时使用最广泛的评价指标。类似地，为了对比评估我们提出的预测链接建立时间的模型的表现，我们也使用F-score作为评价函数。在图3.3中展示了6组在不同时间间隔下，使用不同meta-path相似度测量函数的F-score的实验结果比较。结果说明我们的模型就时间因素而言，在一定程度上有不错的泛化能力。此外，不出意料地，使用*HeteSim*的预测表现要比使用*PathCount*和*RandomWalk*都好。在表3.4中，我们详细的展示了各组实验的正确率和召回率。实验结果说明，*HeteSim*在召回率上的表现要好于*RandomWalk*，而正确率和召回率都要好于*PathCount*。

**图3.4 不同预测模型的F-score比较**

months

F-score

*不同预测模型的预测表现比较。*

在3.6部分我们解释了不同的概率分布假设，指数分布和Weibull分布。为了说明使用基于不同分布假设下，模型对于网络链接建立的时间预测的表现，我们使用了logistic regression模型作为预测结果参照基线。注意，logistic regression的输出值是一个概率值，代表投资者和公司之间的投资关系对在时间间隔建立的概率。而基于指数分布和Weibull分布的模型的输出值是表示投资者和公司之间的投资关系的建立时间。因此，为了使用F-score来评价各个模型的表现，我们定义式(3.13)所示指示函数：

(3.13)

式中表示预测结果正确，错误则为0，并且所有模型都是用*HeteSim*来计算meta-path相似值。

在图3.4中，我们可以看出，在更长时间段的预测表现上，指数分布和Weibull分布的模型要好于logistic regression模型。此外，指数分布和Weibull分布的模型除了能给出某段时间内投资关系发生的概率外，还可以提供更丰富的时间信息。

3.8 本章小结

在本章，我们设计了一个数据驱动的社交投资网络链接预测框架，该框架将投资行为模型为基于异构信息网络下使用meta-path的链接预测问题。特别地，基于实体类型和关系类型，我们首先建立了一个异构关系网络模式；然后在这个异构关系网络上，我们提取了所有的有意义的路径长度不大于4的meta-path，并提供了有效的度量函数来量化meta-path相似度。最后，我们提出了一个有监督的社交投资网络链接预测学习模型，来预测投资行为。实验结果揭示，我们的模型在一定程度上也可以是一个有用的工具，帮助企业(1)更好的掌握投资者是怎样和何时会投资，(2)更好的提供参考信息当他们需要寻求外部投资时。

第四章 P2P信贷网络中的链接预测

4.1 概述

4.1.1 问题简介

P2P信贷是一个新兴的金融市场。近年来，吸引了大量的人们加入到这个互联网在线金融平台。例如在美国影响力最大的两个P2P信贷平台Prosper和LendingClub，它们的月交易量在2014年分别达到5千万美元和1亿美元，并且还处在高速增长阶段。P2P信贷已在全球范围内高速发展。国内成长良好、影响力大的有拍拍贷、人人贷和宜人贷等，欧洲有Zopa等。相较于国内借贷平台，国外借贷平台数据更开放，可获取的信息更全面，所以本章的研究数据是来自美国的Prosper。在Prosper平台，每个人在该平台即可借钱给他人也可以发布标的贷到款。用户可以向平台发送请求，申请贷款金融和愿意支付的贷款利率，平台接收请求并审核通过后，即可发布一条贷款标的。所有的投资者都可以看到哪些发布在平台上的标的，以及这些标的的发起人的身份、信誉和教育背景等信息，投资者可以根据自己了解到的信息来判断是否要给这些标的投标。P2P信贷和传统的商业银行放贷的主要不同是，在P2P信贷平台，所有投资者不仅可以了解到每一项贷款的金融信息，也可以依据借贷者的社交特征来评估投资风险。此外，Prosper作为一个p2p中间平台机构，它可以让大量的投资者和借贷者方便的参与，有效地帮助用户借到钱或通过投资赚取收益。

在Prosper这样的互联网平台，投资者和借贷者构成了一个*M-to-N*的异构关系模型，也称作二部图；在这样一个平台中，一个投资者为了分散风险可以投资*N*个贷款，并且一个项标的可以接受来自*M*个投资者的投标。本章的目的即是通过利用二部图形式的异构网络模型，预测P2P信贷平台上的贷款项目是否值得用户投资。

4.1.2 数据介绍

在我们的实验中，我们使用从Prosper下载得到的数据，并选择从从2009年1月至2010年1月的数据集作为训练数据，并且选择从2010年1月至2010年6月的数据作为测试数据集。之所以选择这段时间是因为Prosper给出的数据集在这段是时间内是相对稳定的。我们从给出的数据集中选出两类最重要的表，bids表和loans表的数据作为训练和测试我们的预测模型。

在bids表中，它主要包含Memberkey，ListingKey，Amount和CreationDate字段。这些字段分别表示投资者ID，贷款项ID，该次竞标的金额，投标时间。每一条竞标记录的产生是在投资者确定了给某一标的投标后生成的。loans表主要包含的字段有ListingKey，Status和CreationDate。这些字段分别表示贷款项ID，贷款状态(如已完成还款、正在还款和正在招标)，贷款项创建时间。利用以上信息和相关关系，我们既可以用于构造投资者和贷款项之间构成的二部投资关系网络。

4.2 相关工作

P2P信贷作为一个新兴的金融市场，已然成为学院研究的新领域。P2P信贷作为互联网形式的创新金融市场，许多研究者主要关注用户的社交关系以及贷款的属性(如贷款金额、还款利率以及贷款时间等)。也有一些研究员设计模型分析怎样能够更成功的申请到贷款。

