可交易数字资产市场分析与预测框架的研究

（申请清华大学工程硕士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系

学　　科：计 算 机 应 用 技 术

研　究　生：姚　文　兵

指 导 教 师：徐　恪　教　授

联合指导教师：李　琦　副　教　授

二〇一九年五月

Research on Market Analysis and Forecasting Framework of Transferable Digital Assets

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Yao Wenbing

(Computer Science and Technology)

Thesis Supervisor : Professor Xu Ke

Associate Supervisor : Assistant Professor Li Qi

May, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

随着区块链、智能合约技术地不断发展，数字资产也逐步发展起来。区块链技术带动了一个新的市场---可交易数字资产市场的兴起。由于发展时间短，缺乏相应的公开、完整的数据集，目前对于该市场的研究较少，市场预测方面的研究也极其缺乏且不充分。而数字资产市场的市场容量却在逐年增加，因此对数字资产市场的研究具有重要的理论意义与现实意义。针对可交易数字资产市场的现状，提出了分析与预测框架。围绕分析与预测两大任务，对于框架进行了系统的设计、实验和原型系统的实现。主要贡献如下：

1. 设计了数字资产分析与预测框架。在分析数字资产市场特点的基础之上，提出了分析与预测的整体结构，给出了分析与预测的工作流程、所需功能部件。框架具有兼容性强、可扩展性高、便于设计开发等优点。
2. 提出了对市场进行多角度分析，包括新闻、价格数据以及链上数据。针对不同类型的信息，采用了不同的特征提取及分析方法。对于新闻数据，引入了不同的文本表征方法。针对链上数据，提出了九种典型的特征。对不同类型的数据，提出采用不同的分析方法 (时间序列分析和机器学习方法)。同时，框架便于引入其它类型的数据、分析方法，具有高扩展性。
3. 提出了一个新的文本表征方法。方法针对以往通用文本表征方法在市场预测方面的弱点，充分利用了文本中实体之间的情感信息，将长文本（新闻文章）表示为一个情感图。该方法在分析新闻文章对市场价格波动影响的准确率较高。
4. 实现了原型系统，并对当前最具代表性且市场份额最大的数字资产、加密货币---比特币为具体案例，进行实证研究。针对市场价格从波动性、可预测性及各类因素与市场价格波动的相关性进行了深入的分析，并基于预测结果进行模拟交易以探究数字资产市场的有效性。

关键词：数字资产；时间序列分析；加密货币；机器学习；数据挖掘；市场预测

Abstract

With the continuous development of blockchain and smart contract technology, digital assets have gradually developed. Blockchain technology has driven the rise of a new market, the tradable digital asset market. Due to the short development time and the lack of corresponding open and complete data sets, there are few studies on this market, and the research on market forecasting is extremely lacking and insufficient. However, the market capacity of the digital asset market is increasing year by year. Therefore, the research on the digital asset market has important theoretical and practical significance. Based on the current status of the tradable digital asset market, an analysis and forecasting framework is proposed. Focusing on the two tasks of analysis and prediction, the system was designed and the prototype system was implemented. The main contributions are as follows:

1. Designed a framework for digital asset analysis and forecasting. Based on the analysis of the characteristics of the digital asset market, the structure of the analysis and prediction framework is proposed, and the workflow and required functional components of the analysis and prediction are given. The framework has the advantages of strong compatibility, high scalability, and easy design and development.
2. Proposed a multi-angle analysis of the market, e.g., news articles, blockchain data and price data. Different feature extraction and analysis methods are adopted for different types of information. For news data, different text representation methods have been introduced. Nine typical features are proposed for the data on the chain. Different analysis methods (time series analysis, machine learning) are proposed for different types of data. At the same time, it is easy to introduce other types of data, different analysis methods.
3. Proposed a new text representation method. The method is based on the weakness of the previous general text representation method in market forecasting, making full use of the emotional information between entities in the text, and representing the long text (news article) as a sentiment graph. This method is more accurate in analyzing the impact of news articles on market price fluctuations.
4. Implemented the prototype system and conducted empirical study on the most representative digital asset and cryptocurrency, i.e., Bitcoin. In-depth analysis

of market price from volatility, predictability and the correlation between various factors and market price volatility are presented. And based on the forecast results, we simulated transactions to explore the effectiveness of digital asset market.

