
# 南京大学

# 本 科 毕 业 设 计

院　　系　工程管理学院

专　　业　计算机与金融工程实验班

题　　目　基金持股关系推断的社会网络与股价崩盘风险

年　　级　2016 级　学 号 161278015

学生姓名　李康

指导教师　方立兵　职 称　副教授

　　职 称

提 交 日 期　2020 年 5 月 21 日

# 南京大学本科生毕业论文（设计、作品）中文摘要

题目：基金持股关系推断的社会网络与股价崩盘风险

院系：工程管理学院

专业：计算机与金融工程实验班

本科生姓名：李康

指导教师（姓名、职称）：方立兵副教授

摘要：随着我国金融市场的快速发展，股价崩盘给金融市场带来了巨大的不确定性，为了稳定金融市场，为金融市场的监管提供更好的指标。我们从基金经理的持股关系出发，利用其持有的股票，推测其私人关系网络。由于信息会在该网络流传，我们提取出了关于股价崩盘的信息，同时分析了其准确度。并用该指标结合减持份额，得到了减持指标。我们利用该减持指标，通过 BP 神经网络、SVR 模型、CART 模型和随机森林模型预测了股价崩盘。并发现该指标的加入对于提升模型预测能力有着显著的帮助。该模型对于后续的风险监控有着很强的参考意义。

关键词：社交关系网络；机器学习；数据挖掘

# 南京大学本科生毕业论文（设计、作品）英文摘要

THESIS：Inferring Social Network and Stock Collapse from Fund Holding

DEPARTMENT：School of Management and Engineering

SPECIALIZATION: Fintech

UNDERGRADUATE: Li Kang

MENTOR: Prof. Fang Libing

ABSTRACT： With the rapid development of China's financial market, the stock price crash has brought huge uncertainty to the financial market. In order to stabilize the financial market and provide better indicators for the supervision of the financial market. We start from the fund manager ’s shareholding relationship and use the stocks it holds to speculate on its personal relationship network. Since the information will be circulated on the network, we extracted information about the stock price crash and analyzed its accuracy. Combined with the reduction of shares, this indicator was used to obtain a reduction indicator. We use this reduction index to predict the stock price crash through BP neural network, SVR model, CART model and random forest model. And found that the addition of this indicator has a significant help to improve the prediction ability of the model. This model has a strong reference significance for subsequent risk monitoring.

KEY WORDS: Social Network; Machine Learning; Data Mining

# 目录

# 1 绪论

## 1.1 研究背景

改革开放以来，我国的金融市场高速发展，随着金融开放的程度不断加大，越来越多的企业选择上市。但是，于此同时，越来越多的股票发生了突然的崩盘，如何对此崩盘现象进行预警，如何对此进行实时的监控成了时代发展的关键问题。

而学术界和实务界都把证券市场的众多波动，尤其是巨幅波动看作是十分值得人们研究的重要的金融问题，李心丹，陈莹（2010）。而这些问题当中便包括股价崩盘，股价崩盘是一种非理性的金融现象。其背后的原因十分复杂，引发了国内外众多学者的讨论。无论是亚洲金融危机，还是互联网泡沫，还是 2008 年的美国金融危机，股价崩盘对于整个经济社会生活是有着巨大的破坏性影响的。它不仅会打击投资者的进场信心，导致整个社会的股权融资的难度提升。同时也会造成资产的错误定价，从而导致资源的错误配置，阻碍整个实体经济的平稳运行。

同时，历年的股价崩盘，背后都可以看到以下行为的影子。比方说，相对一致的投资行为，机构投资者的大幅度集中减持。相对于股价上涨，股价崩盘发生的速度很快，大多数的投资者没有足够的时间去应对。因而，管理者和投资者都迫切地需要预测股价崩盘。以实现对市场的监控。同时对危机做出有效的应对。

Graham （2004)指出，公司的管理者可能会因为自己的薪资激励、职业声誉等原因，会有意去隐瞒一些对于公司不利的私人信号。但是对于公司有利的信号，他们会选择及时披露。而这些对于公司不利的私人信号，会随着时间的推移不断的累积。当外界环境变化触发信息向整个市场公开是，便会导致对股价的冲击，在严重的时候便会导致股价崩盘。

另一方面，近年来机构投资者的一些行为对预测股价崩盘也逐渐收到越来越多的关注。随着信息时代的到来，众多投资者之间的交流也逐渐频繁起来。个人

投资者会根据自己的社交关系网络来进行投资。例如 Liang 和 Guo（2015）根据 2011 年的中国家庭金融调查得出的数据，指出社会互动推动家庭股市参与。而机构投资者之间也会根据自己的社交关系网络中的私人信号来投资股票。同时作为自己增持减持股票的一个重要的依据。

Baker （2012）和 Stefan （2015）建立了投资者相互之间的“相互信任网络”模型，利用仿真，演化了投资者之间的投资行为，研究了股市当中的羊群行为探究了社交关系网络对投资行为的一定影响。

郭白滢（2019）也指出，机构投资者之间除了相互竞争只外，也会选择信息共享。这样的行为提高了市场的定价效率。因而对股价崩盘的风险也有着重要的影响。在相当多的情况下，选择信息共享，其实是机构投资者基于自身的利益，审慎思考后做出的决定，陈新春（2017）。此外 Han 和 Yang（2016）也指出投资者之间的社交关系网络是金融市场私人信息传播的有效途径，同时也是主要途径。网络对信息传播有着决定性的作用。同时，很多机构投资者会通过分享信息，一方面他人对于该信息的反应，可以作为博弈中，检测该信息是否真实的一个手段；另一方面，如果他人同意该观点，Stein（2008）； Crawford（2017）指出，其他人也会选择提供有价值的信息作为回报；此外，其他的投资者如果认同该私人信号，并选择跟随，Gary(2012)指出，会增加该投资的期望收益；最后，通过共享信息，机构投资者扩大了可以选择的资产的范围，Gray(2008)证实，使得他们的投资组合可以足够分散化，进而实现降低投资组合短期波动的目的，最终可以稳定该机构投资者的资金流。

Black，1986; Kumar，2009; 李广川等，2009; 胡昌生和池阳春，2013 指出，长久以来，散户为主的投资者结构是造成资本市场的剧烈股价波动的主要诱导性因素。为提升金融市场的稳定性，我国也在逐步增加机构投资者的数量。证监会在 2001 年就发文并开始实施了“超常规发展机构投资者”的策略。在此背景下，我们的基金行业深入发展。随着我国金融的繁荣程度不断加深，机构投资者也逐步增加。其中不乏公募基金的影子。随着时间的推移，公募基金在整个市场上的影响已经开始举足轻重。因而，研究公募基金的基金经理之间的社交关系网络，

逐渐有了实际意义。同时，作为市场中的重要的私人信号的发掘者，公募基金的私人信号，对于预测某个将要崩盘的公司有着重要的意义。

同时，众多模型中，很多仅仅采用了重复的重仓股票来刻画基金之间的关系。例如：肖欣荣（2012）等。但是 Hong（2005）也指出了，基金经理的信息会在同一个城市内聚集。因而也有必要考虑到基金经理在不同城市的投资组合头寸，对于预测社交关系网络的帮助。

此外，目前的研究模型，大多利用线性回归来辅助预测，但是在不同的场景下未必有着较好的实际预测效果。本文将引入机器学习来着手解决这一问题。

在可以遇见的未来，使用机器学习模型来解决金融问题，尤其是预测和优化问题，将是一个历史潮流。

## 1.2 研究问题

本研究从投资组合的分布在不同城市的权重中推测基金经理在不同城市的私人信号的精确度，其次利用该精确度和对某个特定股票的减持幅度，来预测该股票的崩盘风险。

## 1.3 研究意义

本研究建立在我们公募基金的真实数据之上，并从应对股价崩盘的实际需求出发，结合计算机与金融工程的理论研究，利用机器学习和数据挖掘，来利用基金经理的隐藏信息和市场公开数据，预测某个特定股票的崩盘风险。

该实验实现了股价崩盘的预警，具有较强的实际意义。同时我们的实验结果还表明，一些特定的机器学习方法在预测股价崩盘，监督市场风险方面，具有较强的实际意义。

本模型不仅可以为个人投资者和机构投资者，提供减持、平仓预警。同时还可以为市场监管部分，如证监会等提供对某些股票即将崩盘的预警。对于多只在同一个行业的股票假如同时存在崩盘风险，市场管理者还可以利用本实验来对整

个市场进行宏观调控。同时也为政府出台金融政策提供了有利的帮助。

此外，本研究还可以为公司本身所利用。一些公司在利用本框架监测到股价即将崩盘的之前，可以积极调整战略，控制公司资金风险、运营风险等。为公司的长远发展做出及时积极的调整和打下坚实的基础。

## 1.4 论文结构安排

不断调整与优化，为机器学习在本领域的发展提供了针对现实场景改进的可能性，具有一定的理论意义。

第一章为绪论，具体介绍了本研究的相关背景、具体问题以及实际意义。

第二章刻画了本文的理论和方法。首先，本文从文献出发，提供具体的前人理论支撑。接着，本文介绍了实验的逻辑，阐述了如何实现具体的理论。最后，本文提供了实验所需要用到的模型，阐述了开展实验所需要的具体工具。

第三章描述了实证研究的数据来源及数据处理逻辑，同时展示了本研究所需要用到的实验数据。首先，本文阐明了数据的来源，标明了数据的格式和具体变量。接着，本文介绍了对于一些特殊数据和极端数据的预处理方式，比如一些空缺值的处理。其次，本文描述了数据处理的具体逻辑，提供了具体的复现逻辑。最后，本文展示了实验数据的描述性统计。