近几年，基于P2P信贷平台的社交网络吸引了不少研究员投入研究。Berger和Gleisner[46]研究发现这些网贷市场的中间平台作为金融中介对于提高借贷者的信誉可以起到重要作用，因为P2P信贷平台可以减弱信息不对称的问题，尤其针对那些没有投资吸引力的高风险借贷者。Freedman和Jin[47]分析了Prosper中存在的信息相关问题，并且研究了是否社交网络可以帮助减弱这些信息问题。他们发现在组内的贷款期望回报率要明显低于哪些不属于组内的贷款，其中一部分原因可能是投资者的知识背景，也有一部分是由于Prosper减少了组内投资者的回报。Lin等人[48]验证了是否社交网络可以帮助获取更好的投资收益。Lin等人的方法主要是分析网络中借款成功概率更高的标的、风险更低的标的以及低回报率的标的之间的结构和关系差异。Collier和Hampshire[49]对于如何加强社区名声，从委托代理的视角到定量分析都做了研究。Sergio[50]利用一对一和一对多的关系，定量的分析了社交活动对贷款请求的风险评估作用。他的结果说明，当金融特征不足以保证贷款成功时，培养社交能力可以增加贷款成功的概率。Chen等人[51]提出了一种方法来度量在线P2P信贷平台的企业文化方向。

一些研究者从借贷者的角度出发设计研究模型。Wu和Xu[52]根据P2P信贷平台的智能受理员，为借贷者设计了一个决策支持系统。这套系统可以为借贷者提供个人风险评估，合适投资搜索，债务合并和贷款推荐。[53]的定量分析结果说明，借贷者的决策(如贷款金额、还款利率等)会决定他或她的贷款能否成功完成。Herzenstein等人[54]发现了哪些影响贷款决策的因素以及那些因素会影响贷款申请者申请成功。Pope和Sydnor[55]发现和白人有同等信誉的黑人的贷款标的相比，能够成功申请到贷款的可能性要降低25%到35%。

为了帮助投资者们更好的做决策，Luo等人[56]提出了一个数据驱动的投资决策框架，为了提高P2P信贷中的投资表现，这个决策框架挖掘了每一项标的的投资者构成来指导投资决策。Katherine和Herrero-Lopez[57]通过分析大规模的P2P信贷网络中投资者的行为，发现当存在一些具有很高风险的激进型的投资者时，他们的很多投标决策都是次优决策。Singh等人[58]研究了在Prosper上的投资风险和回报，他们发现对于不同的信用等级的贷款，存在子群组能给出正回报。对于这些子群组而言，他们的收益也是伴随着风险的。此外，那些低信誉等级的贷款，在风险和收益的正相关性上比那些高信誉等级的贷款表现得更明显。Klafft[59]解释了按照一些简单的投资规则，可以提高投资组合的收益率，并且除了哪些很高风险的贷款外，将这些简单投资规则应用于各种等级贷款信誉都可以取得可观的回报。Garman等人[60]提出了一个叫搜索溢价的概念，这一部分溢价不同于借贷者支付的那部分通过无代价搜索的贷款利率，而是借贷者为了(尽快)完成借贷支付最少的利率来减少投资者的搜索代价。Iyer[61]等人评估了在P2P信贷平台上的投资者是否能够利用借贷者的信息来推断他们的信誉。为了验证投资者是否有解读借贷者信息的能力，它们使用了一种利用借贷者类别的方法论，并且发现这些投资者能够利用这些已有的信息来推断贷款信用。

P2P信贷包含许多领域的内容，为我们提供了许多当前未被发掘的研究机会和挑战。本章的目标是基于二部投资网络模型，预测新的投资者的投资能力和预测出新的标的项目是否可以竞标。本章的工作主要是利用平台上老的投资者的历史投资经历并且提供有效的算法对这些老的投资者的投资能力进行评估。利用P2P信贷平台上这些老的投资者的投资行为数据可以帮助我们发现并分析标的的好坏和投资的风险程度。

本章的组织结构如下。第3部分依据P2P信贷平台投资者和贷款的关系，我们首先描述了*M-to-N*关系模型称作二部图；接着，我们定量地分析了新的投资者和未知状态贷款的综合表现。第4部分我们给出了一个完整的决策模型来帮助投资者选取值得信赖的贷款项目进行投资。第5部分我们使用来自Prosper平台的真实数据验证了我们提出的投资决策模型的有效性。通过和使用BayesNet，Logistic和Average方法的实验结果进行比较，说明了我们的计算模型可以作出相对更好的投资决策。此外，我们的实验结果显示Logistic模型和我们提出的迭代计算模型相互之间具有一定的互补特性。在实验结果中，我们发现相较于使用单个模型预测，使用混合模型(Logistic模型和我们提出的模型进行集成)取得了相对更有效且稳定预测效果。最后，在第6部分我们给出了本章总结。

4.3 投资者和贷款的综合评估

4.3.1 二部投资网络

像Prosper这样的P2P信贷平台有点类似于股票市场，在这个网贷平台上，投资者可以将他或她的资金投放到多个标的中，并且相对一个标的也很可能接收来自多个投资者的竞标。这种多对多的异构关系可以被模型为二部图，这种图模型被广泛应用于模型两类实体间的关系。因此，我们将使用二部图来解释投资者和贷款之间的关系。

在P2P信贷所构成的二部关系中，投资者和借贷者是属于完全不同的类型，并且他们之间相连的边是带权值的，权值为投资者的投标金额。特别地，在Prosper的在线平台上，如果一名用户想要通过平台贷款，他或她必须申请发布一个标的。当他或她的申请被平台受理且通过后，就可以被发布到平台上供所有投资者竞标。投资者浏览正在进行的标的列表即可选择是否要为某个标的竞标，直到标的满标为止。

