

**北 京 科 技 大 学**

本科生毕业设计(论文)选题报告

机器学习在股票投资

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

风险评估中的应用

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

计算机与通信工程学院

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

计算机科学与技术

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

马腾

姓　　名： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

41255092

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

指导教师签字： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2016 年 03 月 24 日

目 录

[1 课题背景及研究意义 1](#_Toc446572109)

[1.1 课题背景 1](#_Toc446572110)

[1.2 研究意义 2](#_Toc446572111)

[2 文献综述 3](#_Toc446572112)

[2.1 股市风险的类型 3](#_Toc446572113)

[2.1.1 系统性风险 3](#_Toc446572114)

[2.1.2 非系统性风险 4](#_Toc446572115)

[2.2 股票投资风险的理论背景 4](#_Toc446572116)

[2.2.1 风险衡量指标 5](#_Toc446572117)

[2.2.2 资本资产定价模式（CAPM） 5](#_Toc446572118)

[2.2.3 系统性金融风险压力指标的构建 5](#_Toc446572119)

[2.3 机器学习简介 6](#_Toc446572120)

[2.3.1 机器学习基本结构 7](#_Toc446572121)

[2.3.2 K-means聚类算法 8](#_Toc446572122)

[3 研究内容，预期目标及研究方法 9](#_Toc446572123)

[3.1 研究内容 9](#_Toc446572124)

[3.2 预期目标 9](#_Toc446572125)

[3.3 研究方法 9](#_Toc446572126)

[4 研究进度安排 10](#_Toc446572127)

[参考文献 11](#_Toc446572128)

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. 研究内容，预期目标及研究方法
   1. 研究内容

风险评估是当前金融市场投资分析中一个重要组成部分，然而对于风险程度并没有较好的量化评估和聚类分析的实现。本课题以为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。

机器学习在股票投资风险评估中的应用主要内容包括：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。

* 1. 预期目标
* 理解市场风险的来源。
* 掌握股票市场中的若干技术分析手段。
* 分析市场实时数据与市场风险的内在联系。
* 掌握金融工程学测试与分析的基本方法
  1. 研究方法
* 先理论分析后编程验证

对于工作量比较大的某些工作需要先进行理论推导验证可行性，再实现模型，通过测试验证理论的分析。

* 由局部到整体

先研究股市风险框架之下的某些小问题，再通过整合问题，合并解决整体，最终形成完整的框架。

* 系统科学方法

运用系统科学的理论和观点，将研究对象放在整个系统之中，分别从局部和整体，将系统整合到系统与要素、要素与要素、结构与功能对立统一的环境中去，分析考察已有的研究对象，最终得出最优的解决方案。

1. 研究进度安排

第1周, 查找相关文献，阅读并翻译相关外文文献；学习风险学相关内容

第2周, 学习投资学相关内容

第3~4周, 搜集数据，过滤数据，预处理数据

第5~6周, 完成评估方法的设计

第7~9周, 根据得出的结论对比和理论分析调整相关分析方法

第10~11周, 完成整体风险评估系统结构，由数据得出结论

第12周, 与其他系统进行对比测试

第13~14周, 根据实验数据和最终实现的系统撰论文，准备答辩相关的资料

学生本人签字：

年 月 日

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.

指导教师意见

指导教师签字：

年 月 日