附件 1
总编号 _____

北京师范大学
“本科生科研训练与创新创业”
项目结项报告

项 目 名 称：基于时间序列数据挖掘的量化投资研究

申 请 人：李泽邦

所 在（院） 系：数学科学学院

申 请 人 电 话：18813111766

申请人电子信箱：zebangmail@qq.com

导 师：于福生

导 师 职 称：教授

导师所在单位：数学科学学院

导 师 电 话：13718288960

导师电子信箱：yufusheng@bnu.edu.cn

填 表 日 期：2017 年 5 月 14 日

北京师范大学教务处制表

填 表 说 明

1、 项目负责人按要求如实填写《北京师范大学“本科生科研训练与创新创业”项目结项报告》（简称《结项报告》），并提供必要的附件材料，作为项目验收和评估的主要依据。

2、 填写《结项报告》要求保证内容真实，数据准确。所有栏目必填，不得有空缺，所填栏目不够用时可加附页。

3、 封面总编号由“科研训练领导小组”统一编写。

4、 《结项报告》用 A4 纸打印，一式一份，于左侧装订成册。各单位可以自行翻印，但格式、内容、大小应与原件相同。

一、总结报告（要求字数在 3000－5000 字）

项目综述：

近年来，随着数据不断丰富，人们对强有力的数据分析工具的需求增加,，数据挖掘开始得到广泛的应用。例如对海量数据进行分析处理，挖掘其中蕴涵的各种信息，对于揭示事物发展的规律，发现不同的事物发展之间的相互关系等具有重要的实际意义。其中，时间序列数据会随着时间的推移规模不断扩大。因此针对时间序列数据的数据挖掘研究一直以来受到了学术界和工业界的广泛重视，成为了一个具有重要理论和实际价值的热点研究课题。

特别的，随着金融数据量的与日俱增，常用的一些统计方法已不能满足需求，“数据挖掘”(从大量数据中以非平凡的方法发现有用的知识)成为一种自然的需求。运用数据挖掘方法来分析金融时间序列成为研究的热点之一。作为时间序列挖掘的一个重要研究方向，股指时间序列挖掘，能够为人们正确认识股指和科学决策提供依据。

目前在股指时间序列挖掘这一领域已有一定研究，但由于时序数据本身实际背景的差异，目前还不能有一个统一的挖掘方法。如何针对特定实际意义下的时序数据，探寻能够满足实际需求的挖掘方法，仍是各领域时序挖掘的重点。特别是符合中国股市市场特性的数据挖掘刚刚起步，就此我们尝试从数据挖掘的理论出发，在我组“本基”已有的研究经验基础上，探索适合于中国股指市场特性的，股指预测研究，尝试一些新的方向，为中国市场股指预测提供一些可行的方法。

股指数据的量化投资问题，实质是一个数据挖掘问题，进一步，是一个关联规则挖掘问题。

我们针对股票数据，结合模糊粒化的思想，主要建立了以下两个方面的工作：

1. 构建了一种专门针对股指数据的模糊粒化方法，研究了相关的距离度量，并进行了数据实验，验证了其有效性。
   该种模糊粒化方法不仅仅克服了传统模糊粒化方法不能反映时序的缺点，同时，该种方法克服了对线性分割窗口大小的依赖性问题。并且，该种方法能够代表多种数据类型，特别是在反映价位区间高低上有显著成效。但是其缺点在于，对非常规则的数据类型失效，只能应用于噪声较大的数据类型，例如，股指数据。

2. 建立了时序数据上模糊关联规则的定义，不仅克服了传统方法不能应用与时序数据的问题，还把算法复杂度相较于传统挖掘大大降低。
   特别的，我们的新方法能够识别间断关联规则挖掘，并且是在极小的算法复杂度下实现，这是传统方法远不能实现的。我们经过数据实验验证了该种方法的有效性。这种方法的缺点在于，仍然极大依赖于粒化方法，对窗口划分有严格要求。但是这已经超出了该部分的课题内容。

总体来说，我们的数据挖掘流程分为以下几个步骤：模糊粒化，模糊聚类，关联规则挖掘，基于关联规则的预测。根据这个流程，我们对每一部分都进行了理论创新与技术攻关，不仅建立了自己的新型方法，同时还在程序上实现了我们的算法，最终，编写了 windows 可执行程序，相关理论结果欲整理成论文投稿.

