****

密级：\_\_\_\_\_\_\_\_\_\_

本科生毕业设计(论文)

机器学习在股票投资

题 目： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

风险评估中的应用

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

马腾

作 者： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

41255092

学 号： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

计算机与通信工程学院

学 院： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

计算机科学与技术

专 业： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

成 绩： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

年 月

**本科生毕业设计(论文)**

机器学习在股票投资

**题 目:**

风险评估中的应用

Machine Learning in stock

**英文题目：**

investment risk assessment

计算机与通信工程学院

**学 院:**

4

**班 级:**

马腾

**学 生:**

41255092

**学 号:**

副教授

潘永泉

**指导教师:**  **职称：**

**指导教师:**  **职称：**

声 明

本人郑重声明：所呈交的论文是本人在指导教师的指导下进行的研究工作及取得研究结果。论文在引用他人已经发表或撰写的研究成果时，已经作了明确的标识；除此之外，论文中不包括其他人已经发表或撰写的研究成果，均为独立完成。其它同志对本文所做的任何贡献均已在论文中做了明确的说明并表达了谢意。

学生签名：\_\_\_\_\_\_\_\_\_\_\_ 年 月 日

导师签名：\_\_\_\_\_\_\_\_\_\_\_ 年 月 日

这里的日期为正式论文提交的日期，应在规定的提交日期之前。

不用此信息时，删除此框。

**毕 业 设 计（论 文）任 务 书**  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

一、学生姓名：                 学号：   
二、题目:   
三、题目来源：真实 、    自拟   
四、结业方式：设计 、    论文   
五、主要内容：  
  
  
  
六、主要（技术）要求：  
  
  
七、日程安排：  
  
  
  
  
八、主要参考文献和书目：

本页制作时，可以将网上生成的任务书文档中的内容，全选后，复制到本节。也可分别复制相应内容。

本页的内容，应与本科教学网页上的任务书信息保持一致。

请将任务书页，尽量控制在2页以内。

提交正式论文时，本页需要有学生和老师的签字。

不用此信息时，删除此框。

这里的日期为任务书正式下达的日期，应在规定的任务书下达截止日期之前。

不用此信息时，删除此框。

指导教师签字：             年   月   日

学 生 签 字：             年   月   日

系（所）负责人章：             年   月   日

摘 要

自1989年中国试点股票市场以来，中国的股票市场发展迅猛，截止到2015年9月1日，总共有上海证券交易所和深圳证券交易所两家交易所的共2600余支股票。由于我国股票市场的不成熟和目前股票市场环境的多边形，我国的股票市场具有较高的风险性。本文从量化的角度，通过目前流行的机器学习的方法对于纽约证券交易所的部分股票进行技术层面分析，并对现有的风险种类进行分析和总结，利用相关的投资学方法结合机器学习得到的实际数据建立可靠的风险评估模型。该模型在限定使用范围的情况下，总结了股票投资风险相关的理论，对比了股票的实际的市场表现，为风险评估方向提供了一种定量的评估方法，并提出了部分解决风险的有效方案。其次分析了纽约证券交易所近5年内的数据，分析对比了不同风险计量指标的使用环境。本文的研究意义在于分析了股票市场风险产生的不同因素，并给股票投资者一种新的能衡量股票市场风险的视角。

关键词： 股票市场，投资风险，风险预测，风险控制

**提示信息：**论文中文摘要**字数约为300-600字**，如遇特殊需要字数可以略多。

论文摘要是论文内容不加注释和评论的简短陈述，一般以第三人称语气写成。摘要的编写应遵循下列原则：

1）摘要应具有独立性和自含性，即不阅读论文的全文，就能获得必要的信息。摘要是学位论文的缩影，是学位论文的主要内容、见解、结论简短明了的缩写。

2）摘要应是一片完整的短文，可以独立使用，可以引用。

3）摘要的内容应包含与论文等同量的主要信息，供读者确定有无必要阅读全文，也可供文摘汇编等二次文献采用。

4）摘要一般应说明研究工作的目的意义、研究方法、研究结果、主要结论及意义、创造性成果和新见解，而重点是结果和结论。

5）要用文字表达，不要附图和照片，除了实在无变通办法可用以外，摘要中不用图、表、化学结构式、非公知公用的符号和术语，不要使用表格、公式、上下标以及其他特殊符号，要突出重点，阐述清楚，少用数据表。

论文摘要用语力求简洁、准确。原则上300-600字。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

**提示信息：**中文关键词应为3-5个，用中文逗号“，”分隔。关键词是学位论文的检索标志，从学位论文中选取出来用以表示全文主题内容信息款目的单词或术语。关键词的选择关系到学位论文被检索和利用的效果，因此应全面、准确、规范。并应尽量采用《汉语主题词表》或各专业主题词表提供的规范词。不用此信息时，删除此框。