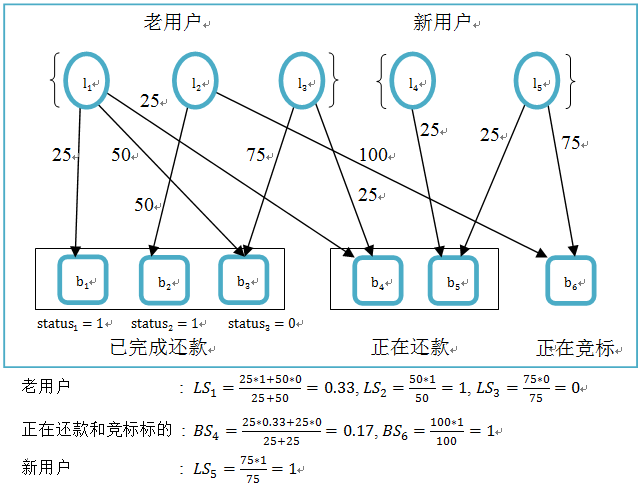
我们使用代表*N*个投资者集合，并且代表*M*个贷款项。表示*M*个贷款项的状态，其中表示完成还款和未完成还款。表示所有竞标记录集，表示投资者在贷款项上的竞标金额。最后，L，B和E构成的二部投资网络，如图1.2所示。

4.3.2 初始评估模型

基于二部投资网络图，我们提出了一个对投资者和贷款项进行综合评估的方法。首先，为了判断一名投资者是否可信赖的，我们需要去评估他或她的投资表现，因此他或她的投资金额和回报的构成将作为决定性评价指标。我们令表示投资者可信赖水平，则表示投资者的完成还款的投标项和他或她的竞标总和的比率。那么可写成等式(4.1)：

(4.1)

式中*m*表示完成还款的投标数目，*n*表示老有过投资记录的投资者总数。



**图4.1 初始评估状态**

当这些有投资历史记录的投资者的投资能力被初始评估后，我们可以反过来利用他们的信赖水平来分析当前正在竞标或正在还款的标的完成还款的概率。令表示贷款项的完成还款的概率，表示在第*i*项贷款上的竞标权重，则贷款项完成还款的概率可以定义成等式(4.2)的形式：

(4.2)

式中*n*表示在贷款项*j*上的投标数量，*m*表示正在还款和正在竞标的贷款项。

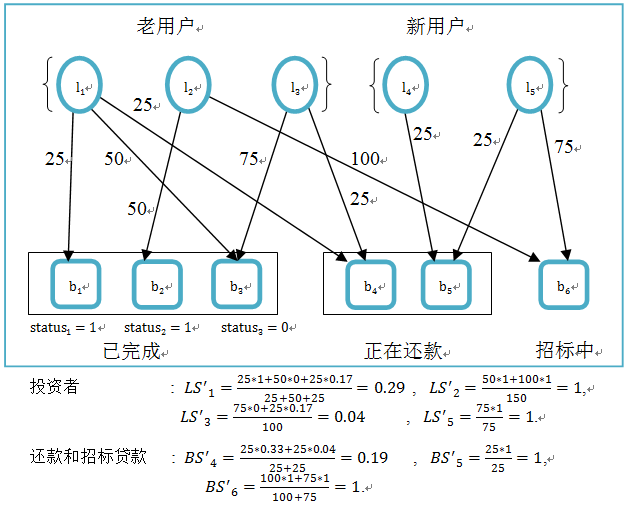
利用正在还款和正在竞标的标的的还款概率，我们又可以去评估新的没有投资历史记录的投资者的信赖水平。计算式定义如等式(4.3)所示：

(4.3)

式中*n*表示正在还款和正在竞标的标的总数，*m*表示新加入的投资者总数。一个简单的初始阶段评估状态如图4.1所示。

完成第一轮评估计算后，不难注意到它存在一个缺陷。如图4.1中的例子，我们注意到投资的信赖水平和贷款项的还款概率都没有被评估。然后，仔细分析初始计算结果状态和二部投资网络结构，我们可以发现如果再进行一轮迭代计算，可以被评估通过。类似地，既然已经知道了，则也可以在下一轮迭代中根据和贷款的关系而被评估。因此，我们需要基于初始计算结果，设计第二轮计算来作评估。

4.3.3 第二轮评估



**图4.2 第二轮评估**

一部分当前正在还款和正在竞标的标的经过第一轮计算后已经被评估，但还有完全未知，需要执行在下一轮计算得到。因此，我们需要再次评估老用户投资者的可信赖度，这时候的评估，我们需要利用上已完成还款、正在还款和正在招标但已得到评分的贷款项目。

(4.4)

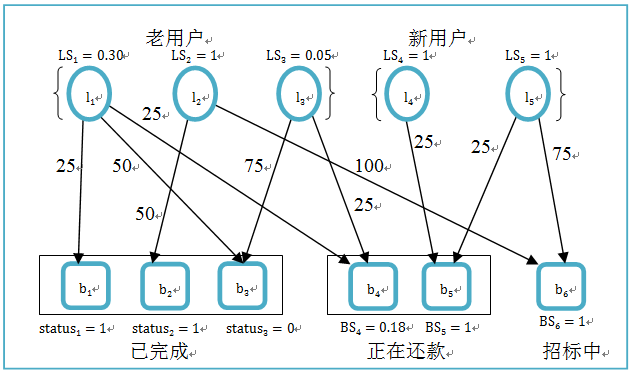
式中*m*表示已完成还款的贷款，*M*表示所有贷款总数，n表示所有投资者数目。因此在第二轮计算贷款项完成还款的概率的评估式定义为等式(4.5)：