项目研究计划要点、执行情况、主要进展：

计划要点：

1) 完成近年若干股指数据的搜集；
2) 建立时序数据模糊粒化方法
(包括隶属度函数的设计，模糊粒之间的距离度量，聚类方法，度量的有效性等工作)；
3) 建立基于模糊粒化的时序频繁项挖掘方法
(包括时序下频繁项的定义，如何进行连续频繁项，特别是不连续频繁项的挖掘方法，如何降低算法复杂度，以及如何在模糊关联规则的意义下进行时序规则的挖掘)；
4) 建立模糊时序关联规则的支持度定义和挖掘方法；
5) 生成强时序关联规则；

执行情况（主要进展）：

对应以上计划要点，执行情况与主要进展如下。

1) 我们经过查阅信息，发现了“通达信金融终端”软件能够完整的提供各个股指数据。同时我们发现了“DataMarket”网站能够提供各类真实的时序数据。这些工作为之后的研究奠定了数据及来源基础。

2) 经过观察，思考，摸索，我们构建了一种针对股指数据的模糊粒化方法。这种模糊化方式克服了传统模糊粒化方法不能反映时序的特点，同时能够反映价格区间的不同，对于金融数据挖掘有良好的适应性。

3) 针对步骤二的构建，进一步，根据 Hausdorff 模糊粒子距离的定义，我们推演了我们所构建的模糊粒的距离度量公式。并且通过数据实验说明了该方法的有效性。

4) 我们将传统的模糊频繁项集支持度定义，推广到时序序列。建立了时序数据中频繁项集支持度的定义。我们不仅考虑了连续型频繁项集，同时，考虑了间断的频繁项集。并且我们的定义计算复杂度非常低。

5) 类似间断频繁项支持度的定义，我们定义了模糊关联规则的支持度，设计了在我们的定义下的关联规则挖掘方法。

6) 我们根据所得频繁模式，确定预测时间跨度，继而计算频繁模式之后相应时间跨度内，关联规则的支持度与置信度。最后根据支持度与置信度阈值生成了强时序关联规则。通过实际数据和人造数据分别检验了我们方法的可靠性。

项目特色、创新点及研究取得的成果：

1) 项目特色：

原创性工作非常多，需要攻坚的环节非常多，使我们项目的特色。包括模糊粒化方法的构建，将模糊关联规则推广到时序数据上进行定义，以及编写相关代码及其程序，均为原创性工作。其次，我们的课题区别于其它课题，我们的课题包含了整套数据挖掘流程，各个步骤都需要原创性方法，任何一步都可以独立成为一个专门课题，任何一个步骤都缺乏合适的现成的理论可以套用，想要得到满意结果，必须建立起原创有效的方法。从数据搜集到程序编写，从理论建立到实际检验，我们的课题都一一涵盖，工作量很大，挑战非常多。相关理论成果已经整理成论文投稿至 ICNC-FSKD 2017。

2) 创新点：

构建了新的针对股指数据的模糊粒化方法；

推广了传统模糊关联规则，定义了时序模糊关联规则的支持度；

克服了一般方法不能挖掘间断规则的缺点，我们的定义及其算法能够挖掘间断规则；

3) 取得的成果：

理论上，构建了模糊粒化方法，定义了时序模糊关联规则的支持度；

应用上，编写程序实现了我们的理论及其算法，并且有可执行的应用程序；

结果上，关于时序模糊关联规则的推广定义，我们欲整理成学术论文投稿。

成果形式（名称）及合作交流情况：

1. Windows 可执行的应用程序；

应用前景评价及必要的说明：

近几年大数据及其挖掘的关注度持续上升，随着金融交易数据观测的愈加精确和完整，数据挖掘理论的不断发展，这两者的结合或许成为一种趋势。基于数据挖掘及关联规则的金融预测将受到越来越多的关注与应用。另外类似基于数据挖掘预测的一步式预测软件和平台，在数据预测的应用领域，必将愈发的具有市场与应用前景。

我们有理由相信，基于数据挖掘的预测理论及其技术，以及基于这些理论的应用平台，软件，公司，必将愈加受到关注，相关研究与市场的前景还远未真正展现。相关领域的前景非常可观，基于数据挖掘的预测技术，将得到充分发展，并渗透到各行各业。在未来这些技术上的改变，将深刻影响人们的决策方式和形式。

同时也注意到，大数据领域的技术尚未完全成熟，大数据的革命才刚刚开始，因此大数据在未来也将带给大家更多的机遇与挑战。

我们的研究是尝试性的，初步的，其中细节，流程的优化，还有诸多工作需要深入完善。但我们这种尝试与初步工作是具有意义的。如果有更进一步的技术支持，能够实现我们的理论，应当可以编写相关金融预测软件。

项目负责人签字：_____ 年 月 日

二、**经费使用情况：**

项目经费支出情况（单位：元）		
经费来源	资助经费	支出经费
教育部经费（国家级）		
市级经费（市级）	10000	
学校经费		
院系部经费		
导师课题经费		
其他经费		

经费支出详细列表（按照项目研究过程中经费支出实际情况，填写项目经费支出明细，对于虚报经费支出或浪费经费的项目，学校将严厉追究项目组责任）

支出经费	支出金额（元）	备注
1、测试费、分析费		
2、材料费		
3、差旅费		
5、出版/文献/信息传播/知识产权事务费		
合计		

请填写如下具体经费支出：

一、材料费：

主要包括在项目实施过程中，项目开发、试验所需的原材料、辅助材料、低值易耗品、零配件的购置费用以及为此发生的运杂包装费用。注意：材料费必须用于专业研究所需材料支出，如实验药品等。不得包含日常办公消耗材料，如硒鼓、打印纸、U 盘等