请不要移动关键词框的位置。该框距页面顶部25cm，距页左边4cm。

Machine Learning in stock investment risk assessment

Abstract

Since 1989 the Chinese pilot stock market, China's stock market is booming, as of September 1, 2015, Shanghai Stock Exchange and Shenzhen Stock Exchange which are two China Mainland‘s stock exchanges totally have more than 2600 kinds of stocks. Since the polygon of the stock market and the immaturity of the current stock market environment, China's stock market has a high risk. From the quantitative point of view, by popular machine learning methods for some New York Stock Exchange of stock technical level of analysis, and the existing risk categories were analyzed and summarized, use of relevant Investment method combines machine learning obtained actual data to establish a reliable risk assessment model. This model in the limited scope of the case, summed up the stock investment risks associated with the theory, compared the actual performance of the stock market, there is provided a method for the quantitative assessment of risk assessments direction, and proposed a partial solution to the risk of effective programs . Secondly, it analyzes the New York Stock Exchange nearly five years of data, analysis and comparison of the different measurement indicators of risk environments. Significance of this study is to analyze the different factors of risk of the stock market, stock investors and give a new energy to measure stock market risk perspective.

**提示信息：**英文摘要和关键词应当与中文在内容上基本相同，英文摘要的实词在300个左右。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

Key Words： Stock market， investment risk， risk forecastin，risk control

当Abstract的正文段落的某一行**出现较多空白**时，应按照参考文献中去掉空白的方法，将空白去掉。否则，太影响美观。

不用此信息时，删除此框。

**提示信息：**英文关键词应为3-5个，用英文逗号和一空格分隔，与中文关键词在内容上基本相同。不用此信息时，删除此框。

请不要移动关键词框的位置。该框距页上顶25cm，距页左边4cm。

目 录

在打印前或需要时，请在目录中点击鼠标右键，选择“**更新域**”中的“**更新整个目录**”，以显示最新目录。不用此信息时，删除此框。

[摘 要 I](#_Toc439577449)

[Abstract III](#_Toc439577450)

[插图或附表清单 VII](#_Toc439577451)

[注释说明清单 IX](#_Toc439577452)

[1 引 言 1](#_Toc439577453)

[2 2](#_Toc439577454)

[2.1 2](#_Toc439577455)

[2.2 2](#_Toc439577456)

[2.2.1 2](#_Toc439577457)

[2.2.2 2](#_Toc439577458)

[2.3 2](#_Toc439577459)

[3 3](#_Toc439577460)

[4 4](#_Toc439577461)

[5 结 论 5](#_Toc439577462)

[参考文献 7](#_Toc439577463)

[附录A 外文文献原文 9](#_Toc439577464)

[附录B 外文文献译文 11](#_Toc439577465)

[附录C 13](#_Toc439577466)

[在学取得成果 14](#_Toc439577467)

[致 谢 15](#_Toc439577468)

插图或附表清单

**提示信息：**插图或附表清单并非必要。论文中如图表较多，可以有此页。图的清单应有图号、图题名和页码。表的清单应有表号、表题名和页码。

根据所列内容，可将本页标题分别更改为“插图清单”、“附表清单”。

此页并非必要。**不用此页时，请删除此页。**

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

注释说明清单

|  |  |
| --- | --- |
| 符号 | 描述 |
|  | 不同影响因素的权重 |
| ***Volume*** | 某只股票的交易量 |
| ***Low*** | 某只股票交易日最低的成交价格 |
| ***High*** | 某只股票交易日最高的成交价格 |
| ***Open*** | 某只股票交易日开盘时的价格 |
| ***Close*** | 某只股票交易日闭盘时的价格 |
| ***Adj Close*** | 某只股票交易日平均成交价格 |
|  | P-J区间内的数据 |
|  | Tuples的集合 |
|  | 第i个聚类的中心 |
|  | 用于计算的特定常数d |
|  | 集合中值最大的数据 |
|  | 集合中值最小的数据 |
| ***CORR2*** | 相关系数的平方 |
| ***CORR*** | 相关系数 |
| ***R*** | 总收益量 |
| ***NPT*** | Average net price of undergraduates in a school |
| ***Rev*** | Total ***NPT*** of all the undergraduates in a school |
| ***I*** | 总投资量 |
| ***Conf*** | Confidence of ***I-R*** function |
| ***ΔR*** | 收益增长量 |
| ***ΔI*** | 投资增长量 |
|  | 系统性风险系数 |
|  | 系统性风险系数 |
|  | 总收益 |
|  | 边际收益 |
|  | 平均收益 |
|  | 数量 |
|  |  |

**提示信息：**此页并非必要。符号、标志、缩略词、首字母缩写、计量单位等的注释说明，如需汇集，可集中置于此页。

根据所列内容，将本页标题分别更改为“符号清单”、“标志清单”、“缩写清单”、“计量单位清单”等。

不用此信息时，删除此框。

此页并非必要，**不用此页时，请删除此页。**

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

1. 引 言

由于股票市场的高风险与其高收益之间存在一定的正比例关系，投资者愿意进行高风险的股票投资。投资者可以得出预期的收益率和风险，通过两者的内在联系进行分析，并且科学合理的设计投资方案。

在描述股票风险与收益率的模型中，比较著名的是CAPM/SLB模型，即有价投资组合理论衍生出的资本资产定价模型。该模型提出来的收益率和实际风险之间的关系，一种近似的线性模型，这种模型的因变量唯一由股票的β量决定。