第四章是本文的实证分析的部分。首先介绍了模型的一些评价的方法。接着，给出了实验的具体结果。最后基于不同的模型，给出了评价和对比分析。

第五章是本文的研究结论，其总结了本文实证研究所得到的具体的结论

第六章是本文的研究展望部分，其综合了以上的实验结果，和国内外已有文献和已有研究，对本实验的未来研究方向给出了一定的展望。

# 2 理论与方法

## 2.1 文献综述

### 2.1.1 国内研究现状

基金经理可以通过自己的努力获得很多信息，按照是否向市场公开，一般分为公有信息和私有信息。其中公有信息指的是投资者可以自由获取的信息。陈新春(2017)指出，私有信息包含两大部分，一方面是投资者自己的主观信息，另一方面是投资者的社会关系网络中流存的信息。

申宇(2013)指出了私有信息是影响机构投资者持仓行为的主要原因。国内研究肖欣荣（2012）指出，“在熊市和震荡市中，基金重仓股票仓位的变化与其基金的网络结构有着显著的正相关关系。” 而投资者的行为具有传染性。投资者一方面会从市场资产价格中学习信息，同时也会观察网络（注：社交关系网络，下同）中其他人的行为和结果，彼此之间相互交流。因而两个投资者的重仓股票如果十分相似，或者在某一段时间内，增持减持步调保持一致。这很有可能意味着，他们获得了类似甚至相同的信号。也有可能意味着他们之间存在着直接或者间接的关联甚至交流。申宇等（2013）认为基金具有信息优势，私有信息促进了基金的隐性交易，并提高了业绩。

肖欣荣（2012）基于 Pareek（2011）也通过实验证实了基金经理的行为传染与来自基金经理网络中的私人信息有关。“基金经理会根据来自网络中的信息而进行交易”，同时也证实了网络结构中投资者的羊群效应。Scharfstein 和 Stein，1990；Prendergast(1996)； 罗真和张宗成 2004 证实，基金业会存在排行榜现象。假如某只基金的表现是落后于市场或者同行业的。会收到相关投资者和同行的质疑和责备的。这样的压力，也会导致基金经理放弃原先的投资规则，在投资行为上呈现某种跟风。呈现处一种同行博弈行为，例如羊群效应。刘京军，苏楚林(2016)通过空间计量，检验了基金现金流的网络溢出对基金自身业绩的影响。

张雪峰（2017）指出：“大多数基金投资者只寻求短期的收益，缺乏长期投

资的耐心，同时也缺乏长期投资的风险承受能力。”同时基金公司施行的末尾淘汰的竞争制度，也会使得基金经理由于职业竞争的压力，而寻求短期利益。追逐市场热点。另外“基金经理处于自身职业发展以及自身的薪资的考虑，会去追求自身利益的最大化。在追求业绩排名和薪资收入的双重刺激下，很难进行长期投资”，这些都会导致，基金经理在得到一个关于股票价格崩盘或者抛售的具体信号时，会在短期将其抛售，而不会等到整个市场全部反应过来，价格已经触底时才展开抛售行为。因而，公募基金的大规模减持行为，一般会发生在股价完全崩盘之前。所以，对于基金经理的减持行为的观测，有着于研究股价崩盘。同时其私人信号的准确度，也于预测股价崩盘的准确与否息息相关。

对于预测崩盘，国内的研究，大多基于基金经理的网络关系和股价崩盘之间的线性关系。但是，较为明显的时，有必要将在考虑到线性关系的基础上，同时考虑非线性关系，以获得更为准确的预测结果。

### 2.1.2 国外研究现状

由于社交关系网络存在的信息传递的作用。一个信息通过投资者的社交关系网络，例如一种主流观点，或者市场上的一种策略，会在不同的机构投资者之间相互传播。因而重仓持有的相同的股票的基金经理之间会彼此存在相互联系，Pareek(2012)。

国外的学者 Bushee & Goodman（2007）；Jiang（2010）指出：基金经理持有的较大仓位的特定股票，主要源于基金经理的私人信息。投资者之间的社会互动与其之间的信息交流也会对其投资行为产生影响。Shiller 和 Pound（1986）证实了“基金经理倾向于预期持有相同股票的其他基金经理之间进行交流，并且其实际的投资行为也会受到他们的影响。” Hong、Kubik 和 Stein（2004）证实了投资者的投资决策是会显著收到其社会互动的影响的。

另一方面，社会网络会显著影响投资者的行为。Cohen、Frazzini 和 Malloy（2008）证明了基金经理更加倾向于投资他们教育网络中存在的公司。例如基金经理会更加愿意去投资公司高层于基金经理有着相同教育背景的公司，例如公司

高层和基金经理毕业于同一个大学同一个专业。同时，。Cohen、Frazzini 和 Malloy（2008）也证实了上述依赖社交关系网络的投资行为会有着很好的投资收益。在另一个角度上看，这样的更好的收益，也会促使投资者最终的投资行为是和自己的社交关系网络紧密相关的，尤其是重仓股票。投资者甚至会为了获得更好的投资收益去获得一些与重仓股票紧密相关的社交关系网络关系。其持仓的大幅变化也会参考该社交关系网络。

这就导致了，即便两个基金经理原先与其重仓股票背后并没有社交关系网络关系。更好的投资收益，也会促使他们去与该股票相关的社交关系网络节点，建立社交关系。从而两位基金经理，最后很可能间接，甚至直接地建立了社交关系。

Hong(2005)的研究指出，基金经理的持仓决策非常容易收到同一个城市内的基金经理的影响。他们在许多情况下都会通过口口相传来传递关于股价的信息。投资者个体在这样的一个行为下，最终形成了一个投资者的信息网络。成为帮助私人信息传播的一个有效的工具。Colla 和 Mele( 2010)中也证实了，社交关系网络中联系“紧密”投资者的投资行为之间会存在显著正相关性。

我们需要从一些可以预测投资行为的信息中去预测基金经理的社交关系网络。比方说从基金经理所持有的重复重仓股票上，或者同一个城市的重仓股票的投资上 Harrison Hong 和 Jiangmin Xu(2019)。尽管有一些社交软件，但是其中关于投资的信息还是一直都很少。在社交账号上的伙伴也未必是投资上的伙伴，也未必对分析他的投资策略有帮助。

Harrison Hong 和 Jiangmin Xu(2019)指出私人信号的精确度会随着他在一个城市的 non-iid（非独立同分布，下用 non-iid 替代）的关联数的增加而增加。关联数也就是他在该城市的社交关系网络的大小。

这种关联数意味着，一个基金经理在该城市的人际关系模型并不是简单的 Erdo¨s and Re´nyi (1959) null 模型（表征普通两个随机路人之间的关联），在这个 null 模型中，人与人之间的关联是独立同分布的。加入我们观测到一些隐藏网络的参数估计是显著地不平均分布，那就意味着在这群人里面，有些管理人对于某个特定的城市的关系是 non-iid 的。

这种非独立同分布的关系，是怎么来的呢。可能是他的亲戚，或者某个朋友在那个城市。可能在那个城市工作。他们也可能是大学的校友，因为他们有的人在那个城市工作，或者知道某个人在那边工作。

举例来讲，一个在哈佛上学的人，可能在波士顿有着不成比例的联系。由于他们是同样的水平同样的专业，所以关系更加稳定，也更可能会对他们的未来决策起着关键的作用。

而在Harrison Hong和Jiangmin Xu(2019)中,投资者是面临着short-sales constraints （卖空约束）和 quadratic trading costs（二次交易成本）的，这就意味着，投资者并不能很随意的得到资金或者随便处置自己管理的基金，在面临着投资哪一个股票时需要认真考虑,同时也不能频繁地更改自己的投资头寸。这就意味着,投资者会尽可能地利用自己的信息优势去投资自己有着丰富而又精确的私人信号的股票。而 short-sales constraints （卖空约束）和 quadratic trading costs（二次交易成本）的假设是合理的，因为Chen et al.（2002)和Gârleanu and Pedersen（2013）中证实了基金经理是被机构禁止shorting（卖空）的。并且较大的风险头寸会导致价格的波动，进而触及机构所设置的风险管理约束。

Harrison Hong和Jiangmin Xu(2019)证明,依靠基经理在不同城市的关系，可以得到很精确的私人信号，同时依赖此指标投资，可以获得更多的收益。这也催使基金经理更多地利用自己在不同城市的关系，去发掘更加精确的私人信号。

在研究基金经理在不同城市的投资组合的权重,可以发掘其在该城市的隐藏社交关系网络，同时利用其网络关系的大小，来刻画其在不同城市的私人信号的相对精确度。

Pareek(2012)利用了基金的重仓股票的网络密度，发现在社交关系网络中，私人信息广泛传递，基金经理有着较为明显的羊群效应，同时这样的现象与基金经理各自的投资风格无关。这从另一方面也意味着，持有相同的重仓股票的基金经理之间具有较强的相互认识，并时常相互沟通的可能性。

Ozsoylev、Walden、Yavuz 和 Bildik（2011）研究证实了，股票市场上的投

资者网络中，信息会通过网络扩散，使得一些投资者在市场其他人交易之前有所行动。进而获得较高的收益。这也意味着有着交为精确的大量私人信号的投资者，一般会在信息在整个市场中大规模扩散之前有所行动。进而该基金经理的调仓行为在某种程度上可以预示，将要到来的股价的变化。比方说大规模减持所预示的股价崩盘。因而对于一个有着较为精确信号的社交关系网络节点，其大幅减持行为一般是先与股价崩盘的。我们在选取时间区间时，可以选取一个时间区间，例如一年，来同时研究减持情况与崩盘，而不需要一定选取两个时间区间。