Key Words: digital asset; time series analysis; cryptocurrency; machine learning; data mining; market prediction

目 录

主要符号对照表

Digital Asset	数字资产
Time Series Analysis	时间序列分析
Bitcoin	比特币
Ethereum	以太坊
Blockchain	区块链
Cryptocurrency	加密货币
AR	自回归模型
MA	移动平均模型
ARMA	自回归滑动平均模型
ARIMA	差分整合移动平均自回归模型
ARCH	自回归条件异方差模型
GRACH	广义 ARCH 模型

第 1 章　引言

1.1　研究背景及意义

1.1.1　研究背景

区块链是一种分布式账本，最初是比特币的底层技术。它可以用于记录数字信息，比如用户之间的交易信息、用户资产所有权信息等等。在其上可以构建智能合约，以实现自动化的买卖交易。它不依赖于中心化的服务器或机构，可靠性高，并且全天候在线。区块链的兴起与飞速发展为各类数字资产的交易提供了非常大的便利。由于区块链（公有链）去中心化特点带来的鲁棒性，任何用户无论何时何地，只要可以接入互联网，就可以参与链上交易。因此，基于区块链的数字资产交易容易形成自由市场。目前，区块链上最典型的数字资产应用是加密货币。加密货币是一种数字资产、货币，它让用户可以在互联网上安全、匿名的进行交易。近几年来，加密货币吸引了大量的投资者、消费者以及投机者的注意。

2009 年，中本聪提出了第一个去中心化的加密货币：比特币[1]。它采用工作量证明机制 (Proof of Work)，来保证比特币以一个固定的速度被“挖”出来。它的总量固定，任何人都不能随意发行比特币，总量为 2100 万，最初的发行速度为没 10 分钟产生 25 个，随后每四年产生速度减半，直至挖完。比特币的这种机制，可以有效的防止通货膨胀。其底层是区块链技术，通过密码学来保证其交易的安全性。用户进行的交易都会被记录到区块链上去，区块链技术通过共识来解决双花问题，以保证用户交易的安全性。与中心化的数字货币系统和银行系统不同，它的工作方式是去中心化的。底层的区块链就是它存储交易数据的数据库。

比特币[1] 问世至今已经过去了十年。在此期间，加密货币发展迅速，涌现出了一大批其它类型的加密货币。典型的代表有：以太坊[2]、瑞波币[3]、莱特币[4]、达世币[5] 等。它们均在某一方面上对现有的比特币的系统进行了改善，促进了加密货币的发展。以太坊为例，它是第一个将加密货币与智能合约相结合的，它使用股权证明机制 (Proof of Stake) 机制来进行权益（以太币）的分配。在其上，用户可以轻松的开发中各种各样的智能合约来满足不同需求。比如用户可以通过以太坊来轻易地发行自己的加密货币。以太坊在结构上与区块链也有所不同，它改变了区块的大小与生成方式，提高了交易速率。截止 2019 年 4 月 2 日，交易所中的加密货币的数量已经发展到 2139 种，总市场容量已超 1473 亿美元，日交易额超过 490 亿美元。其中比特币的交易总额占到了总交易额的一半以上[6]。全世界的

矿池、交易所的数量也是与日俱增。加密货币的影响力在不断扩大，全世界许多国家都可以自由交易加密货币，如：美国、日本、加拿大、澳大利亚、芬兰等。美国一些机构对利用加密货币进行非法交易出台了一些政策进行约束，但整体上还是对加密货币的态度是积极的。一些主流的企业，比如微软在其在线商店中支持用比特币进行交易[7]。同样加拿大对加密货币的立场也是相对友好的，但也有一些限制。比特币被加拿大税务局认定为是一种商品，比特币交易所获的的收入也是需要交税的；加拿大将加密货币交易视为货币服务业务，因此它受到反洗钱法的约束；在加拿大，一些银行也禁止使用信用卡进行比特币的交易。

市场价格预测一直是一个非常具有挑战性的任务，它对许多研究者来说都非常具有吸引力。目前，仍然没有一个精度较高的通用算法可以精确预测一些商品比如股票的价格。传统的股票预测已经吸引了许多研究者，但是新兴的可交易数字资产市场缺乏深入的研究。数字资产市场与传统的股票市场存在很大的不同之处，比如数字资产等不受中心化机构的控制，比如比特币由全球近万个节点共同维护。并且数字资产如加密货币等可以进行 7*24 小时交易，相比于股票市场而言，数字资产市场更加的自由。

许多因素会引起数字资产的价格波动。一些对数字资产产生消极影响的事件（如 2013 年，Mt.Gox 交易所比特币被盗），可能会引起加密货币价格的大幅下降。数字资产作为一个新兴事物，它的价格还受到社会接受程度的影响。典型的数字资产---加密货币具有隐私保护的特性，使用它进行支付与信用卡支付不同。使用信用卡支付会在银行和商家留下交易记录，而使用加密货币进行交易虽然同样会在区块链上留下记录，但是它们是匿名的。在合理使用的情况下（比如一次一密），很难将加密货币的账户地址与现实生活中人的身份联系到一起。因此，加密货币目前常在非法商品交易或者犯罪活动中使用。由此可以看出，加密货币的价格还受到某些非法活动的影响。数字资产的价格还受到许多其它因素影响，比如公众观点、公众情感、政府政策以及拥有大量比特币的持有者的影响。