股票市场风险理论的主要目的在于从理论上解释在既定的市场结构之下，投资者对股票买入卖出导致的实际损失。在这里我们对于模型进行了简化，只研究系统性风险对股票市场的影响。从定性分析而言，目前股市系统性风险的主要类型分为政策风险、利率风险、购买力风险和市场风险，作为之中最重要的组成部分利率风险和市场风险起到了决定性的影响作用。

对于股票市场理论而言，风险评估机制作为股票市场理论的重要组成部分，在追求市场稳定性的今天更是热点中的热点。另一方面，计算机领域中的机器学习方法更多的应用于金融工程的量化分析，众多国外研究者将机器学习的方法应用于金融工程领域。本文的研究基于某一时间段内股票交易市场对象，在样本模型构建上采用了静态分析法，并运用回归分析对风险的分类进行比较。最后，应用聚类分析对象之间的相似程度来无指导的分组到不同的簇，通过对比不同聚类之间的相似和不同之处进行分析。

本课题以股票市场中的应用作为背景，基于目前计算机科学中常见的机器学习算法进行设计，相关实现和对比测试。主要内容包括：搜集股票市场相关数据并进行过滤和预处理，定性分析风险因素的来源与相关性，设计定量分析的机器学习算法进行分析，分横向纵向对于数据集定量对比测试。

**提示信息：**引言简要说明研究工作的目的、范围、相关领域的前人工作和知识空白、理论基础和分析、研究设想、研究方法和实验设计、预期结果和意义等。应言简意赅，不要与摘要雷同，不要成为摘要的注释。一般教科书中有的知识，在引言中不要赘述。

本科生毕业论文需要反映出作者确已掌握了坚实的基础理论和一定深度的专业知识，具有开阔的科学视野，对研究方案作了充分论证。因此，有关历史回顾和前人工作的文献综合评论，以及理论分析等，可以在正文中单独成章，用足够的文字叙述。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

1. 课题背景及研究意义

股票市场理论一直以来都是现代金融理论研究的热点领域，而股票的风险评估机制作为股票市场理论的重要组成部分，在追求市场稳定性的今天更是热点中的热点。另一方面，计算机领域中的机器学习方法更多的应用于金融工程的量化分析，本文将利用机器学习的思想对股票市场的风险进行建模和分析。

* 1. 课题背景

股票市场风险理论的主要目的在于从理论上解释在既定的市场结构之下，投资者对股票买入卖出导致的实际损失。在这里我们对于模型进行了简化，只研究系统性风险对股票市场的影响。从定性分析而言，目前股市系统性风险的主要类型分为政策风险、利率风险、购买力风险和市场风险，作为之中最重要的组成部分利率风险和市场风险起到了决定性的影响作用。

从股票的不同类型的角度探究股票市场的风险性是今年来分类方法流行的必然趋势，它主要解释的问题是可以将股票定量问题转化为定性问题，再通过对于定性问题的分析和规范将问题的结果定量化。

目前该研究领域中都仅仅对于股票市场的风险进行定性分析，我们希望通过建立某种模型对不同中股票进行分析得出一个“股票风险指数”，从而能够辅助投资者对于股票风险的分析。根据先定性后定量的方式，首先需要通过对于股票市场中若干只样本股票进行分类，对于得到的不同类股票进行经济学方面的定性分析，再通过定性分析的结果转化为初级模型，通过初级模型进行自学习完善模型，最终得到一个完整的分类系统，而这种模型首先是要建立在数据分类的基础之上的，所以我们需要能进行自主学习并正确分类的数据分析方法。

对于计算机科学中的机器学习领域，许多算法都提供了分类的方式处理数据，类如支持向量机（SVM）、大规模聚类算法（CURE）、GRGPF算法，这里我们准备使用K-均值算法（K-means）的改进算法K-means++，该算法对于大规模数据有较好的分类效果，并且能够通过试探K的值确定最优的分类数目。

通过引入机器学习的算法解决了分类的问题，在定性分析方面，通过对于不同影响因子设定不同的权重，最终得出一个可以全面表现风险性的指数。如果需要设定影响因子的权重，只通过分析是不能准确设定的，可以通过模型自学习的方式进行在线动态设定，再通过蒙特卡洛方法进行置信度分析，通过多次反馈得出最优的模型。

* 1. 研究意义

由于股票高风险及其高收益之间的对应关系，投资者愿意进行高风险的股票投资，透过预计投资回报率我们可以得出科学的投资决策（投资组合）。然而另外一方面，如何规避其中的风险并且获得较高的投资回报需要投资者进行慎重的衡量，这就需要对于不同股票的投资风险进行分析评估。目前已存在风险评估系统大多建立在主观评价之上，本文设计了一种基于机器学习的股票风险评估系统。

通过该模型的评估结果，向投资者提供风险分析的渠道和辅助手段，对于中短期投资者具有较好的参考价值，我国的股票市场往往被称之为政策市场，这是由于我国股票市场投机性占主导因素和缺乏回避机制造成的，众多券商和融资机构借助内幕消息进行投资套利，加剧了我国股票市场的混乱程度。通过我们模型的聚类分析，我们可以很容易将具有高风险和人为操作痕迹的劣质股票挑选出来，对于投资者而言股票市场更加公开透明，对于我国股票市场的健康发展是极为有利的。