而根据私人信号的基金经理的减持行为是先于整体市场的行为的。因而我们要考虑的是，关于股票利空的私人信号对于基金经理的行为的影响。

## 2.2 实验逻辑

### 2.2.1 整体实验结构

本研究将利用基金经理在不同城市的私人信号的精确度，以及在不同城市减持的情况来预测股价本盘。总体上本研究分为以下部分的实验。

首先是通过基金经理所持有的股票，利用股票所在相同城市和股票持有权重，来得到不同基金经理之间的关联程度。并通过标准化处理，用关联程度衍生的指标来刻画不同基金经理在不同城市的私人信号的精确度。

其次，由于本实验是检测基金经理的社交关系是否能很好地反馈股价崩盘。实验的本质目的是探究一个变量对另一个变量的相关性。同时给出一些刻画相关性的模型。因而实验并没有对股价崩盘的参考因素，做出时效性的考量。本文设置的时间界面的长度为一年。来检测指标之间的相关性。

此外，本文通过整体市场的收益率和单只股票的收益率之间关系，联系一只股票在一年内的收益率高于平均收益率的天数和低于平均收益率的天数，得到股价崩盘指标。

最后，本文对于研究基金经理的网络指标和股价崩盘指标之间的关系，将会通过机器学习的方法，把线性和非线性关系同时纳入其中。使得本研究结果有着

较强的实际应用价值。

### 2.2.2 基金经理的私人信号

在本文中，我们利用基金经理的私人信号精确度和股票本身的性质来预测股价崩盘。我们对基金经理在某个城市的关联数进行标准化，来表征其私人信号的准确度。

### 2.2.3 使用私人信号预测股价崩盘

在本文中，我们利用基金经理的私人信号精确度和股票本身的性质来预测股价崩盘。

首先是该股票的基本信息，诸如振幅，区间换手率，非系统风险，波动率等。用于预测该股价是否会崩盘。本实验将基金经理的数据进行分类，一部分用于训练模型，而另一部分用于测试模型。

我们在前人研究的基础上加了了基金经理对于一只股票是否会崩盘，以及崩盘程度的判断。

如果实验选取的是整个市场的基金经理，那么他们的私人信号的精确度，对于整个市场是有效的。同样的，选取的是特定市场的基金经理，那么其私人信号的相对大小对该特定市场有效。

本实验根据 Wind 基金分类，选取了中国公募基金市场作为研究。同时排除了被动跟踪的的 ETF 基金，因为该类基金调仓的依据是跟踪具体指数，而不是在卖空约束和二次交易成本下的依赖自己的私人信号和择时判断。由于选取的是公募基金市场，因而每个基金经理的私人信号的准确度大小，在该市场内是有效的。本文用于预测崩盘的股票也选取了公募基金池中的股票。

基金经理根据自己私人信号的相对准确度，减持一定量的股票，对于预测股价崩盘是有着内在联系的。在大幅减持一只特定股票的情况下，一个基金经理的私人信号的相对精确度越大，说明他越有把握该股票会有着崩盘风险。在减持一

只特定股票的情况下，减持幅度越大，说明基金经理对于该股票将要崩盘的倾向越大。

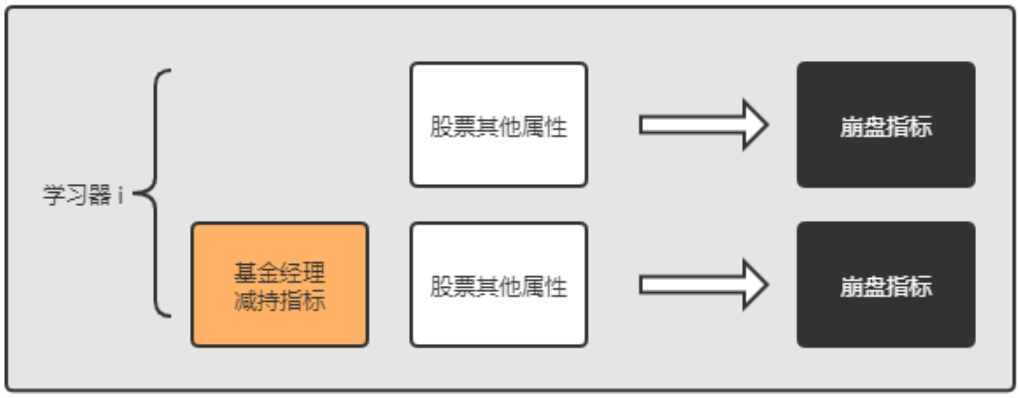
此外， 我们采用了对照组实验，比较在预测回归任务中，加入了基金经理崩盘指标和不加入，对回归的损失函数的影响。从而来反映，该崩盘指标对于实际预测股价崩盘的效果。

### 2.2.4 预测效果比较

本实验中，通过横向比较 4 个学习器，来确定那个学习器更加适合在实际应用中预测股价崩盘。

其中，股票的其他属性为本实验的控制变量，对于控制变量的选取，一定程度上参考借鉴了郭白滢（2019）。

控制变量的指标见数据处理，控制变量的示例见描述性统计。




不同学习器的对照实验示意图

## 2.2 模型简述

本研究采用的机器学习，机器学习为一门学科，机器学习利用研究如何通过计算的办法，使用已有的经验来改善整个系统自身的性能。而在计算机系统中，经验一般是以数据的形式存在的。机器学习的主要任务，就是在计算机的数据中产生“模型”，这样一个算法，也就是我们常说的“学习算法”。通过学习算法，

我们可以把经验数据提供给计算机，让计算机执行学习算法，这样他就可以通过数据来产生新的模型。当这个算法在面对新的数据，也就是新的情况的时候。模型可以为我们提供一个基于学习算法的对应判断。如果说计算机学科是研究算法的，那么机器学习就是研究学习算法的。

机器学习的训练集为，输入数据，让模型通过计算得到模型具体参数的数据集。而机器学习的测试集为，输入数据，通过已经计算好的模型，得到预测结果，与现有结果进行比较，从而测试机器学习效果的数据集。

本研究的学习器均为有监督学习，通过输入不同的属性以及标签，让学习器学习，在输入未知的属性的情况下，如何输出一个正确的标签值。对于学习器的训练集而言，如果同时存在输入属性与输出标签，我们称该学习任务为有监督学习，典型的有监督学习有分类和回归。对于学习器的训练集，如果只存在输入属性而不存在输出标签，我们称之为无监督学习，典型的无监督学习有聚类等。

对于输出的标签值为离散数据时，我们称之为分类问题；对于输出的标签值为连续数据时，我们称之为回归问题。

监督学习的难点在于处理不属于训练集的数据。当输入的数据集，是学习器从未见过的时候，如何得到正确的预测结果。此问题，即如何提高模型泛性。归纳与演绎是科学研究与科学推理的两大基本手段。前者就是是从“特殊”到“一般”的“泛化”过程。即从具体事实到归纳处一般规律。

### 2.2.1 股票减持指标

第一，在本文中，我们通过基金经理之间持有的相同的股票，来预测该基金经理在该城市的关联数。

在 Harrison Hong 和 Jiangmin Xu(2019)中指出，如果基金经理受到卖空约束和二次交易成本，那么基金经理，会更加愿意持有自己有着相对更加精确的私人信号的城市的股票。而这种精确的私人信号，在 Harrison Hong 和 Jiangmin Xu(2019)中指出，是由于其有着相对于其他城市的更多的关联，Harrison Hong 和 Jiangmin Xu(2019)中给出的例子是，大学同学在这个城市工作。所以与投资

关联很大的私人关系的数量，可以在某种程度上很好地表征，该基金经理在该城市的私人信号的精确度。Harrison Hong 和 Jiangmin Xu(2019)中通过实验证明，基金经理之间大量持有的相同的城市的相同的股票，在某种程度上意味着他们很可能直接相互认识，或者通在该城市之间的某个中间人相互认识。两个基金经理通过持有的重仓股票之间的关联，可以作为描述他们与投资相关联的私人关系的指标。本文中，将采用 Harrison Hong 和 Jiangmin Xu(2019)中给出的指标，用来刻画一个基金经理 $i$ 在城市 $k$ 的私人信号的精确度 $y_{i,k}$ 。从而得到所有基金经理在所有城市之间的私人信号的精确度集合 $\{y_{i,k}\}$

对于 $\{y_{i,k}\}$，其具体计算逻辑如下：

本算法是一个三层循环算法

（1） 首先对于一个基金经理 i：

（2） 在某个城市 k:

（3） 对于在城市 k 的其他基金经理 l：

（4） 通过基金经理 i 在该城市 k 的所有投资组合权重 $\bigcup_j \{w_{i,j,k}\}$ 与 基金经理 i 在该城市 k 的所有投资组合权重 $\bigcup_j \{w_{i,j,k}\}$，比较两者之间的较小值（如果有一个权重为 0，则返回 0，意味着通过该股票，无法为两基金经理之间建立关联）求和后得到相对权重

（5） 将（4）中得到的权重求和，即可得到基金经理 i 在城市 k 的关联数

（6） 将关联数，对于整个集合标准化，即可得到基金经理的私人信号精确度 $y_{i,k}$

第二，我们选取了基金经理在过去一年中，减持幅度最大的股票作为研究对象。

对于该股票，我们记减持份数为 $MaxreduceShare$ ，其对应的股价为 $Maxreduce\_price$ ，而该基金经理管理的基金总份额为 $Total\_volumn$ ，则其减