(4.5)

式中n代表投资者数量，m表示当前正在还款和正在招标的贷款项。因此，使用等式(4.5)，我们可以计算得到。第二轮计算后的结果状态如图4.2所示。

4.3.4 多轮迭代至收敛

经过第二轮计算后，我们发现值还没有计算出来。但是这个问题并不麻烦，因为我们可以很简单地预测的值，只要我们利用式(4.5)的计算结果代入式(4.4)中即可得到。我们得到的计算结果是。从式(4.4)和式(4.5)的定义可知，它们的定义是递归的。因此，我们它们可以从给定的初始状态开始迭代计算至收敛状态。最后，我们在图4.4展示了经过多轮迭代计算后达到稳定状态的投资者信赖水平和贷款项完成还款的概率。



**图4.3最终的收敛状态**

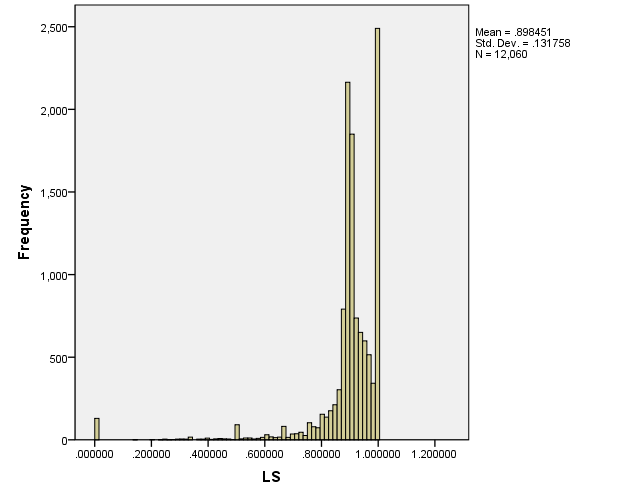
4.3.5 收敛特性

在图4.4中，我们知道所有投资者的信赖水平和贷款项完成还款的概率都到达了稳定状态。我们使用来自Prosper平台的真实数据，其中包含超过300,000个投标，超过4,000个标的，以及超过12,000个投资者用户。在我们对这些投资者和贷款最终的评估结果中，我们将投资者的信赖水平和贷款完成还款的概率的计算值保留小数点后3位，最终在可接受的25次迭代后达到收敛状态。图4.4展示了具有不同数量的贷款项的二部投资网络最终达到收敛状态的迭代次数。图4.4的结果说明，计算模型的拓展性很好，因为无论在多大规模的数据集中，它的复杂度因子都几乎是线性的。这一计算过程和PageRank[3]算法过程很相似，[62]给出了类似PageRank一类的计算模型的收敛性描述。不同的是，PageRank是一个同构网络算法，而我们的计算模型是在二部图模型的异构网络上。既然我们的计算模型最终是在对数时间复杂度达到收敛的，那也意味着二部图同样拥有一个很好的扩张因子。

**图4.4 不同贷款数的迭代计算次数**

4.4 实验分析

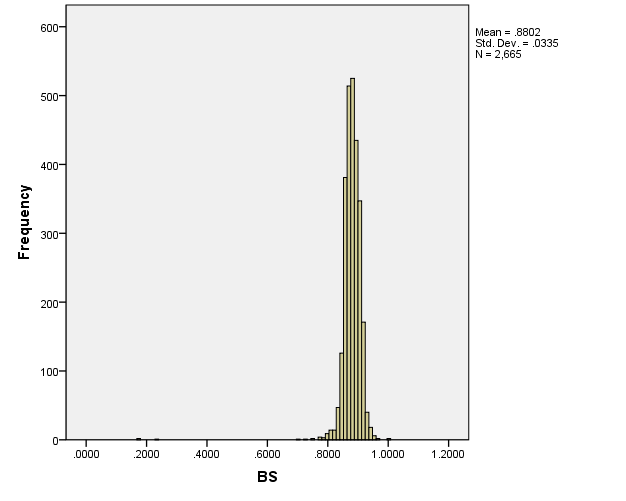
在这一部分，我们描述了我们通过使用P2P信贷平台的真实数据来验证我们的模型的有效性。在接下来的部分，我们将详细地阐述我们提出的模型预测效果。最终实验结果发现，我们的预测模型在一定程度上可以有效地预测贷款项的还款概率。

****

**图4.5 投资者信赖水平的统计分布**

4.4.1 结果分析

在这一部分，我们将通过分析模型的计算结果，来说明我们的预测模型在预测贷款能否完成还款上的有效性。图4.5展示了我们的预测模型在经过24次迭代计算后达到收敛状态时，投资者信赖的水平的统计分布。从图4.5可以看出具有高信赖水平(大于0.9)的投资者比较少，但是具有很低信赖水平(小于0.6)的投资者也很少，大多数人的信赖水平还是集中在0.8到0.9之间的。我们的计算模型认为，投资者信赖水平*LS*越高，则该投资者投资决策越可靠；反之，认为他或她的决策越错误。



**图4.6贷款完成还款的概率的统计分布**

图4.6展示了计算收敛后，贷款完成还款的概率的统计分布。从图4.6的统计分布情况可以看出，贷款完成还款的概率呈现了明显的均值为0.88、标准差为0.034的高斯分布。此外我们使用式(4.6)来计算我们预测结果在给定不同概率区间值下的正确率。具体结果如表4.1所示。