另外一方面，通过宏观数据的统计，政府可以对于不同聚类的股票进行宏观调控，使得产业结构向着有利的方向发展，从微观角度而言，由于非系统风险在投资总体风险之中的比例很小，在完全不考虑非系统性风险的情况之下，投资者仍然需要承受相当高的市场性风险。而上述模型可以理性的分析系统性风险，投资者可以利用风险分析的结果进行投资组合，分担相应的系统性风险。



1. 文献综述

对于本文所述的股票风险评估模型，我们需要使用到风险分析的基础方法，数据的预处理方法，资产组合理论，资本资产定价理论和K-means算法。

* 1. 股市风险的类型

股市风险可以划分为系统性风险和非系统性风险，两者在分析方法和影响方式上均大相径庭。

* + 1. 系统性风险

系统风险又称市场风险，也称不可分散风险。是指由于某种因素的影响和变化，导致股市上所有股票价格的下跌，从而给股票持有人带来损失的可能性。系统性风险主要是由治、经济及会环境等宏观因素造成，投资人无法通过多样化的投资组合来化解的风险。主要有以下几类：

1、政策风险

经济政策和管理措施可能会造成股票收益的损失，这在新兴股市中表现得尤为突出。如财税策的变化，可以影响到公司的利润，股市的交易策变化，也可以直接影响到股票的价格。此外还有一些看似无关的策，如房改策，也可能会影响到股票市场的资金供求关系。

2、利率风险

在股票市场上，股票的交易价格是按市场价格进行，而不是按其票面价值进行交易的。市场价格的变化也随时受市场利率水平的影响。当利率向上调整时，股票的相对投资价值将会下降，从而导致整个股价下滑。

3、购买力风险

由物价的变化导致资金实际购买力的不确定性，称为购买力风险，或通货膨胀风险。一般理论认为，轻微通货膨胀会刺激投资需求的增长，从而带动股市的活跃；当通货膨胀超过一定比例时，由于未来的投资回报将大幅贬值，货币的购买力下降，也就是投资的实际收益下降，将给投资人带来损失的可能。

4、市场风险

市场风险是股票投资活动中最普通、最常见的风险，是由股票价格的涨落直接引起的。尤其在新兴市场上，造成股市波动的因素更为复杂，价格波动大，市场风险也大。

* + 1. 非系统性风险

非系统性风险一般是指对某一个股或某一类股票发生影响的不确定因素。如上市公司的经营管理、财务状况、市场销售、重大投资等因素，它们的变化都会对公司的股价产生影响。此类风险主要影响某一种股票，与市场的其它股票没有直接联系。主要有以下几类：

1、经营风险

经营风险主要指上市公司经营不景气，甚至失败、倒闭而给投资者带来损失。上市公司经营、生产和投资活动的变化，导致公司盈利的变动，从而造成投资者收益本金的减少或损失。例如经济周期或商业营业周期的变化对上市公司收益的影响，竞争对手的变化对上市公司经营的影响，上市公司自身的管理和决策水平等都可能会导致经营风险，如投资者购买垃圾股或低价股(\*ST)就可能承担上市公司退市风险。

2、财务风险

财务风险是指公司因筹措资金而产生的风险，即公司可能丧失偿债能力的风险。公司财务结构的不合理，往往会给公司造成财务风险。公司的财务风险主要表现为：无力偿还到期的债务，利率变动风险，再筹资风险。形成财务风险的主要因素有资本负债比率、资产与负债的期限、债务结构等因素。一般来说，公司的资本负债比率越高、债务结构越不合理，其财务风险越大。

3、信用风险

信用风险也称违约风险，指不能按时向股票持有人支付本息而给投资者造成损失的可能性。此类风险主要针对债券投资品种，对于股票只有在公司破产的情况下才会出现。造成违约风险的直接原因是公司财务状况不好，最严重的是公司破产。

4、道德风险

道德风险主要指上市公司管理者的不道德行为给公司股东带来损失的可能性。上市公司的股东与管理者之间是一种委托代理关系，由于管理者与股东追求的目标不一定相同，尤其在双方信息不对称的情况下，管理者的行为可能会造成对股东利益的损害。

* 1. 股票投资风险的理论背景

股票投资风险一般指未来投资收益的不确定性 ,即实际收益率可能偏离期望收益率的幅度。

* + 1. 风险衡量指标

1952年，芝加哥大学的亨利·马科维茨教授在其发表的《资产组合选择》一文中，首次采用股票投资收益率历史数据的方差，作为风险衡量指标，并将投资总风险划分为系统风险和非系统风险两类，指出与证券市场的整体运动相关联的宏观系统风险，如购买力风险、利率风险、政策风险、市场风险等不能通过投资分散化加以消除；而只影响某一具体证券的微观非系统风险，如公司破产风险、流动性风险、违约风险、管理风险等却可以通过同时投资于多种股票加以弱化。以此同时，马科维茨教授在投资者效用最大化的基础上，将复杂的投资决策问题简化为一个风险（方差）—收益（均值）二维问题，即在相同的期望收益条件下，投资者选择风险最小的证券（组合）；或者在相同的投资风险下，选择预期收益率最大的证券（组合）。