持幅度为：

$$reduceExtent_k = MaxreduceShare_k * Maxreduce\_price_k / Total\_volumn \quad (1.1)$$

第三，本文结合基金经理的信息准确度 $y_{i,k}$ ，对于基金经理减持的最大幅度股票 k，得到信息精确度 $y_k$ 。一支股票的所有基金经理的减持幅度的加总，并进行标准化，则可以得到该股票的减持力度。

可以得到其对于股票 k 的减持力度：

$$reduce_k = reduceExtent_k * y_k \quad (1.2)$$

线性回归是一般量化分析问题中常用的一种统计方法，对于分析各解释变量的影响权重有直观的表述，但由于线性回归的因变量通常是连续的数值，对于一些常见的分类问题并不直接适用。Logistic 回归可以被认为是线性回归模型的一个特例，它的输出值不再是一个连续的数值，而是通过一个 sigmoid 函数，被转化为 0-1 之间的一个接近 0 或 1 的 y 值，因此可以很好地应用于二分类问题，其回归式如下：

$$y = \frac{1}{1 + e^{-(w^T x + b)}}$$

但是，二元因变量违反了一般回归模型的正态性假设。Logistic 回归模型表明，事件拟合概率是解释变量的观测值的线性函数。这种方法的主要优点是它可以产生一个简单的分类概率公式，缺点是无法正确处理解释变量的非线性关系和交互效应问题。

### 2.2.2 DUVOL 指标

DUVOL 指标为股价崩盘指标。国内国外研究者对于股价崩盘的指标都有着一定的研究，Hong and Stein(2003)指出了刻画股价崩盘的标准：

（1） 股价反常的剧烈波动

（2） 负向的价格变动

（3） 崩盘的大规模传染

在数值方法刻画上，国内外也逐渐有了较为公认的指标。其中以收益上下波动比率的应用最为广泛。崩盘指标来自 Chen(2001).作为本研究采用股价崩盘指标。

公司股票收益率的上下波动的比率的计算方法如下：

首先，通过公式（1）来消除市场整体收益率的影响。其中 $r_{i,t}$ 为股票 i 在 t 日的收益率； $r_{M,t}$ 为整个市场在 t 日的平均收益率；残差项表示整个该股票的收益率当中，不能被市场的收益率所解释的部分。

$$r_{i,t} = \alpha_i + \beta r_{M,t} + \varepsilon_{i,t} \tag{1.3}$$

接着，我们定义

$$W_{i,t} = \ln\left(1+\varepsilon_i,\right) \tag{1.4}$$

为该股票该日的特有收益率。

最后，定义一个股票的收益率的上下波动比率。设为 DUVOL ，其具体实现如下：

$$DUVOL = \log\left\{\left[\left(n_u - 1\right)\sum_{Doum} W_{i,t}^2\right] \Big/ \left[\left(n_d - 1\right)\sum_{U_p} W_{i,l}^2\right]\right\} \tag{1.5}$$

对于上述模型，其中 $n_u\left(n_d\right)$ 为股票在第 i 日的回报率低于整年的总回报率的天数。

总体而言，DUVOL 的值越大，该股票收益左偏的程度越大，股价的崩盘风险越高。

如果实验选取的是整个市场的基金经理，那么他们的私人信号的精确度，对于整个市场是有效的。

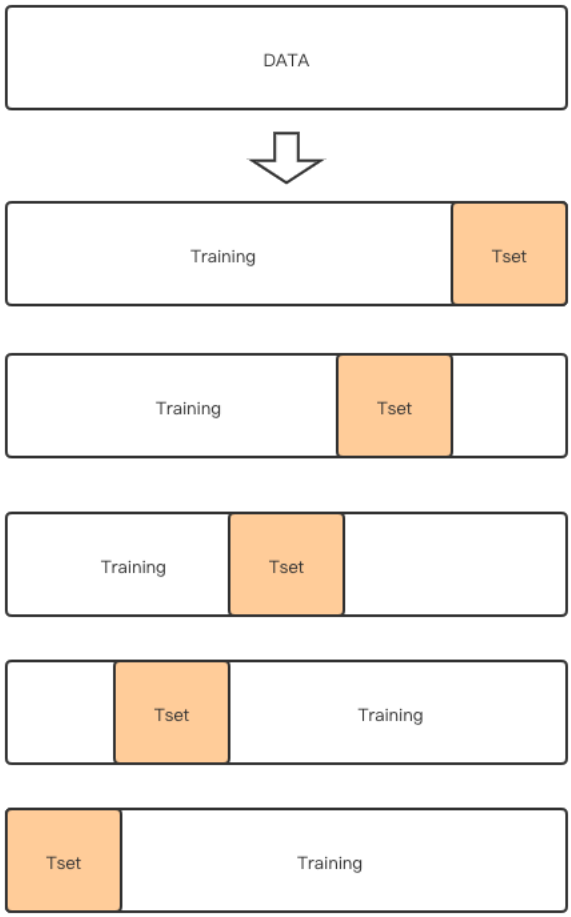
### 2.2.3 K-折交叉验证

K-折交叉验证模型（k – fold cross validation），有被称作循环估计。是一种可以充分利用数据集对算法测试作用的模型。

K-折交叉验证的主要目的在于，减少过拟合和选择偏差的问题，使得整个模型可以在一个新的独立的数据集上实现通用化。

K-折交叉验证模型首先将原数据集划分为测试集和训练集，接着将原来的数据集划分为相等的 K 部分，也就是 K 折。然后，按顺序依次将数据的第 i 部分作为测试集，其余的作为训练集。i 从 1 至 K，一共循环 K 次，每次训练模型后，测试模型训练结果。利用测试集获得训练的准确度。记录每次训练的准确度，将平均准确度作为模型的最终准确度。举例如下（五折交叉验证模型）

K-折交叉验证模型，主要有两个作用。首先，可以充分利用实验数据，实验的训练集相当于增加了 K 倍。在实验数据较为紧张的情况下，K-折交叉验证模型可以较好的实现数据的充分利用。其次，K-折交叉验证模型可以更好地评估算法性能的好坏。可以防止由于选择训练集和测试集有误，而导致地无法较为准确地反应算法的性能。




5-fold cross-validation

### 2.2.4 BP 神经网络

BP 神经网络，即 error Backward Propagation Neural Network. 是一种基于误差逆传播算法的神经网络模型。神经网络并不是我们常说的生物学意义上的神经网络，本文提到的是人工神经网络，与上一个概念存在较大的差距。

神经网络是一个实际应用价值非常大的一种机器学习模型。在金融市场中有着广泛的应用。很多人用其来预测高频交易的交易趋势，也有很多人应用神经网络来实现股价收益的预测。谢衷洁(2001)便指出了神经网络可以用于金融预报，并在很多场景下有着不错的效果。

对于非线性系统，神经网络可实现很好的拟合并同时保存很好的泛性。神经网络具有自主学习的能力，我们在使用的时候，只需要将大量的学习样本放入学

习模型中，神经网络便会通过计算，不断更改自己的连接权重，来学习样本中的数据。而有着自学习的能力对实际的应用有着重要的意义和价值。

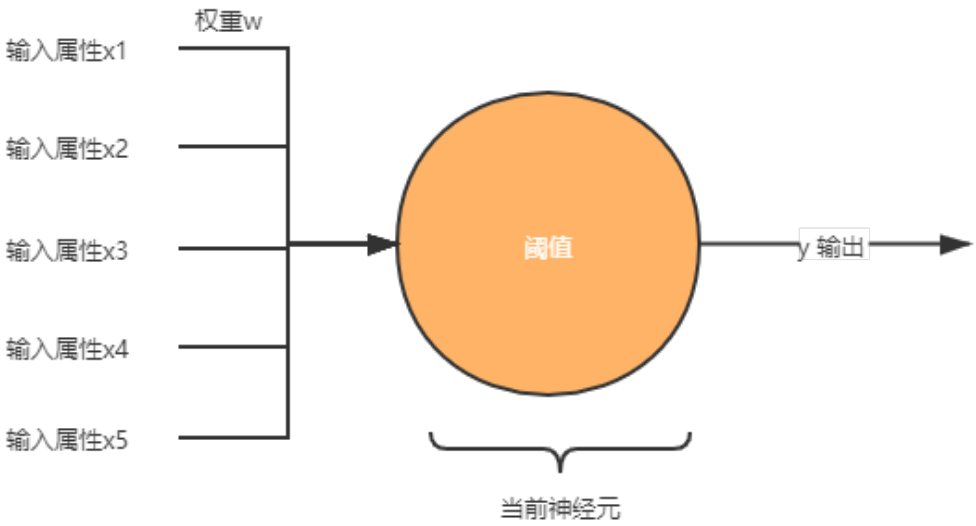
相比于其他的学习器，神经网络具有很强的泛化能力和记忆能力。其中便得益于，其网络结构中包含着大量的参数可以调整，正是这些参数，给了神经网络学习的能力。不同的连接权重和阈值的组合可以抽象出各种复杂的函数。

当有着无限层级的网络在学习时，可以学得所有的函数映射。但是我们在实际应用时。会在满足实际需求的基础上，仅选择较小的层数，以节省计算力，并防止过拟合。

神经网络的网络拓扑结构包括三样：输入层、隐层和输出层。我们在实际应用的过程中，随着任务量的增加，比方说数据量的增加、特征维度的增加，会选择增加隐层的层数，而不是一个隐层的大小。因为提升隐层的层数可以提升计算模型的抽象能力，而增加隐层的大小，只是简单增加了线性关系。