序号	名称	数量	单价（元）	总价（元）	支出理由
1					
2					
3					
4					
5					
	合计：				

二、 测试化验加工费/被试费：

测试化验加工费是指在项目研究过程中支付给外单位（包括项目承担单位内部独立经济核算单位）的检验、测试、化验及加工等费用；如项目招募被试，被试费包含在此项之中。

序号	内容	数量	标准（元）	总价（元）	测试时间
1					
2					
3					
4					
5					
	合计：				

三、差旅费：

差旅费是指在项目实施过程中，开展业务调研、学术交流等所发生的差旅费。差旅费的开支标准应当按照国家有关规定执行（学生外出不能乘坐飞机，可乘坐火车硬座）。

序号	事由	目的地	人数	天数	交通费	食宿费	金额（元）
1							
2							
3							
4							
5							
	合计：						

四、出版/文献/信息传播/知识产权事务费：

出版/文献/信息传播/知识产权事务费是指在项目实施过程中，需要支付的出版费、资料费（**图书和复印费**）、专用软件购买费、文献检索费、专利申请及其他知识产权事务等费用。

1、 图书费

序号	书名	数量	单价	总价
1				
2				
3				
4				
5				
6				
7				

2、 复印费

序号	复印材料内容		金额
1			
2			
3			
4			
	合计		

3、出版费：_____元（拟出版文章_____篇）

出版费一般国内 1000 元/篇，国外 3000 元/篇；专著出版费 3 万元/本。

4、文献检索费：_____元

根据相关文献检索付费标准进行合理预算。

5、专利申请费_____元

项目负责人签字：_____ 年 月 日

三、成员合作情况

成　员		学号	部院系专业	实际的任务分工	承担的任务占总任务的百分比	是否为项目开题后新加入人员	本人签字
主持人	李泽邦	201411132065	数科	论文等	50%	否	
参加人	卜凡	201411132101	数科	编程等	30%	否	
	李东升	201411132055	数科	文献等	20%	否	

四、项目组对项目完成情况的自我评价

（项目投入情况、是否达到预期目标、解决了哪些问题、还存在哪些问题、有何建议和意见等）

项目投入情况而言，我组经过广泛搜集资料，数据，查阅文献，从初涉数据挖掘领域，到完成 matlab 程序的编写，中间经历了许多波折，我组投入大量精力克服种种苦难，边学边做，边做边摸索，中间还克服了初始项目申请人变更的问题，这一过程中，我们付出了很多努力，从大三上接手项目开始，一点点摸索，从零做起，最终给出成果，这一点应当值得肯定。

是否达到预期而言，我组最终给出了预测方法及其程序，达到了项目预期，

解决的问题而言，我组在研究过程中，解决了数据搜集，模糊粒化，聚类，确定模式与规则发现的阈值等理论性问题（具体问题在学术报告中有详细提及），也解决了 Matlab 编程实现中遇到的种种技术性问题。一点点摸索，从零做起，经过克服这些困难，我们最终给出了预测程序的编写。

还存在的问题来说，总体而言，主要是窗口划分的问题，我们只是采取等长的窗口划分，实际上窗口划分是一个非常关键且复杂的问题，这是主要的问题。另外一点是有预测的问题，由于时间关系，我们开发的算法对于预测没有深入研究。这些问题的解决理论上讲，不是苦难的事情。但是由于时间，人力，客观条件的种种限制，进一步的工作尚未进行。希望我们未来能够完善这些工作。

进一步的建议来说，主要是我们在研究过程中，体会到传统方法很多时候不能适应频繁波动的数据集，应当考虑其它数据压缩方法。另外，仅仅基于频繁模式给出关联规则有一定局限性，应当考虑周期模式等。由于是一整套流程，要完成这些工作，必须在相当充裕的时间及精力的条件下。因此，未来有机会，这些方面的完善都是值得考虑的。

总体而言，但我组从初涉数据挖掘领域，到程序的编译完成，克服了重重困难，包括理论性的算法问题，也包括技术性的编程问题，甚至包括投稿的种种问题，但这些问题都没能妨碍我们给出最终成果，从大三上接手项目开始，一点点摸索，投入了相当精力，我们对自己的工作感到满意和自豪。

五、指导教师意见：

<table>
<tr><td colspan="4">指导教师对项目的评语：

项目完成情况：☐合格　　☐不合格

指导教师签字：____________　　　　年　　月　　日</td></tr>
<tr><td colspan="4">各成员成绩及评语</td></tr>
<tr><td></td><td>姓名</td><td>成绩</td><td>评　语</td></tr>
<tr><td>主持人</td><td>李泽邦</td><td></td><td></td></tr>
<tr><td rowspan="4">参加人</td><td>卜凡</td><td></td><td></td></tr>
<tr><td>李东升</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="4">备注：由指导教师对所有参加人员在项目中的表现按百分制分别给出成绩以及相应评语。</td></tr>
</table>

六、单位意见：

项目所在单位意见	部院系教学部长/院长/系主任 签字：_____盖章 年　　月　　日
学校评定意见	签字：_____　盖章 年　　月　　日