1.1.2　研究的意义

数字资产的兴起为投资者们提供了新的投资选择，它们可以为投资者们带来巨大的经济回报。虽然目前数字资产的投资仍然是高风险的，市场的价格波动非常大，但其仍然吸引了一大波投机者的注意。加密货币作为一种最典型的数字资产，它的兴起为实时在线支付、跨境支付提供了一种新的解决方案，可以促进电子商务的发展。从目前的发展趋势来看，数字资产的交易额每天都在不断增长。数字资产市场的发展时间短，从比特币诞生至今只有短短的 10 年时间，因而目前人

们缺乏对该市场的深入分析，缺乏对其价格波动的影响因素的分析。

对数字资产市场的分析的研究有助于帮助人们加深对数字资产的了解，及时了解数字资产的动向，有助于为国家制定数字资产相应的政策提供借鉴。对数字资产市场的预测的研究有助于帮助投资者们了解市场动向，规避市场大规模动荡的风险，从而帮助人们理性投资。同时对数字资产市场的正确预测还可以带来巨大的经济效益。

1.2　国内外研究现状

本节从典型的数字资产---加密货币、市场预测以及其它等方面对当前的研究现状进行综述。

1.2.1　加密货币市场研究

加密货币具有数字金融资产的属性。Dyhrberg[8] 使用 GARCH 模型研究了比特币作为金融资产的能力。在最初的模型中，比特币显示出了和黄金以及美元的相似之处，展现出了作为交易媒介的对冲能力和优势。非对称 GARCH 模型显示：比特币可能有助于投资者进行风险管理，并且是风险厌恶投资者对抗负面市场冲击的理想选择。他们的研究结果表明，比特币由于其去中心化的属性以及其有限的市场大小，它的位置是介于货币与商品之间的。这种分类给投资者带来了更多、更加明智的投资组合的选择。

Gandal 等人[9] 从两方面对加密货币市场中的竞争进行了研究。他们研究了两个时间段内不同加密货币之间的竞争情况。第一阶段从 2013 年 5 月到 2013 年 9 月，第二阶段从 2013 年 10 月到 2014 年 2 月。他们发现：在第一阶段，比特币的受欢迎程度相比于其它加密货币不断增加，而在第二阶段，其它加密货币的受欢迎程度相比于比特币而言，有过之而无不及。主要原因是：比特币打开了加密货币市场，因此占据了“先发优势”。但是，从长期来看，比特币的这种优势是否可以保持还是未知的。他们对交易所之间竞争的分析指出：市场几乎很少或不存在套利的机会。而且从第一阶段到第二阶段，套利的机会正在不断减少，鉴于此，不同交易所之间可能长期共存并且可能在各种费用上产生竞争。

ElBahrawy 等人[10] 研究了加密货币市场的动态演变过程。他们对从 2013 年 4 月到 2017 年 5 月之间的加密货币市场进行了调研。调研指出：从 2016 年 4 月开始，加密货币市场容量进入了指数增长阶段，而比特币的市场份额正稳定减少。从他们研究的时间段开始，加密货币的种类、市场份额分布以及排名变化都较为稳定。他们从生态学的角度出发，指出进化的自然选择模型可以解释市场中一些可

以观察到的属性。

Cocco 等人[11] 对比特币市场的交易进行了建模。该模型精确地反应了真实市场的许多特性。它对不同的交易策略、符合实际的订单交易规则、和实际一致的比特币产生规律、新交易者的到来以及用户的财富分布等均进行了建模。模型给出的结果很符合比特币实际的价格规律。他们的计算结果表明他们无法拒绝比特币价格变化符合随机游走规律的假设。同时，实验结果还表明波动聚类现象的存在。模型中的绝对收益和现实中的一样，都呈现出了厚尾 (Fat Tail) 现象。

Caporale 等人[12] 使用 R/S 分析技术分析了四种加密货币 (比特币、莱特币、瑞波币、达世币) 的持久性及其随时间演化的情况。研究结果表明：加密货币市场仍然是一个非有效市场，但是随着时间的推移，这种非有效性正在逐渐减弱。该现象在莱特币市场上表现的尤为明显。加密货币市场的这种持久性表明我们可以通过某种交易策略来获取额外的收益。Sashikanta 等人[13] 对比特币市场回报的可预测性以及适应性市场假说分别进行了评估。他们发现比特币市场的效率是随着时间变化的，并且在比特币的市场中验证了适应性市场假说。Wang[14] 对加密货币市场中的流动性以及市场有效性进行了研究，他们检验了 456 种加密货币的流动性，结果表明具有高流动性的加密货币回报的预测性消失了。研究指出比特币市场中有效性是存在的，但是大量的其它种类的加密货币仍然展现出自相关性和非独立性。研究还表明流动程度对市场有效性和新的加密货币回报的可预测性有很大影响。