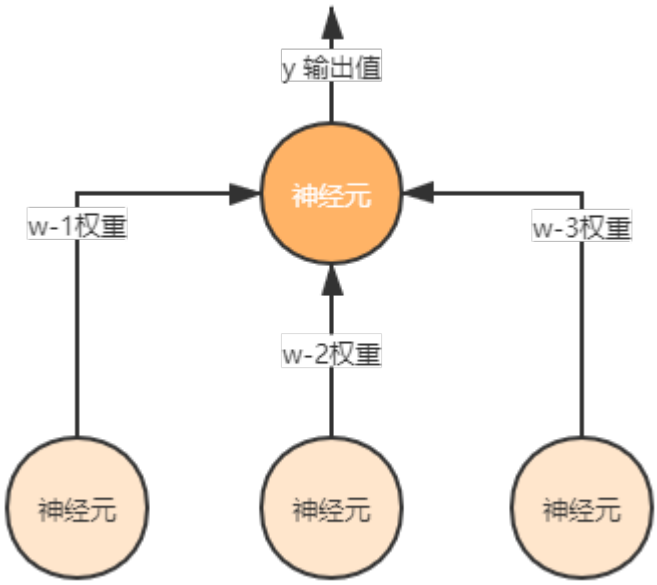
在神经网络中，其基本单元为神经元模型。神经元模型其输入为数据中的特征，在经过权重加权后，与神经元的阈值相比较，然后通过激活函数。为了一般的激活函数为阶跃函数，即将输入的特征值输出为 0 或者 1，代表神经元的激活或者兴奋。但是阶跃函数的连续性太差，在本实验中，采取的是 sigmoid 函数。其表达式为：

$$\mathrm{sigmoid}(x)=\frac{1}{1+e^{-x}} \tag{1.6}$$




M-P 神经元模型

复杂的神经网络是由很多隐层组成的。而为了讲清原理，便不得不提最简单的反馈网络结构，感知机。感知机是一个两层神经元组成的一个结构。其基本逻辑如下：




感知机模型示意图

在训练模型的过程中。连接节点之间的权重以及最终的阈值可以通过学习过程中的计算得到。对于一个感知机，其学习的过程是：

对于训练的样本，设属性为 x，标签为 y，那么感知在输出为 $\hat{y}$ 时，整个感知机的权重将会调整为：

$$\begin{aligned} w_i &\leftarrow w_i + \Delta w_i \\ \Delta w_i &= \eta(y - \hat{y})x_i \end{aligned} \tag{1.7}$$

其中，$\eta$ 便是整个学习过程中的学习率。对于一个感知机，其所有能学到的“知识”也就包含在他的权重和阈值当中。

为了增强学习的能力，我们在单层感知机的基础上，逐渐增加网络的层数。在训练的过程中，我们采用的调整连接权重和阈值的方法是误差逆传播算法。BP 神经网络，它是 Rumelhart 和 McCelland 在 1986 年提出的一种多层反馈神经网络。也是目前应用最为广泛的神经网络模型之一。BP 神经网络可以储存大量的映射关系，随着隐层的增加，它的抽象能力也在飞速增加。BP 神经网络可以描述线性的映射关系，同时也可以描述非线性的映射关系。因而在本实验中有着重要的应用价值。

BP 算法是梯度下降法的一种变体。BP 神经网络的原理可以简单归纳为“信息沿网络传播 – 计算每一层的输出 – 计算误差 – 误差反向传播 - 各层修改权重 – 学习收敛”。每次学习的权重的调整，根据梯度下降法可以算出为：

$$\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}} \tag{1.8}$$

其中 $E_k$ 为训练的均方误差：

$$E_k = \frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k - y_j^k\right)^2 \tag{1.9}$$

整个算法的过程是对于误差函数求导，利用得到的梯度，来设置调整连接权的步长和方向。

对于整个的 BP 神经网络，其具体的算法实现如下：

（1） 输入训练集，和学习效率指标

（2） 在（0，1）之间随机地初始化整个神经网络中所有的连接权重和所有的阈值

（3） 对于所有在训练集中的特征指标和标签指标

（4） 根据当前的所有参数计算当前样本的输出值 yhat

（5） 计算输出层所有的神经元的梯度

（6） 计算隐层中所有的神经元的梯度

（7） 计算连接权与阈值，并将其更新

（8） 重复（3）~（7）直至满足停止条件，如误差低于某个指标，或者迭代次数大于某个指标

（9） 输出所有的连接权，和所有的阈值，得出训练好的神经网络。

### 2.2.5 SVR 模型

SVR 模型，即 Support Vector Regression 模型。又被称作为支持向量回归模型。其最初是 SVM 模型，为一种分类模型，随着时间的推移，逐步被改进，进而使用在回归任务上。为一种监督式的学习模型。

对于一个回归问题，给定一个模型的训练样本 $D=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots(\mathbf{x}_m,y_m)\},y_i\in\mathbb{R}$ ，目标是得到一个 $f(x)=w^{\mathrm{T}}x+b$ 的回归模型。使得 $f(\mathbf{x})$ 和 $y$ 足够接近。其中，$\mathbf{w}$ 和 $b$ 是模型参数，需要通过计算来确定。

那么问题可以转化为：

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}\ell_c\left(f\left(\mathbf{x}_i\right)-y_i\right) \tag{1.10}$$

其中 C 是正则化参数，而 $\ell_\epsilon$ 是损失函数，其具体表达形式为：

$$\ell_\epsilon(z)=\begin{cases}0, & \text{if } |z|\leqslant\epsilon\\ |z|-\epsilon, & \text{otherwise}\end{cases} \tag{1.11}$$

经过计算，SVR 的解为

$$f(\mathbf{x})=\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)\mathbf{x}_i^{\mathrm{T}}\mathbf{x}+b \tag{1.12}$$

带入核函数：

$$\kappa\left(\mathbf{x}_i,\mathbf{x}_j\right)=\phi\left(\mathbf{x}_i\right)^{\mathrm{T}}\phi\left(\mathbf{x}_j\right) \tag{1.13}$$

则最终的 SVR 可以表示为：

$$f(\mathbf{x})=\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)\kappa\left(\mathbf{x},\mathbf{x}_i\right)+b \tag{1.14}$$

### 2.2.6 CART 模型

CART 模型，即 Classification And Regression Tree 模型，又被称作为分类回归树模型。

CART 模型可以检测到输入输出变量之间的非线性关系。使得在一些即便在 OLS 检测下，线性无关的变量，也可以在预测问题中发挥很大的作用。

此外，CART 模型，由于采用了决策树算法，相比深度学习模型，比如 ResNet 还具有较好的可解释性，这也使得决策者可以打开模型“黑箱”，进一步去观测事件发生的原因。也可以利用该实验的最终结果，分析决策树的各个分支，来实现对模型的理解和调控，以实现控制的目的。

其次，CART 模型，在多属性问题中，具有较好的表现。这就意味着，在满足模型的数据量的基础上，决策者可以在模型中加入其他次重要的属性，来实现更为精准的预测效果。这对于本实验而言，是有着重要的意义的。

CART 模型的算法具体流程为：

输入：训练集，基尼系数阈值，样本数阈值。

输出：CART 决策树

算法采用递归，从 CART 决策树的根节点开始

（1） 对于当前节点对应的数据集，若样本数小于阈值，则返回决策树，当前节点递归停止。

（2） 计算样本数据集的基尼系数，如果基尼系数小于阈值，那么就返回 CART 决策树，当前节点递归停止。

（3） 计算当前节点各特征对应的基尼系数。

（4） 对于计算得出的众多基尼系数，选取基尼系数最小的特征。并根据该特征将数据集划分为两个部分。

（5） 对左右两个节点递归调用以上四个步骤，生成 CART 决策树。

对于上述算法中的基尼系数，其定义为：

$$G(\tau)=\sum_{k=1}^{K}P_{\tau}(k)(1-P_{\tau}(k)) \tag{1.15}$$

本实验采用的是 CART 回归树模型。

模型的表达式为：

$$f(x)=\sum_{m=1}^{M}c_{m}I(x\in R_{m}) \tag{1.16}$$

其中真实值与模型的输出值之间的误差可以表示为：

$$\sum_{\text{ithere}}\left(y_{i}-f(x_{i})\right)^{2} \tag{1.17}$$

目标是上述的误差最小。

上述的模型的具体算法步骤可以表示为：

输入：训练集 D

输出：回归树

（1） 首先选择最优切分点，求解：

$$\min_{j,}\left[\min_{\mathbf{q}}\sum_{x\in\mathbb{R}(U,x)}\left(y_{i}-c_{1}\right)^{2}+\min_{\sigma_{i}}\sum_{x,e\mathbb{R}_{i}(j,n)}\left(y_{i}-c_{2}\right)^{2}\right] \tag{1.18}$$

（2） 用上述的求解结果对数据进行划分，并决定其相对应的输出：

$$R_{1}(j,s)=\left\{x\,|\,x^{(n)}\leqslant s\right\},\quad R_{2}(j,s)=\left\{x\,|\,x^{(0)}>s\right\} \tag{1.19}$$

$$\hat{c}_{m}=\frac{1}{N_{m}}\sum_{x_{i}\in R_{s}(j,y)}y_{i},\quad x\in R_{m},\quad m=1,2 \tag{1.20}$$

（3） 对上述的子集继续迭代步骤（1）以及步骤（2），直到满足（1）最小。

（4） 生成决策树：

$$f(x)=\sum_{m=1}^{M}\hat{c}_m I\left(x\in R_m\right) \tag{1.21}$$

### 2.2.7 随机森林回归

随机森林模型，即 Random Forest 模型，是一种典型的 Bagging 算法。同时也是 Bagging 算法的进一步延申与拓展。本文用到是处理回归任务的随机森林模型，即随机森林回归：Random Forest Regression.