* + 1. 资本资产定价模式（CAPM）

威廉·夏普教授在马科维茨均值方差模型的基础上，建立了均衡的证券定价理论，即著名的资本资产定价模式(Capital Asset Pricing Modlel)，简称 CAPM，其数学表达形式为：

其中，E(Ri)为股票(组合)i的预期收益率，Rf 代表无风险利率，E(Rm)为市场组合的预期收益率β，i = Qim/ Qm2，其中Qim是股票 i 收益率与市场组合收益率的协方差，而Qm2是市场组合收益率的方差，所以βi 用于表示股票 i 收益率变动对市场组合收益率变动的敏感度，即可以用βi 系数来衡量该股票系统风险的大小。

CAPM的核心思想是：在证券市场上，由于非系统风险可以通过投资多元化加以消除，所以市场参与者对该种风险不会给予收益补偿，而对预期收益产生影响的只能是无法分散的系统性风险。自70年代以来，对股票投资风险的分析及对CAPM的实证检验已成为现代金融理论研究中最活跃的领域之一，并在世界各国股票市场上获得了不同程度的验证。

* + 1. 系统性金融风险压力指标的构建

借鉴现有理论研究和相关监管标准对上述指标体系进行压力指数的量化。 通过设定最优值和预警值，运用转换函数来衡量指标反映的压力大小，最终保证压力值 F(x)∈[0,1]。股价指数、实际利率和实际汇率等部分指标需计算指标变动率再结合转换函数量化压力大小。在权数计算方面，可以通过专家调查问卷、层次分析法等主客观相结合的赋权方法，也可以通过计算各指标的精度来确定相对权重。 通过权重汇总各项指标得分，计算各项压力指数，最终衡量区域系统性金融风险的大小及主要影响因素如下：

1、宏观经济。

宏观经济是金融体系稳健运行的根本条件，故宏观审慎指标是金融稳定评估的先行指标。 由于金融体系对宏观经济运行的过度杠杆化暴露在宏观经济逆转时期，去杠杆化过程中资产价格的螺旋下降可能危及整个金融市场的安全，这通常被称为宏观区域系统性金融风险。选取GDP增长率、出口变动率和通货膨胀率等项指标汇总合成区域经济压力指标Et：

2、金融机构。

系统重要性金融机构的破产容易引发人们对类似或相关金融机构的信任危机，从而对整个区域金融产生冲击，这被称为微观区域系统性金融风险。 鉴于我国金融体系仍以银行为主导，采用信贷比率（I1）（反映信贷过度风险）、贷存比（I2）（反映流动性风险）以及不良贷款率（I3）（反映金融机构资产质量）的加权值来衡量银行业机构压力指数It：

3、金融市场。

Allen、Gale（2000）研究表明金融危机经常伴随资产价格中出现的泡沫，风险转移能增加投资者的资产回报，也造成投资者投机性投资哄抬资产价格使其远高于基本价值。金融自由化使这种投机性投资的借贷量极度膨胀，导致资产价格泡沫出现。股票市场和房地产市场是资产市场泡沫的最主要载体，两者相互联系，都是产生经济危机的重要根源。为此，本文采用股票价格指数（M1）变化程度和房屋销售价格指数（M2）变化情况来反映资本市场的波动程度和压力大小：

* 1. 机器学习简介

机器学习(Machine Learning, ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

* + 1. 机器学习基本结构

表示学习系统的基本结构。环境向系统的学习部分提供某些信息，学习部分利用这些信息修改知识库，以增进系统执行部分完成任务的效能，执行部分根据知识库完成任务，同时把获得的信息反馈给学习部分。在具体的应用中，环境，知识库和执行部分决定了具体的工作内容，学习部分所需要解决的问题完全由上述3部分确定。下面我们分别叙述这3部分对设计学习系统的影响。

影响学习系统设计的最重要的因素是环境向系统提供的信息。或者更具体地说是信息的质量。知识库里存放的是指导执行部分动作的一般原则，但环境向学习系统提供的信息却是各种各样的。如果信息的质量比较高，与一般原则的差别比较小，则学习部分比较容易处理。如果向学习系统提供的是杂乱无章的指导执行具体动作的具体信息，则学习系统需要在获得足够数据之后，删除不必要的细节，进行总结推广，形成指导动作的一般原则，放入知识库，这样学习部分的任务就比较繁重，设计起来也较为困难。

因为学习系统获得的信息往往是不完全的，所以学习系统所进行的推理并不完全是可靠的，它总结出来的规则可能正确，也可能不正确。这要通过执行效果加以检验。正确的规则能使系统的效能提高，应予保留；不正确的规则应予修改或从数据库中删除。

知识库是影响学习系统设计的第二个因素。知识的表示有多种形式，比如特征向量、一阶逻辑语句、产生式规则、语义网络和框架等等。这些表示方式各有其特点，在选择表示方式时要兼顾以下4个方面：(1)表达能力强 (2)易于推理 (3)容易修改知识库 (4)知识表示易于扩展。

对于知识库最后需要说明的一个问题是学习系统不能在全然没有任何知识的情况下凭空获取知识，每一个学习系统都要求具有某些知识理解环境提供的信息，分析比较，做出假设，检验并修改这些假设。因此，更确切地说，学习系统是对现有知识的扩展和改进。