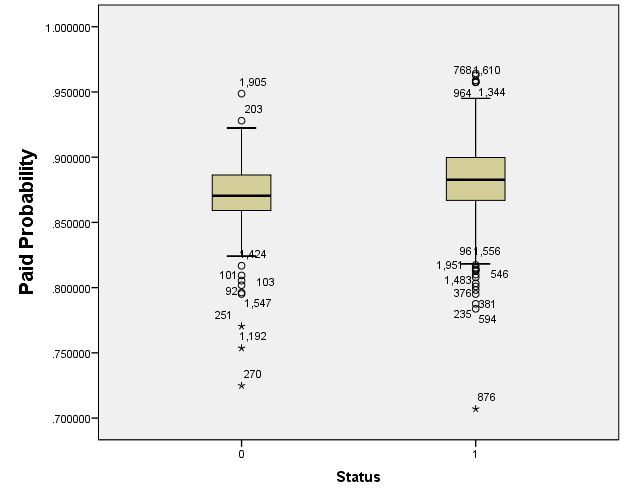
(4.6)

式中，表示实际完成还款的贷款数，表示实际未完成还款的贷款数。

**表4.1 不同概率区间值的预测正确率**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | <0.8 | 0.8to0.82 | 0.82to0.84 | 0.84to0.86 | 0.86to0.88 | 0.88to0.9 | 0.9to0.92 | 0.92to0.94 | >0.94 |
| Paid | 5 | 13 | 63 | 322 | 484 | 456 | 328 | 51 | 14 |
| Total | 10 | 16 | 75 | 389 | 557 | 508 | 344 | 52 | 15 |
| *AR* | 0.5 | 0.81 | 0.84 | 0.83 | 0.87 | 0.90 | 0.95 | 0.98 | 0.93 |

表4.1说明了随着选取的预测概率值增大，预测的正确性也随之提高，并且预测的会完成还款的贷款数和实际会完成还款的贷款数之间的差值也越来越小。



**图4.7 实际完成还款和未完成还款的预测概率箱图**

图4.7展示的，我们的预测模型对于实际完成还款和未完成还款的贷款的概率预测值的统计箱图，可以明显地看出实际会完成还款的贷款的中位值要高于实际未完成还款的。这也是一个很好的信号说明，我们的预测模型在一定程度上的有效性。

4.4.2 多种预测模型比较

在P2P信贷中，我们的预测模型在选取有价值贷款项时，是依据我们对它们预测完成还款的概率值从高到低排序后的。图4.8中展示了不同预测模型的预测结果，在各自预测的排名靠前的贷款项中选取前*α*比例的贷款项目，其中各个模型预测的贷款的还款率如图中各曲线所示。图中使用平均还款率作为参考基线，由图可以看出平均还款概率大约为88.3%。除了和平均还款概率比较，我们还和logistic regression和BayesNet模型进行了比较。从图中我们可以看出，相较于我们的计算模型，logistic regression和BayesNet模型的预测准确率随着*α*值得增大下降得较快。我们也可以明显地看出，在一定程度上我们的计算模型的表现要优于其它三类预测做法。虽然，我们的计算模型在α=0.1时比Logistic Regression模型有轻微下降，但是从整体而言，我们的计算模型的表现仍然是更优的；Logistic Regression模型和我们的计算模型的在部分上表现出来的优劣，也提示了我们使用混合的预测模型。

Top Rate α

Paid Rate

**图4.8不同模型预测结果的还款率比较**

4.4.3 和Logistic模型的混合

在图4.8中，我们看到BayesNet和我们的计算模型在α=0.1时的走势都是急速下降；而相反的是，Logistic表现出了上升的趋势。因此，我们猜想Logistic或许可以取得更好的表现，如果混合上我们的计算模型的预测结果。为了集成两个分类模型的的预测结果，我们使用式(4.7)。这样，贷款项完成还款的概率就由我们的预测模型和Logistic模型的结果构成。

(4.7)

式中表示对我们的预测模型的结果的置信权重。

Top Rate α

Paid Rate

**图4.9 混合模型和其他模型的预测结果对比**

在图4.9比较了混合模型的预测结果中还款项所占的比率。结果说明，相较于其它几类单独的预测模型，我们的预测模型和Logistic模型构成的混合预测结果取得了更有效且更稳定的表现。在我们的实验中，我们令θ=0.6，因为我们对比了多组实验结果最终得出该值的效果最佳。基于图4.9所示的实验结果，我们认为我们的预测模型和Logistic模型，至少在我们实验的数据集中，一定程度上相互之间是具有互补效应的。此外，我们在图4.10中给出了当时，混合模型和其它几个模型的预测结果对比。从结果图同样可以看出，相较于其它几个单独的预测模型，混合模型取得了更优的表现。α=0.1是表现最显著的点，在该点我们可以看出，模型混合后得到了很大的提升，并且表现了和Logistic相似的预测特性。可见，Logistic模型在提高我们的预测模型的准确性上表现了非常重要的促补充作用。因此，也更加说明，我们的预测模型和Logistic模型具有一定的互补特性；而相较于独立的模型，集成两类模型的预测结果也能取得相对更好的表现。