首先，对于 Bagging 算法。在集成学习中，如果想要得到更强的泛化能力。那么集成学习中的个体应该尽量相互独立，但是独立这件事在现实当中很难实现。为了解决这个问题，我们设法让基学习器有着尽可能大的差异。但是为了集成的效果，我们同时也需要学习器的不能太差，如果采样的过程中，每个子集都是完全不同的，那么每个基学习器就只能使用到一小部分的数据，很多甚至都无法学习，也就无法保证学习的效果了。

为了解决这个问题，我们采用了 Bagging 算法。它基于自助采样法。给定数据集，我们随机采样一个样本放入采样集。并放回，进行下一次采样。多次采样之后，得到多个样本的采样集。重复上述的步骤，我们可得到多个采样集，然后通过每个采样机来训练对应的基学习器。最后再将这些基学习器相结合，就可以得到最终的学习器。

随机森林算法，建立在 Bagging 算法的基础上。再决策树的训练的过程中，我们加入了随机属性选择的过程。相较于传统的决策树，在选择划分属性的时候是在当前的集合中经过计算得到一个最优的属性。在随机森林中，对于整个学习的每个基学习器，是一个基决策树，对于这个树的每一个节点，我们都先从该节点的所有的属性中随机选取一个包含 k 个特征属性的一个属性子集。接着再用这个子集，从中选取一个最好的属性进行划分。这里的 k 是用于控制随机性的引入的。假如 k 等于原来属性集合的元素总个数，那么这就是传统的决策树算法。

随机森林可以用于处理分类问题，也可以用于处理回归问题。在处理分类问题的时候。对于样本数据，随机森林中的每一棵树都会给出类别判定，最终结合

随机森林中所有树的结果，通过投票法来得出样本的最终类别。在处理回归任务时，对于样本数据，随机森林中的每一棵树都会给出类别判定，最终结合随机森林中所有树的结果，最终判定均值，返回为最终结果。

具体的实现为：

(1) 从数据集中抽取一定数量的样本，作为每一棵树的根节点

(2) 通过计算建立决策树，随机抽取指定数目的候选属性，计算得到其中最优属性作为分裂节点。

(3) 利用上述（2）步骤，重复运行，直至建立好随机森林。通过均值法，结合随机森林中每一棵树的结果，得出最终的回归结果。

随机森林计算开销较小，正好适用于本实验。随机森林中基学习器具有很强的多样性，一方面，这样的多样性来自于样本扰动，另一方面，这样的多样性来自于属性的扰动。这样就可以进一步提升模型的泛化能力。

对于随机森林的收敛性，它其实和 Bagging 算法很相似。虽然通过引入属性的扰动，个体学习器的能力在降低，但是随着学习器数量的增加，整体算法的泛化能力在增加，随即森林会收敛到更低的误差，具有较强的泛化能力。

# 3 数据处理

## 3.1 数据来源与数据示例

本研究的实验数据来源于 Wind 数据库和优矿数据库。

数据来源的主要数据表如下表所示：

外部数据

| 数据表 | 名称 | 释义 | 算法 |
| --- | --- | --- | --- |
| reduceShare | 重仓证券持仓变动 | 报告期末投资组合中重仓证券的 | 报告期末重仓证券的持仓数量-上 |

| | | | |
|---|---|---|---|
| | | 持仓变动 | 个季度末重仓证券的持仓数量 |
| stock_province | 股票省份 | 股票与其对应的省份 | 一一映射 |
| fund | 基金重仓 | 基金重仓以及对应的持股比重 | 按持仓股票市值和基金净值计算的 |
| stock_chara | 股票属性 | 股票的一些用于预测崩盘的属性 | 一一映射 |

数据示例：

reduceShare 表

| 指标名称 | 具体数值 |
|---|---|
| 证券代码 | 673090.OF |
| 证券简称 | 西部利得个股精选 |
| 任职期限最长的现任基金经理 ［名次］第 1 名 | 刘荟 |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 1 名 | 601318.SH |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 1 名 ［单位］份 | 4,300.0000 |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 2 名 | 600519.SH |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 2 名 ［单位］份 | 200.0000 |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 3 名 | 600036.SH |
| 重仓证券 Wind 代码［报告期]去年年报［名次］第 3 名 ［单位］份 | 4,400.0000 |

| | |
|---|---|
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 4 名 | 600276.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 4 名 [单位] 份 | 1,200.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 5 名 | 600030.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 5 名 [单位] 份 | 3,300.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 6 名 | 600887.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 6 名 [单位] 份 | 2,600.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 7 名 | 601601.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 7 名 [单位] 份 | 1,300.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 8 名 | 601328.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 8 名 [单位] 份 | 12,700.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 9 名 | 600016.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 9 名 [单位] 份 | 11,600.0000 |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 10 名 | 600000.SH |
| 重仓证券 Wind 代码 [报告期]去年年报 [名次] 第 10 名 [单位] 份 | 5,100.0000 |

stock_province 表

| 股票代码 | 省份 |
|---|---|
| 600519.SH | 贵州省 |

fund 表

| 指标名称 | 具体数值 |
| --- | --- |
| 基金代码 | 006692.OF |
| 基金简称 | 金信消费升级 A, |
| 任职期限最长的现任基金经理 | 杨仁眉, |
| [名次] 第 1 名","重仓股股票 Wind 代码 | 600519.SH |
| [名次] 第 1 名","重仓股市值占股票投资市值比 | 10.6185 |
| [名次] 第 2 名","重仓股股票 Wind 代码 | 300015.SZ |
| [名次] 第 2 名","重仓股市值占股票投资市值比 | 10.5753, |
| [名次] 第 3 名","重仓股股票 Wind 代码 | 600763.SH |
| [名次] 第 3 名","重仓股市值占股票投资市值比 | 10.4619 |
| [名次] 第 4 名","重仓股股票 Wind 代码 | 300347.SZ, |
| [名次] 第 4 名","重仓股市值占股票投资市值比 | 10.4069 |
| [名次] 第 5 名","重仓股股票 Wind 代码 | 002607.SZ |
| [名次] 第 5 名","重仓股市值占股票投资市值比 | 10.3881 |
| [名次] 第 6 名","重仓股股票 Wind 代码 | 000333.SZ |
| [名次] 第 6 名","重仓股市值占股票投资市值比 | 10.3325 |
| [名次] 第 7 名","重仓股股票 Wind 代码 | 002044.SZ |
| [名次] 第 7 名","重仓股市值占股票投资市值比 | 10.0591 |

| | |
|---|---|
| ［名次］第 8 名","重仓股股票 Wind 代码 | 600276.SH |
| ［名次］第 8 名","重仓股市值占股票投资市值比 | 7.1349 |
| ［名次］第 9 名","重仓股股票 Wind 代码 | 000651.SZ |
| ［名次］第 9 名","重仓股市值占股票投资市值比 | 6.6148 |
| ［名次］第 10 名","重仓股股票 Wind 代码 | 300001.SZ |
| ［名次］第 10 名","重仓股市值占股票投资市值比 | 5.5535 |

stock_chara 表

| 名称 | 解释/算法 | 示例 |
|---|---|---|
| 股票代码 | 股票代码 | 300003.SZ |
| 年振幅 | 一年内的最高收盘价减去最低收盘价<br>/最低收盘价 | 37.3942 |
| 区间换手率 | Σ[单个交易日成交量(股或份)<br>/ 当日股票自由流通股本数 | 336.33 |
| 非系统性风险 | {(Σ(Yi^2)-α*ΣYi-β*Σ(XiYi))/(N-2)}^(0.5) | 0.0497348 |
| 波动率 | {Σ[(Ri-ΣRi/N)^2]/(N-1)}^0.5 | 39.6333 |

## 3.2 数据预处理

### 3.2.1 特殊值处理

在本实验中，有许多股票在从 wind 的数据库中导出时，并没有标明所在的城市。我们查询了其总部所在地，手动标明了其所在城市。

### 3.2.2 缺失值处理

对于 reduceShare 表中，有些股票在公示阶段并没有给出减持幅度。本文的解决方案是，只对持仓变化公示的股票进行排序，选取其中减持幅度最大的股票。

例如，对于 0005.HK，实验开展过程中，并未在数据库中得到其总部城市地址，经过查询，手动设置其总部为“香港特别行政区”。

### 3.2.3 重复值处理

对于基金而言，很有可能会存在同一个基金前缀，但是后缀是 A，B，C 的情况、或者不同基金名称，在此情况下，基金经理是相同的。对于本实验而言，要研究基金经理在一个特定城市的信息的精确度。因而把他看作是不同的基金是不合理的。有两种解决办法：（1）将其合并成一个基金，作为一个基金处理。（2）只选取一个基金对其进行研究。本文采取的是第二种解决办法。本文对一个基金经理名下，不同的基金，只选取其中一个具有代表性的进行研究。

### 3.2.4 数据标准化

本研究，进行数据标准化处理，将实验数据均归于一个指定区间。这样可以去除数据绝对大小对实验的影响，同时去除不同单位的影响。在本研究中，对基金经理的减持指标进行了标准化处理，使之全部都介于 0 到 1 之间。