执行部分是整个学习系统的核心，因为执行部分的动作就是学习部分力求改进的动作。同执行部分有关的问题有3个：复杂性、反馈和透明性。

* + 1. K-means聚类算法

k-means 算法接受输入量k；然后将n个数据对象划分为k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来进行计算的。

工作过程说明如下：首先从n个数据对象任意选择k个对象作为初始聚类中心；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

k-means算法是一种基于样本间相似性度量的间接聚类方法，属于非监督学习方法。此算法以k为参数，把n个对象分为k个簇，以使簇内具有较高的相似度，而且簇间的相似度较低。相似度的计算根据一个簇中对象的平均值（被看作簇的重心）来进行。此算法首先随机选择k个对象，每个对象代表一个聚类的质心。对于其余的每一个对象，根据该对象与各聚类质心之间的距离，把它分配到与之最相似的聚类中。然后，计算每个聚类的新质心。重复上述过程，直到准则函数收敛。k-means算法是一种较典型的逐点修改迭代的动态聚类算法，其要点是以误差平方和为准则函数。逐点修改类中心一个象元样本按某一原则，归属于某一组类后，就要重新计算这个组类的均值，并且以新的均值作为凝聚中心点进行下一次象元素聚类；逐批修改类中心：在全部象元样本按某一组的类中心分类之后，再计算修改各类的均值，作为下一次分类的凝聚中心点。

1. 股票数据的选取与处理
   1. 各项样本及其数据的选取

对于的股票种类的选取来源于eoddata网站中纳斯达克（NASDAQ）列表中按字母A-Z排序的列表。为了简化计算，减少计算量，我们仅仅选取了A-C之间的股票代码缩写的列表。为了抓取列表信息，实验使用了正则表达式过滤无效信息，表达式如下。

本文的股票数据来源于雅虎财经（YAHOO FINANCE）的数据接口，我们通过使用Python语言实现的爬虫程序从网络上进行数据抓取。在时间范围的选择上面，选取了2009年1月1日至2014年1月1日的数据，避开了无效数据较多的2008年（注：由于2008年经济危机的影响）。在保存为CSV格式之后，数据的格式如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Volume | Adj Close |
| ######## | 4.68 | 4.77 | 4.64 | 4.73 | 95500 | 4.73 |

由于节假日，某只股票停盘等不同因素的影响会导致股票不能呈现出周期性，对于部分无效的数据我们采取了以下的措施：

如果对于数据中的一周5天内开盘的时间少于4天，那么这一周的数据被标定为无效。如果有任意一年的有效有数的天数少于220天，那么这只股票就被标定为无效数据。

另外我们通过从<http://www.nasdaq.com/>这个网站中寻找数据，得到了比较权威的相关公司的市值，IPO时间等一系列信息，得到的数据信息的格式如下图 ：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Symbol | Name | | | | LastSale | MarketCap |
| AAAP | Advanced Accelerator Applications S.A. | | | | 36.81 | 198513826.9 |
| ADR TSO | | **IPOyear** | **Sector** | **Industry** | | **Summary Quote** |
| 5392932 | | 2015 | Health Care | Major Pharmaceuticals | | <http://www.nasda>  q.com/symbol/aaap |

* 1. 各项样本及其数据的预处理

本文在建立股票风险模型之前，需要有相关数据进行预处理的过程。由于开盘价（OPEN）、闭盘价（CLOSE）、最高价（HIGH）、最低价（LOW）、成交量（VOLUME）都应该作为实际的研究结果的影响因素。受到计算机计算的精度的显示，过大或者过小的数据会超出CPU的精度，产生相应的误差，并在以后的处理过程中积聚误差，对于模型的可靠性产生极大的影响。所以我们需要进行相关处理。

* + 1. 数据的归一化

运用归一化的方法进行数据简化，将数量极大的数据和数量级小的数据进行同一化，有利于减少模型的预测误差。

在这里本实验进行了化简，将常量MAX和MIN分别设定为1和0，于是得到以下的公式：

* + 1. 数据的聚集化

本实验通过对多组数据进行聚集达到简化数据的目的，举个例子，我们可以使用某些相近的股票的某些属性，对于不同股票赋予不同的权重，求出一个具有代表性的平均量。

* + 1. 预测模型的评价标准

为了能够检测模型的拟合和预测效果，本文采用了均方根误差的误差评判标准：

通过均方差的误差分析能够确定模型的实际置信度，并能够反馈改进风险模型。该公式得出的值越小表明预测的效果比较好，越大表明预测的结果比较差。

* 1. 评估指标量的计算

综合文献综述中对于股票收益和投资风险之间关系的分析，我们可以归纳出若干影响因素对于实际模型具有影响作用，实验选取以下几个指数作为不同方面指标的衡量标准，分别是上市公司规模（SCALA）、账面/市值比(BV/MV)、股价的年化价格波动率。

* + 1. 账面/市值比

通常观点认为股票投资收益率通账面/市值比（BV/MV）存在正相关的关系，从而作为影响市场预期风险的一个重要指标。另外从直观上看，如果一个公司的股票具有较高的账面/市值比，该公司在市场中的成长预期相比其他公司就会低，而另外一方面该公司的风险就会变高，所以账面/市值比作为衡量股票投资收益率的重要标准列入了指标之中。