Top Rate α

Paid Rate

**图4.10 θ=0.5时的预测结果对比**

4.5 本章小结

在本章，我们提出了一个P2P信贷的投资预测模型。我们挖掘老的投资者用户的投资表现，评估他们的可信赖水平，并基于他们的信赖水平去预测或评估正在还款或正在招标的贷款完成还款的概率。具体地，我们使用老的投资者的历史投资记录去计算他们的信赖水平；然后依据他们的信赖水平去预测未结束的贷款项完成还款的概率，同时反过来利用这些估计的完成还款的概率值去评估新的投资者的信赖水平；接着，我们使用已完成的和已评估的贷款项去重新估所有投资者的信赖水平，随后又再使用重新估计得到的投资者信赖水平值去再次估计所有未结束的贷款完成还款的概率；最后，我们迭代上述计算过程直到所有的投资者和未结束的贷款项的概率水平达到稳定状态。使用来自P2P信贷平台上的真实数据实验，结果说明我们的预测模型在一定程度上可以帮助投资者过滤出好的投资项目进行投资。多次实验结果也表明，使用我们的预测模型和Logistic模型集成的混合结果，可以取得更有效且稳定的表现。实验结果也在一定程度上说明，我们提出的预测模型和Logistic模型在具有良好的互补特性。因此，相对于使用单一模型的预测结果，使用混合模型可以取得更佳的预测表现。

第五章 总结和展望

在本文，我们提出了使用异构信息网络来模型现实问题中的数据，并且介绍了一些方法来处理基于异构网络的链接预测问题。虽然同构网络已得到广泛的应用和研究，但是在现实世界中各个对象通常是通过复杂的网络方式将各种异构类型连接起来的，这种异构形式的网络覆盖了丰富的语义和关键性信息，例如前文介绍的致病基因网络链接预测问题。显然地，异构信息网络可以蕴含数据中更丰富的语义信息，基于异构信息网络而进行挖掘的算法通常可以更深入地发掘数据中不同实体类型之间的关系，实现更好的预测结果。

5.1 总结

在本文中，我们针对异构网络链接预测做了如下几项工作：

(1)致病基因网络中的异构网络链接预测问题。伴随着生物实验技术的开展，大量生物信息数据被收集得到。因此，针对这些已知的生物数据，如何借用数据挖掘的技术来探索人类疾病和基因之间的关系，正成为生物信息学的新的研究任务。本文在第二章中，就关于致病基因网络的异构链接预测问题，提出了针对链接预测问题而设计的使用异构相似约束的隐因子关系预测模型。具体地，我们首先将预测疾病和基因之间关系的问题转化为半监督学习问题。接着，根据隐因子模型的关于对象之间在相同特征空间里，距离越近则越相似的算法思想，我们将疾病和基因之间的关系预测转化为二分类概率模型：疾病和基因如果在隐含特征空间的距离越近它们存在关系的概率越高；反之，若距离越远，则存在关系的概率越低。最后，为了充分利用异构信息网络所覆盖的丰富语义而更准确的挖掘基因和疾病之间可能存在的相关性，我们在基础模型的基础上加入了异构相似约束，即平均异构相似约束隐因子模型和个性化异构相似约束隐因子模型。

(2)社交投资网络中的异构网络链接预测问题。随着互联网技术的发展和普及，社交网络、众筹等以互联网为基础的社会活动正在迅速发展。在高速向前的技术和创业热潮下，无论是众筹、风投还是企业融资，在当前信息化背景下，充分利用已有的相关投资记录，对于公司而言可以更准确地定位投资者，而相反对于投资者而言是可以更有效地寻找出有价值投资项目。在本文第三章中，就关于社交投资网络中的异构链接预测问题，我们设计了一个数据驱动的社交投资网络链接预测框架，该框架将投资行为模型为基于异构信息网络下使用meta-path的链接预测问题。特别地，基于实体类型和关系类型，我们首先建立了一个异构关系网络模式；然后在这个异构关系网络上，我们提取了所有的有意义的路径长度不大于4的meta-path，并提供了有效的度量函数来量化meta-path相似度。最后，我们提出了一个有监督的社交投资网络链接预测学习模型，来预测投资行为。

(3)P2P信贷网络中的异构网络链接预测问题。P2P信贷作为一种以互联网为基础的新兴金融市场，近年来在国内外迅猛发展。在P2P信贷这个在线平台中，投资者们最希望的是能够投资哪些能保证收益的贷款，而作为中间平台而言，他们也希望过滤掉那些差的借贷者们。因此，无论是对投资者还是第三方平台，为他们预测有价值的贷款项目正是他们所需。在本文第四章，我们提出了一个P2P信贷的投资预测模型。我们挖掘老的投资者用户的投资表现，评估他们的可信赖水平，并基于他们的信赖水平去预测或评估正在还款或正在招标的贷款完成还款的概率。具体地，我们使用老的投资者的历史投资记录去计算他们的信赖水平；然后依据他们的信赖水平去预测未结束的贷款项完成还款的概率，同时反过来利用这些估计的完成还款的概率值去评估新的投资者的信赖水平；接着，我们使用已完成的和已评估的贷款项去重新估所有投资者的信赖水平，随后又再使用重新估计得到的投资者信赖水平值去再次估计所有未结束的贷款完成还款的概率；最后，我们迭代上述计算过程直到所有的投资者和未结束的贷款项的概率水平达到稳定状态，从而为投资者预测出可投资的贷款项。

5.2 展望

本文对异构网络链接预测进行相关的研究工作，具体的针对不同应用问题提出几种算法，并在实际提供的数据上进行了算法实验，实验结果说明了文中提出的算法在一定程度上的有效性。此外，本文仍然还有需要改进和进行深入的研究地方：

(1)在关于致病基因网络的链接预测问题中，我们只使用了通过同构网络计算得到的基因之间和疾病之间的异构相似约束，可以考虑使用通过meta-path计算得到的相似值做约束。

(2)在本文中的半监督学习中，对于负样本的选取策略，在很大程度上会影响后续算法学习的准确性，所以探索计算效率更高且选取更精确的策略极其必要。

(3)加入更多的相关数据，提高预测准确性。异构网络相较于同构网络的另一个优势是，可以非常便利地加入更多其它相关的对象类型和关系类型。例如对于本文第三章的问题，如果我们可以获得投资者相关的数据，并加入到网络模式中，对提高算法的预测准确性势必有益。