通过标准化处理，我们可以很好地消除原来的数据中单位和数量级的对模型预测能力的影响。使得模型的预测精确度提升。

## 3.3 数据处理逻辑

首先，由于实验的样本量较小，因而本文选取了十折交叉验证。

其次，对于所有的基金经理，我们选取了减持最大的股票的代码，以及其所对应的减持份额，通过其减持份额乘以其平均价格，在对应其基金净值，得到了减持幅度属性。最后生成了[基金经理，减持最大股票代码，减持幅度]数据表。

接着，我们将实验分成了对照组和实验组。对照组只有股票的属性，实验组还加入了减持指标。其中实验精度提升百分比计算逻辑为：

（对照组误差-实验组）/实验组

最后，对于市场收益率指标，我们选取了两个市场的指数收益率作为代替，A 股市场选取万得全 A，香港市场选取恒生综指：

表一 指数信息表

| 指数名称 | 指数代码 |
| --- | --- |
| 恒生综指 | HSCI.HI |
| 万得全 A | 881001.WI |

## 3.4 描述性统计

本章节中，我们对实验所需要的实验数据进行了描述性统计，我们计算了岩本的均值、方差、最大值和最小值。

首先是市场收益率的描述性统计，这里我们使用两个指数的收益率作为代替。

表一 指数数据描述性统计

| | 万得全 A | 恒生综指 |
| --- | --- | --- |

| | | |
|---|---|---|
| 计数 | 331 | 331 |
| 最大值 | 5.6213 | 5.5112 |
| 最小值 | -8.1583 | -4.9381 |
| 标准差 | 1.4426 | 1.2603 |

表二 基金经理、股票和城市描述性统计

| 变量名称 | 数量 |
|---|---|
| 基金经理 | 99 |
| 城市 | 29 |
| 股票 | 99 |

表三 股票控制变量描述性统计

| 变量名称 | 变量代码 | 数量 | 最大值 | 最小值 | 均值 | 标准差 |
|---|---|---|---|---|---|---|
| 年振幅 | x1 | 99 | 170.0300 | 15.3046 | 47.7153 | 24.59426 |
| 区间换手率 | x2 | 99 | 1899.6569 | 14.6128 | 352.4197 | 302.3727 |
| 非系统性风险 | x3 | 99 | 0.1454 | 0.0151 | 0.0475 | 0.0205 |
| 波动率 | x4 | 99 | 103.7768 | 13.6112 | 39.3523 | 13.8744 |

表四 duvol 指标描述性统计

| 指标名称 | 数值 |
| --- | --- |
| 最大值 | 7.3331 |
| 最小值 | -120.5346 |
| 均值 | -37.9436 |
| 标准差 | 23.7336 |

表五 input0 表描述性统计

| 指标 | 释义 | 最大值 | 最小值 | 均值 | 标准差 | 计数 |
| --- | --- | --- | --- | --- | --- | --- |
| x1 | 年振幅 | 170.0300 | 15.3046 | 47.7153 | 24.59426 | 99 |
| x2 | 区间换手率 | 1899.6569 | 14.6128 | 352.4197 | 302.3727 | 99 |
| x3 | 非系统性风险 | 0.1454 | 0.0151 | 0.0475 | 0.0205 | 99 |
| x4 | 波动率 | 103.7768 | 13.6112 | 39.3523 | 13.8744 | 99 |
| shock | 崩盘指标 | 7.3331 | -120.5346 | -37.9436 | 23.7336 | 99 |

表六 input1 表描述性统计

| 指标 | 释义 | 最大值 | 最小值 | 均值 | 标准差 | 计数 |
| --- | --- | --- | --- | --- | --- | --- |
| reduceIndex | 减持指标 | 648.6880 | 0.1982 | 79.3581 | 108.1374 | 99 |
| x1 | 年振幅 | 170.0300 | 15.3046 | 47.7153 | 24.59426 | 99 |
| x2 | 区间换手率 | 1899.6569 | 14.6128 | 352.4197 | 302.3727 | 99 |

| | | | | | | |
|---|---|---|---|---|---|---|
| x3 | 非系统性风险 | 0.1454 | 0.0151 | 0.0475 | 0.0205 | 99 |
| x4 | 波动率 | 103.7768 | 13.6112 | 39.3523 | 13.8744 | 99 |
| shock | 崩盘指标 | 7.3331 | -120.5346 | -37.9436 | 23.7336 | 99 |

对于表一，可以看出，所选的指标基本已经覆盖所有的省份。对于

对于表五，其除了 reduceIndex 指标其余与表五相同，在一般情况下，白哦六的信息包含表五，可以用作某种替代。

对于表六，崩盘指标是越小越好，一般都会为负数。但是考虑到，有些股票并不一定会有崩盘倾向，所以也会有正值存在的可能性。

# 4 实证分析

## 4.1 模型评价方法

### 4.1.1 回归误差

我们选取了多个计算回归的函数损失的指标：

| 误差指标 | 英文名称 | 备注 |
| --- | --- | --- |
| 平均绝对误差 | Mean Absolute Error, MAE | y 与 yhat 之间差的均值 |
| 均方误差 | Mean Squared Error, MSE | y 与 yhat 之间差的平方的均值 |
| 均方根误差 | Root Mean Squared Error, RMSE | y 与 yhat 之间差的平方均值的根式 |

他们的具体计算方式如下：

记 m 为样本数量，则

平均绝对误差：

$$\frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i-\hat{y}_i\right)\right| \tag{1.22}$$

均方误差：

$$\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2 \tag{1.23}$$

均方根误差：

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{1.24}$$

对于以上的三个指标，数值越低，说明模型预测得越精准。

## 4.2 假设检验

我们对 input1 表中的 shock 指标和 reduceIndex 指标进行回归，在选取适当数据量后，可以得到回归结果，见下表

$$shock = a + b * reduceIndex \tag{1.25}$$

H0：shock 和 reduceIndex 无相关性

H1：shock 和 reduceIndex 有相关

表三 相关性检验

| | std err | t | P>\|t\| |
|---|---|---|---|
| Intercept | 9.234 | -6.507 | 0.000 |
| reduceIndex | 0.077 | 1.902 | 0.099 |

在 0.9 的置信水平下，拒绝原假设，认为 shock 变量与 reduceIndex 有相关性。

## 4.3 模型结果

### 4.3.1 基于 BP 神经网络的股价崩盘风险预测

基于 BP 神经网络的股价崩盘风险预测的预测结果如下表所示。

表一 基于 BP 神经网络的股价崩盘风险预测结果

| 模型 | MAE | MSE | RMSE |
|---|---|---|---|
| 对照组 | 11.5130 | 286.5441 | 16.9276 |
| 实验组 | 9.2503 | 175.1463 | 13.2342 |
| 准确度提升百分比 | 24.5 | 63.6 | 27.9 |

从表中可以看出，BP 神经网络的对照组的实验的均方误差为 286.5441，而实验组的均方误差为 175.1463.预测准确的提升高达 63.6%。由此可见减持指标的对于预测股价崩盘是有着较好的表现的。不同于线性回归，BP 神经网络模型，可以学习到股价崩盘指标和减持指标之间除了线性相关性之外的非线性相关性。

我们同时可以看到 BP 神经网络的均方误差保持在一个较低水平。在真实的应用场景中，使用 BP 神经网络模型，也可以使用该模型来实现股价崩盘的预测。

### 4.2.2 基于 SVR 的股价崩盘风险预测

在使用 SVM 模型时，我们选择了 RBF 核函数，我们设定了惩罚系数 C=0.02，gamma=defualt。我们将训练得到的 SVR 模型应用于测试集上作预测，其结果如下。此时将训练得到的模型在数据集上做预测，此时，模型平均绝对误差为 15.6.

基于 SVR 的股价崩盘风险预测的预测结果如下表所示。

表二 基于 SVR 的股价崩盘风险预测结果

| 模型 | MAE | MSE | RMSE |
|---|---|---|---|
| 对照组 | 15.57978359011374 | 382.0934414199486 | 19.54721057900458 |
| 实验组 | 15.579782833815282 | 382.0934362466902 | 19.5472104466773 |
| 准确度提升百分比 | 4.9e-6 | 1.4e-6 | 6.8e-7 |

由表二可见，SVR 模型对于预测的提升能力不高。或者说并不能声称，在 SVR 模型下，加入了减持指标和没有加入有什么显著的区别。表明，该模型对于特征

的提取能力并不是很强。在四个模型中属于最弱的层次。

但是与此同时，我们从表中可以看出，SVR 模型的平均绝对误差在四个模型中处于中等水平。模型本身的预测能力不低。

其特征敏感度低也有其特殊的用途。在实际应用中，如果模型的属性中我们并不知道什么属性有着必然的关联，可以使用 SVR 模型，使得模型受其他不相干的变量的干扰下降。

### 4.2.3 基于 CART 的股价崩盘风险预测

CART 分类回归树模型对于股价崩盘的预测结果如表六所示，此时，模型的对照组均方误差为 614，而实验组的均方误差为 499，预测的精准度提升了 22.3%

表三 基于 CART 的股价崩盘风险预测结果

| 模型 | MAE | MSE | RMSE |
|---|---|---|---|
| 对照组 | 20.3645 | 613.6992 | 24.7730 |
| 实验组 | 17.4250 | 499.3726 | 22.3466 |
| 准确度提升百分比 | 16.9 | 22.3 | 10.9 |

由表三可见，CART 模型的预测度，并不是很高，可能是由于只是单个的学习器，对于数据的敏感度很高。可能会存在过拟合的情况，模型的泛化能力不高。但是我们同时也看到加入了减持指标后，模型的预测能力在提升。