根据数据的范围，我们采用了以下步骤对于账面/市值比进行计算：

首先计算股票的每周的收益的比率，根据周收益率利用前面数据聚合的方法等权重计算出平均股价。

根据所有周的周收益率计算得出不同年份的周平均收益率。

通过平均股价和年化周平均收益率相除，近似计算出的值作为账面/市值比的值。

* + 1. 上市公司规模

公司的规模（SCALA）可以从多个角度来衡量。如果是比较注册资本规模，如果都是上市公司，可以比较它们的股本；如果是比较经营规模，那就需要比较主营业务收入。上市公司的股票具有流通性，但是股票价格中并不含有控制权；标的企业收购一般属于控制权收购，但股权的流通性较差。因此，在确定标的企业乘数的过程中需要将控制权与流通性因素予以充分考虑。

由于本文进行分析的都是目前已有的上市公司，于是就可以通过所有公司在某一个时间的市值作为衡量标准去划分上市公司的实际规模。

* + 1. 股价的年化价格波动率

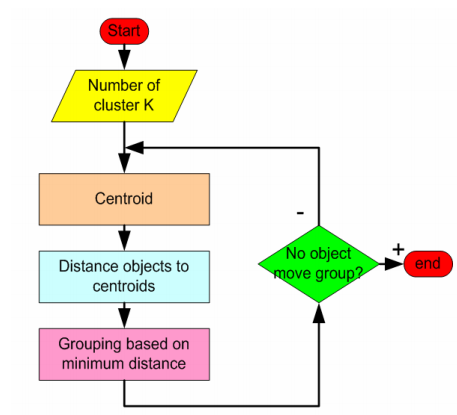
波动率分为两种，一种是回望型波动率（backward looking），另外一种是前瞻波动率（forward looking）。前者是用历史数据算出来的波动率，后者是根据现在的期权价格，用 B-S 期权定价模型反推出来的波动率。前者是已经发生了的历史价格的波动，我们算一个波动率，后者是我们对未来一个价格的波动率的预测，未必准确的。

历史价格波动率最准确的计算方式是把各个时间段的标准差的平方，然后加起来平均，最后年化。

1. 通过聚类分析对所选样本股票进行风险分类
   1. K聚类算法的具体设计

在数据挖掘中，K-Means算法是一种cluster analysis的算法，其主要是来计算数据聚集的算法，主要通过不断地取离种子点最近均值的算法。K-Means的算法如下：

1. 随机在图中取K个种子点。
2. 然后对图中的所有点求到这K个种子点的距离，假如点Pi离种子点Si最近，那么Pi属于Si点群。
3. 接下来，我们要移动种子点到属于他的“点群”的中心。
4. 然后重复第2和第3步，直到种子点没有移动。



在整个算法中，我们需要输入的数据有聚类的数量K，以及相关的原始数据N组及他们的不同属性（SCALA、BV/MV、Di），我们可以得出这N组数据分为K类的方法。

对于整个程序中比较难以确定的是K的数目，因此我们采用了一种自适应的方法，通过令K值随机化，再通过反馈聚类方法的好坏（通过反馈评估函数得到），进一步变化K的值，通过二分的方法最终得到最优的K的值。

通过进行归一化的处理和聚合我们得到的部分结果如表3所示，将不同种影响因素归一化。

1. 结 论

**提示信息：**论文应有结论。论文的结论是最终的、总体的结论，不是正文中各段的小结的简单重复。

结论应包括论文的核心观点，列出论文的创新之处，交待研究工作的局限，提出未来研究工作的意见或建议。

结论应该观点明确、严谨、完整、准确、精炼。文字必须简明扼要。如果不可能导出应有的结论，也可以没有结论而进行必要的讨论。

结论是论文的“收尾之笔”，应是“点睛之笔”，应认真阐明本人在科研工作中创造性的成果和新见解，在本领域中的地位和作用，新见解的意义。结论中不要简单重复罗列实验结果，要对存在的问题和不足作出客观的叙述，并提出进一步的设想。应严格区分自己的成果与他人（特别是导师的）科研成果的界限。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