(4)在关于P2P信贷网络的链接预测问题中，投资者和借贷者除了有借贷关系外，还存在社交关系；同样的，投资者之间也存在好友关系、社区或组关系等社交关系。因此，考虑如何深入地融合以上社交关系到信贷网络中，也是我们需要进一步研究的课题。

(5)在实验过程中，算法在单机环境下运行很耗时，尤其在实际应用环境中，随着数据量的增长，将制约算法的应用。因此，针对现行环境，将算法并行化或基于hadoop/spark等流行的大数据计算框架的分布式算法，也是我们需要深入的课题。

参考文献

1. S. Roy, T. Lane, and M. Werner-Washburne. Integrative construction and analysis of condition specific biological networks. In Proc. 2007 AAAI Conf. on Artificial Intelligence (AAAI＊07), Vancouver, BC, July 2007.
2. Berger S C, Gleisner F. Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending[J]. Business Research, 2009, 2(1):39-65.
3. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 30(1-7):107-117, 1998.
4. Radicchi F. A paradox in community detection[J]. Epl, 2013, 106(3):228-232.
5. C. C. Aggarwal, editor. Social Network Data Analytics. Springer, 2011.
6. Sun Y, Han J, Yan X, et al. Mining knowledge from interconnected data: a heterogeneous information network analysis approach[J]. Proceedings of the Vldb Endowment, 2012, 5(12):2022-2023.
7. Meta-Path-Based Search and Mining in Heterogeneous Information Networks. TSINGHUA SCIENCE TECHNOLOGY. 4 (2013) 329-338.
8. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research 21: 1109-21.
9. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome research 13: 2363-71. Karni S, Soreq H, Sharan R (2009) A network-based method for predicting disease-causing genes. Journal of Computational Biology 16: 181-189.
10. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology 314: 1041-52.
11. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, et al. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proceedings of the National Academy of Sciences of the United States of America 107: 6544-9.
12. Online Mendelian Inheritance in Man, OMIM (2011) Available: http://omim. org/. Accessed: 2011 Aug.
13. Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation[J]. PLoS computational biology, 2010, 6: e1000641.
14. McGary KL, Park TJ, Woods JO, et al. Systematic discovery of nonobvious human disease models through orthologous phenotypes[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107: 6544-9.
15. Goh K, Cusick M, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proceedings of the National Academy of Sciences 104: 8685.
16. Tian W, Zhang LV, Taan M, Gibbons FD, King OD, et al. (2008) Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biology 9 Suppl 1: S7.
17. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and highthroughput data. BMC systems biology 1: 8.
18. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4: 189.
19. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics: 949-958.
20. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS computa- tional biology 6: e1000641.
21. Li Y, Patra JC (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics/computer Applications in The Biosciences 26: 1219-1224.
22. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555-1558.
23. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. Genome Research 19: 1093-1106.
24. McGary KL, Lee I, Marcotte EM (2007) Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. Genome biology 8: R258.
25. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research 21: 1109-21.
26. Mordelet F, Vert JP (2011) Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics 12.
27. U Martin S B, Nagarajan N, Ambuj T, et al. Prediction and validation of gene-disease associations using methods inspired by social network analyses.[J]. British Journal of Dermatology, 2013, 8(5):66-68.
28. Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008. 2008:426-434.
29. Liu B, Dai Y, Li X, Lee WS, Yu PS (2003) Building text classifiers using positive and unlabeled examples. In: Intl. Conf. on Data Mining. pp. 179-188.
30. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proc. of UAI ＊98, 1998.
31. P. Giot, U. Hege, A. Schwienbaher. Expertise or Reputation? The Investment Behavior of Novice and Experienced Private Equity Funds. Available at (www.fbfcorporatefinance.fr/medias/cahiers\_de\_recherche/VCFLOWS\_2011-09-19.pdf)
32. M. Grinblatt and M. Keloharju. The investment behavior and performance of various investor types: a study of Finland's unique data set. Journal of Financial Economics. 55 (2000) 43-67.
33. J. S. Doran, D. R. Peterson, C. Wright. Confidence opinions of market efficiency and investment behavior of finance professors. Journal of Financial Markets. 13.1 (2010) 174-195.
34. A. Barnea, H. Cronqvist and S. Siegel. Nature or Nurture: what determines investment behavior? Journal of Financial Economics. 98 (2010) 583-604.
35. Kleinberg, Liben Nowell Jon. "The Link-Prediction Problem For Social Networks." J American Society for Information Science & Technology. 58.7 (2003) 1019-1031.
36. Y. E. Liang, S. D. Yuan. Investors Are Social Animals: Predicting Investor Behavior using Social Network Features via Supervised Learning Approach. Proceedings of the Workshop on Mining and Learning with Graphs(MLG-2013), 2013, pp. 305-312.
37. Y. E. Liang, D. Y. Sor-Tsyr. Where's the Money? The Social Behavior of Investors in Facebook's Small World. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 158-162.
38. G. Xiang, Z. Zheng, M. Wen, J. Hong, C. Rose, and C. Liu. A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and New Articles on TechCrunch. ICWSM'12, 2012.
39. Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, J. Han. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks. 2011 International Conference on Advances in Social Networks Analysis and Mining, 2011, pp. 121-128.
40. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in Heterogeneous Information Networks. In Proc. 2011 Int. Conf. on Very Large Data Bases (VLDB'11), Seattle, WA, August 2011, pp. 992-1003.
41. R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining(KDD'10), Washington D.C, July 2010, pp. 243-252.
42. C. Shi, X. Kong, P. S. Yu, S. Xie, B. Wu. Relevance Search in Heterogeneous Networks. In EDBT 2012, 2012, pp. 180-191.
43. Y. Sun, J. Han. Meta-Path-Based Search and Mining in Heterogeneous Information Networks. TSINGHUA SCIENCE TECHNOLOGY. 4 (2013) 329-338.
44. A. W. Marshall, I. Olkin. A multivariate exponential distribution[J]. Journal of the American Statistical Association. 62.317 (1967) 30-44.
45. A. C. Cohen. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. Technometrics. 7.4 (2012) 579-588.
46. Berger SC and Gleisner F. Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. BuR Business Research Journal 2009; 2; 39-65.
47. Freedman S and Jin GZ. Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.Com. NET Institute Working Paper 2012, pp. 8-43.
48. Lin M, Prabhala NR and Viswanathan S. Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. Management Science 2013; 59; 17-35.
49. Collier BC and Hampshire R. Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets. Proceedings of the 2010 ACM conference on Computer supported cooperative work, 2010, pp. 197-206.
50. Herrero-Lopez S. Social interactions in P2P lending. Proceeding SNA-KDD '09 Proceedings of the 3rd Workshop on Social Network Mining and Analysis, Article No. 3, 2009.
51. Chen X, Yang L, Wang P and Yue W. An Effective Interval-Valued Intuitionistic Fuzzy Entropy to Evaluate Entrepreneurship Orientation of Online P2P Lending Platforms. Advances in Mathematical Physics. Article ID 467215, 2013.
52. Wu J and Xu Y. A Decision Support System for Borrower＊s Loan in P2P Lending. Journal of Computers 2011; 6; 1183-1190.
53. Li S and Qiu J. Do Borrowers Make Homogeneous Decisions in Online P2P Lending Market? An Empirical Study of PPDai in China. Service Systems and Service Management (ICSSSM) 2011, pp. 25-27.
54. Herzenstein M, Andres R, Dholakia U and Lyandres E. The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-To-Peer Lending Communities. Working Paper, University of Delaware, 2008.
55. Pope DG, Sydnor JR. What＊s in a picture? Evidence of discrimination from Prosper.com. The Journal of Human Resources 2011; 1; 53-92.
56. Luo C, et al. Enhancing investment decisions in P2P lending: an investor composition perspective. KDD＊11(Knowledge Discovery in Databases), New York, 2011, pp. 292-300.
57. Krumme K and Herrero-Lopez S. Do Lenders Make Optimal Decisions in a Peer-to-Peer Network?. Web Intelligence and Intelligent Agent Technology. IEEE Computer Society 2009, pp. 124-127.
58. Singh H, Gopal R and Li X. Risk and Return of Investments in Online Peer-to-Peer Lending. In Proceedings of the 19th Workshop on Information Technologies and Systems (WITS 2009). Available at: http://misrc. umn. edu/wise/papers/p1-5.pdf, 2009.
59. Klafft M. Online peer-to-peer lending: A Lenders＊ Perspective. Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, IEEE 2008, pp. 371-375.
60. Garman S, Hampshire R and Krishnan R. A search theoretic model of person-to-person lending. Available at: http://heinz.cmu.edu/research/244full.pdf, 2012.
61. Iyer R, Khwaja AI, Shue K et al. Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? NBER Working Paper No. 15242, 2009.
62. Rajeev M and Prabhakar R. Randomized Algorithms. Cambridge University Press, 1995.