### 4.2.4 基于随机森林的股价崩盘风险预测

使用随机森林模型的预测结果如表七所示，此时，模型在训练集上的预测准确率为 0.9989，在测试集上的预测准确率为 0.9995。

表四 基于随机森林的股价崩盘风险预测结果

| 模型 | MAE | MSE | RMSE |
|---|---|---|---|
| 对照组 | 14.7953 | 376.7759 | 19.4107 |
| 实验组 | 13.4600 | 259.4108 | 16.1062 |
| 准确度提升百分比 | 9.9 | 45.2 | 20.5 |

由表四可见，随机森林模型相较于上一个 CART 模型，在各个方面都有着显著的提升。同时泛化能力更强。在加入了减持指标后，模型预测准确度也有着显著提升。因而，在计算能力允许时，相较于当个的树模型，例如 CART，更推荐集成学习模型。

## 4.3 模型比较

在上一个章节的模型结果，我们可以看出，本研究的四个机器学习模型对照组，在加入了减持指标后，预测能力都有了不同程度的提升。

尤其是神经网络模型，其本身的预测能力在四个模型中便是最高的，在加入了减持指标后，更是显著提升了预测能力。其模型预测准确度，远远高于其它模型。

首先，BP 神经网络模型作为一种神经网络模型，具有很强的泛化能力。模型在学习了训练集的样本后，在测试集上有着很优秀的表现。

其次，SVR 模型，对于特征属性并不敏感。可以在无法判断数据集中那些属性有较好的预测效果的时候，使用该模型，来降低无关变量的干扰度。

接着，CART 模型，作为一种树模型，与随机森林的预测结果类似，原因在于，其中的当中的学习器都是树模型。在该模型中，加入了减持指标，模型的预测能力有着一定程度的提升。

另外，随机森林模型，作为一种集成学习，同时还有随机属性选择，具有着良好的泛化能力。在两个基于树模型的学习器中，随机森林模型的表型最好。接近于神经网络模型。值得注意的是，在加入了减持指标后，随机森林模型的预测

准确度提升非常大，仅次于神经网络模型。

最后，在实际应用中，推荐使用神经网络模型和随机森林模型。但是当模型的特征很多，并不知道要选择哪一个特征时，选用 SVR 模型。

# 5 研究结论

预测股价崩盘对于金融市场有着重要的意义和价值，对于稳定经济社会生产活动，有着不可估量的影响。本文利用股价减持指标，通过四种机器学习方法，对股价崩盘进行了预测。最后，本文通过已有的文献和实证结果出发得到了以下的结论。

1. **减持指标可以用于预测股价崩盘风险。**

本文中，通过回归分析，假设检验，证实了减持指标对于预测股价崩盘的有效性。

此外，我们还训练了可以实际用于预测不同输入集的机器学习模型用于预测股价崩盘风险。从实验的结果我们可以看出，在加入了股价崩盘指标后，大多数的机器学习模型的预测准确度，都有了显著的提高。

2. **非线性模型，可挖掘属性之间的更多关系，自动化程度高，适用于大数据集。**

本实验采用的非线性模型，可以学习到除了线性关系之外的非线性关系。同时，利用机器学习模型，可以应对更多的数据集。值得注意的是机器学习模型在实际使用的过程中自动化程度高，处理大数据能力强。有利于用于生产实践。

3. **相同模型对于不同特征的表现不同，对于选择模型需要考虑到数据集的特征属性。**

从本实验可以看出，机器学习模型，对数据的要求很高。本质上，机器学习只是一种处理数据提取特征并进行泛化的一种工具而已。本研究中，SVR 模型对于特征的敏感度不高，在加入了减持指标后，SVR 模型的股价崩盘预测能力，并没有显著的提升。而 BP 神经网络模型对于数据集的特征敏感度更高。

4. **神经网络模型更适用于实际中的预测股价崩盘。**

在本研究中。我们发现神经网络模型更适用了实际生产生活。其预测预测精度高，同时在加入了减持指标后，模型的预测能力也有了显著的改善。

# 6 研究展望

1. 由于本实验是检测基金经理的社交关系是否能很好地反馈股价崩盘。实验的本质目的是探究一个变量对另一个变量的相关性。同时给出一些刻画相关性的模型。因而实验并没有对股价崩盘的参考因素，做出时效性的考量。本文设置的时间界面的长度为一年。来检测指标之间的相关性。同时，由于很多基金经理的大幅减持的行为都发生在一年中相对较短的一段时间内。因而在实际应用中，可以在本实验的基础上，适当地缩短时间截面，来达到预测的效果。但是如何选择一个相对恰当的时间截面是一个深入研究的问题。

2. 可以参考 Hong and Jiang（2019）中的方法，利用股票的自身的属性，来推测基金经理之间的关系。

3. 回归所选取的数据量较小，虽然证实了相关性，但是仅是在较小的数据集上的结果。例如，基金经理所选的个数较少，实验会存在一定误差。未来的研究可以加大数据量，并研究不同数据集对应的较为合适的机器学习模型。

4. 本文中的比较分析，所选取的机器学习模型只有五个，未来研究，可以深入探究其他机器学习模型。

5. 本文所考虑的数据集建立在一年的股票属性指标上。所能提取出的信息较少。如果可以增加数据量，可以以周或者月为一个周期，研究时间序列数据。更高的数据量可以发掘更多的信息，未来也可以更好地提升模型的预测效果。

# 参考文献


[1] 周志华. 机器学习[B]. 2019

[2] 罗真，张宗成，职业忧虑影响基金经理投资行为的经验分析[j]. 财经研究，2004，30（12）：111-120

[3] 张雪峰. 开放式基金投资行为及其影响因素研究[D].天津商业大学,2017.

[4] 申宇, 赵静梅, 何欣. 基金未公开的信息: 隐形交易与投资业绩 [J]. 管理世界,2013 (8): 53-66.

[5] 陈莹,袁建辉,李心丹,肖斌卿.基于计算实验的协同羊群行为与市场波动研究[J].管理科学学报,2010,13(09):119-128.

[6] 楼晓霞. 社交关系网络影响股市崩盘的机理[D].浙江大学,2019.

[7] 郭白滢,李瑾.机构投资者信息共享与股价崩盘风险——基于社会关系网络的分析[J].经济管理,2019,41(07):171-189.

[8] 陈新春,刘阳,罗荣华.机构投资者信息共享会引来黑天鹅吗?——基金信息网络与极端市场风险[J].金融研究,2017(07):140-155.

[9] 谢衷洁,黄香,叶伟彰, 等.人工神经网络及其在金融预报中的应用[J].北京大学学报(自然科学版),2001,37(3):421-425. DOI:10.3321/j.issn:0479-8023.2001.03.021.

[10] Jiang H.，2010，“Institutional Investors，Intangible In-

formation and the Book-to-Market Effect”，Journal of Financial Economics，Vol. 96，Iss.1，pp. 98~126.

[11] Bushee，B. and T. Goodman，2007，“Which Institution-

al Investors Trade Based on Private Information about Earnings and Returns?”，Journal of Accounting Research，Vol. 45，pp. 289~ 321.

[12] Pareek，A.，2011，“Information Networks: Implications

for Mutual Fund Trading Behavior and Stock Returns”，Working Paper.

[13] Ozsoylev，H.，J. Walden，M. D. Yavuz，R. Bildik，2011，

“Investor Networks in the Stock Market”，Working Paper.

[14] Scharfstein, Stein, 1990, “Herd Behavior and Investment”, American Economic Review, Vol.80(3): PP465-479.

[15] Prendergast, 1996, “Stole Impetuous Youngsters and Jaded Old-Times: Acquiring a Reputation for Learning”, Journal of Political Economy, Vol.104(6): PP1105-1134.

[16] Chen, J., Hong, H., & Stein, J. C. Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices[J]. Journal of Financial Economics, 2001, 61(3): 345-381.

[17]Hong, H., Stein, J. C. Differences of Opinion, Short-Sales Constraints, and Market Crashes[J]. Review of Financial Studies, 2003, 16(2): 487–525.

# 致谢

我首先要感谢我们班的创始人李心丹老师和我的指导老师方立兵老师，在选题初期，两位老师就给了我很多建议和帮助，确定毕业设计题目后，他们及时跟进我的研究进度，多次与我沟通交流，并在遇到问题和困惑时第一时间帮我解决。

李心丹老师创办了计算机金融工程实验班，给我一生的轨迹带来了幸运的改变。我在这个班级中学习到了太多的能力。也得到了太多。

另外，我要尤其感谢我的指导老师方立兵老师，他在大学四年中两次教过我的核心课程。为我的职业发展，打下了坚实的学术基础。此外，在这四年中还细心地带领我一篇一篇地读论文，研究论文。教会了我金融市场微观结构和中级微观经济学。他拿出宝贵的私人时间，带领学生学习知识，传授给学生研究问题能力的精神令人敬佩！在研究本课题的过程中，方立兵老师还不断地给我细心的建议，每一个方向性的问题方老师都细心地给出了建议。很多细节我没有注意到的，他也认真地提了出来。

其次，我还要感谢为我的实证研究提供帮助的 Wind 数据库，以及与我并肩作战的研究生同学谢贤，本研究的顺利完成离不开他几个月来无私的奉献与支持。最后，我要特别感谢南京大学计算机与金融工程实验班的全体老师和同学们，感谢大家一路以来的帮助与支持，让我的大学生活更加丰富而难忘！