参考文献

1. 石予友，仲伟周，马骏，陈燕. 股票的权益比、账面市值比及其公司规模与股票投资风险——以上海证券市场的10只上市公司股票投资风险为例[J/OL]. 金融研究，2008,06:122-129.
2. 施东晖. 上海股票市场风险性实证研究[J/OL]. 经济研究，1996,10:44-48.
3. 黄峰，杨朝军. 流动性风险与股票定价:来自我国股市的经验证据[J/OL]. 管理世界，2007,05.
4. 张雅慧，万迪昉，付雷鸣. 股票收益的媒体效应:风险补偿还是过度关注弱势[J/OL]. 金融研究，2011,08:143-156.
5. 张瑾. 基于金融风险压力指数的系统性金融风险评估研究[J/OL]. 上海金融, 2012,09.
6. 李毅学. 供应链金融风险评估[J/OL]. 中央财经大学学报, 2011,10.
7. 周浩, 康建伟, 陈建华, 包松. 蒙特卡罗方法在电力市场短期金融风险评估中的应用. 中国电机工程学报. 2004,12.
8. Poon, S.H., M. Rockinger, and J. Tawn (2004), “Extreme Value Dependence in Financial Markets: Diagnostics, Models, and Financial Implications,” Review of Financial
9. Studies, 17, 581–610Yuqing Dai, Yuning Zhang. A machine learning approach for stock price prediction. IDEAS '14, 2014.
10. Bjoern Krollner. Risk Management in the Australian Stockmarket using Artificial Neural Networks. Austr. J. Intelligent Information Processing Systems 13(2). 2012.
11. Amit Agarwal, Elad Hazan, Satyen Kale, and Robert E. Schapire. Algorithms for portfolio management based on the newton method. In Proceedings of International Conference on Machine Learning, pages 9-16, 2006.
12. 机器学习. 百度百科(EB/OL). http://baike.baidu.com/view/7956.htm
13. Eyal Gofer. Machine Learning Algorithms with Applications in Finance. Senate of Tel Aviv University. March 2014
14. Maheu, J.M. and T. McCurdy (2011), “Do High-Frequency Measures of Volatility Improve Forecasts of Return Distributions?” Journal of Econometrics, 160, 69–76.\
15. Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2010), “Cascades in Networks and Aggregate Volatility,” Manuscript, MIT.
16. Alizadeh, S., M.W. Brandt, and F.X. Diebold (2002), “Range-Based Estimation of Stochastic Volatility Models,” Journal of Finance, 57, 1047–1091.
17. 外文文献原文

**提示信息：**此附录部分，包括两部分内容：（1）英文文献的文献著录资料，基本同参考文献列表中相关信息，可适当拓展；2）英文文献的原文。

对于英文原文，应基本保持原样。具体做法有：1）制作多页含页眉页脚的空白页，将打印好的英文原文剪贴于该空白页上，然后复印；2）将电子版的英文原文（或扫描制作电子版英文），通过图片剪截放置于本部分；3）根据英文原文，将英文原文的内容，输入到本页；4）或者其他方法等。

无论采取任何方法，请尽量保持本页的内容清晰，字号和图表等不要太大或太小。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

1. 外文文献译文

**提示信息：**此附录部分，包括两部分内容：（1）英文文献的文献著录资料，基本同参考文献列表中相关信息，可适当拓展；2）英文文献的中文译文，含论文题目、作者信息等。

本部分的译文应在内容上与原文基本一致，在格式上应与本毕业论文的正文相协调，应保持基本美观。文字基本采用“b正文”样式、“b图标题” 样式、“b表标题” 样式等，1级标题建议采用“黑体、小三”，2级、3级标题建议采用“黑体、四号”，图、表、公式等应清晰、字体不易过大过小，行距基本为1.3倍行距。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）



**提示信息：**根据需要，可以再增加附录，比如“附录C”、 “附录D”、“附录E”等，比如文中的原始数据、文中的调查问卷样板、文中的程序清单、文中的公式推导等。

本部分在格式上应与本毕业论文的正文相协调，应保持基本美观。文字基本采用“b正文”样式、“b图标题” 样式、“b表标题” 样式等，1级标题建议采用“黑体、小三”，2级、3级标题建议采用“黑体、四号”，图、表、公式等应清晰、字体不易过大过小，行距基本为1.3倍行距。

1级标题采用“C.1”、 “C.2”、“C.3”等编号方式，2级标题采用“C.1.1”、 “C.1.2”、“C.2.1”等编号方式，3级标题的编号方式类似。图表标题的编号类似，如“图C.1”、“表C.3”等。公式编号类似，如“（式C.1）”。

不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）

在学取得成果

1. 在学期间所获的奖励

（应注明奖励名称、授奖机构、授奖时间等）。不用此信息时，请删除本段。

1. 在学期间发表的论文

（应按照参考文献的格式来填写）。不用此信息时，请删除本段。

1. 在学期间取得的科技成果

（应注明课题名称、参加身份、通过时间、通过方式、评定机构等）。不用此信息时，请删除本段。

此部分不是可选部分，若无内容，请保留本部分的这三个标题，并在各标题间插入两行空行。

不用此信息时，删除此框。

致 谢

**提示信息：**论文作者可以在致谢页对下列方面致谢：

* 国家科学基金，资助研究工作的奖学金基金，合同单位，资助或支持的企业、组织或个人；
* 协助完成研究工作和提供便利条件的组织或个人；
* 在研究工作中提出建议和提供帮助的人；
* 给予转载和引用权的资料、图片、文献、研究思想和设想的所有者；
* 其他应感谢的组织或个人。

学位论文的致谢用语应规范，感情真挚，言辞恳切，谦虚诚恳，实事求是。致谢辞体现着作者的学术修养和文化修养。

致谢辞主要感谢导师和对论文工作有直接贡献及帮助的人士和单位。学位申请人的家属及亲朋好友等与论文无直接关系的人员，一般不列入致谢的范围。

致谢请控制在一页以内。不用此信息时，删除此框。

（鼠标移到此框四边，鼠标变为十字箭头，点击边框选中此框，然后按Del删除）