攻读硕士学位期间发表论文及科研情况

1. Investment Behavior Prediction in Heterogeneous Information Network, *Neurocomputing* , accepted.
2. A Decision Support Model for Investment on P2P Lending Platform, *Expert Systems with Applications*, under review.

致 谢

时间似白马，转身一望，三年就要过去了。或许是一种缘分，也或许是一种机遇，我来到了中国最美的大学，经历了人生美好的三年时光，体验了厦大师生的温情关怀和同学之间的深厚友谊。感谢厦大给了我这样一个平台，我收获了凤凰花情谊、良好的师生关系以及爱情。

回想起来，厦大海韵园是一个非常棒的科研地点，坐山傍海。每每工作劳累了，在窗前放眼一跳，即可望见广阔的大海，还能享受海风拂面的舒适感，身体瞬间就轻松下来了。更加难得的是，在当前国内各大城市环境污染泛滥成灾的环境下，厦门仍然空气质量良好。食堂晚饭后，天气好的时候，去海边散步溜达一圈，往往都能领略到大城市难得绚丽晚霞。

说完环境，那一定少不了，要说陪伴我度过硕士三年时光的人们。首先当属曾老师啦。真的要感谢有缘和曾老师成为师生。对曾老师的学生来说，既可是学生又可是兄弟，因为他不仅可以像老师一样传授你很多学术相关的计算机知识，也能和学生像哥们一样，玩在一起，谈论生活。然后，就是计算智能组内的所有成员，哪些毕业的师兄和还在一起的可爱的师弟师妹们。很幸运、也很幸福能和师兄、师姐、师弟和师妹们一块儿玩耍，哪些嬉笑恶搞的记忆和图片，每每再看起都让我欣喜不已。

最后，我真的需要太感谢我的父母了，这么多年了，我的母亲抚养我长大，父亲供我上学。因为父母的勤劳和关爱，让我的生活无忧。回忆过往，内心不得不为父母为家庭和我们所做的努力和牺牲所感动。他们与我，比所有伟人都伟大，谢谢他们为我做了好榜样，我努力做得比他们更好。

感谢我的朋友们，不管发生什么事，我都知道有你们在背后支持我。衷心感谢所有帮助和爱护我的人！