加密货币市场中，用户的行为对市场的走向有着直接的影响。不同于传统的股票市场，加密货币市场中的交易更加自由。Krafft 等人[15] 对加密货币市场中的用户行为进行了分析。他们采用了实际在线实验的方法，在交易所中进行了实时交易，从而从交易行为角度去分析投机者的行为对其他交易者的影响 (peer influence)。他们设计了交易机器人，在为期 6 个月的时间内，对 217 中加密货币进行了总数超过 10 万笔的交易（每笔交易的费用低于 1 美分）。他们发现个人买入的行为会导致接下来短期内相对于干涉量数百倍的买方效应。他们发现交易所的设计策略加强了这种影响。

Bouri[16] 等人研究了主要的加密货币交易量与回报、价格波动之间的关系。他们研究了 7 种主要的加密货币（比特币、瑞波币、以太币、莱特币、新经币、达世币和恒星币）。研究发现：交易量是所研究加密货币的极端负回报和正回报的格兰杰原因。但在波动较小的情况下，交易量仅是其中三种加密货币（莱特币、新经币、达世币）价格波动的格兰杰原因，而在 GARCH 模型下，这种因果关系并不存在。

Caporale 等人[17] 研究了周内效应对加密货币市场的影响，他们所使用的方法主要有：平均分析、学生 t 检验、方差分析、Kruskal−Wallis 检验、虚拟变量的回归分析以及交易模拟方法。他们发现，周内效应对比特币之外的加密货币都没有影响。比特币在周一的回报率要明显高于一周内的其它时间段。在这种情况下的交易模拟实验表明市场中存在可利用的获利机会。但是，实验结果大多数与随机生成的没有大的差别，因此，周内效应并不能成为证明市场无效的证据。

Bouri 等人[18] 对不同加密货币之间价格剧烈变化之间的关系进行了研究。他们发现一个加密货币价格的剧烈波动一般依赖于其它加密货币价格的剧烈波动，并且这种联合爆发性和加密货币的大小关系不大。

1.2.2　市场预测研究

Alessandretti 等人[19] 利用机器学习方法来预测加密货币的相对价格，以利用市场的非有效性，从而获得收益。他们分析了从 2015 年 11 月到 2018 年 4 月间的 1681 类加密货币的每日数据。他们的研究结果表明通过简单的交易策略在机器学习算法的辅助下，收益可以超过基准。实验结果表明一个重要但是简单的算法可以帮助对短期内加密货币市场进行预测。他们主要使用了两种机器学习算法：梯度提升决策树和 LSTM（长短期记忆），其中 LSTM 的表现最好。在实验中，他们用比特币作为单位而不是常用的法币（如人民币、美元等）。实验结果表明使用比特币作为基准来预测所得到的收益比使用法币要高。

Catania 等人[20] 在他们的论文中研究了四种加密货币（比特币、以太币、莱特币以及瑞波币）的条件波动性。他们研究了波动过程中长记忆的计算效应及其对过去序列的非对称反应，以预测波动率。他们的研究表明包含杠杆和时变偏度的更加复杂的波动模型可以提高波动预测的准确度。

情感因素在股票市场预测中的研究非常广泛。Bukovina 等人[21] 研究了情感因素与比特币波动性之间的影响。他们的研究表明比特币的吸引力和对比特币的投机性投资对比特币价格波动会产生影响。该研究指出了情感因素与比特币价格波动之间关系的经济因素。同时，作者们还提出一种独特方法对比特币的价格进行了分解，将价格分解为了理性部分和非理性部分。他们使用了 2013 年 12 月 12 日到 2015 年 12 月 31 日间的情感数据对他们提出的理论进行了测试。他们发现情感因素对比特币价格的波动影响只占一小部分。在价格波动比较大的情况下，情感因素的解释性更强，特别是在 2013 年末到 2014 年初之间的泡沫阶段。研究还发现积极因素对比特币的价格波动性影响比消极因素更大。在所提出的价格分解模型下，他们通过对 2015 年的数据进行分析，发现来自实体经济的投机性投资需求

减少了，比特币的价格变动的主要驱动因素为供求关系。

在过去的 10 年里，博客、微博、聊天等 web2.0 服务发展迅速，这加速了人与人之间的交流。人的情绪也很容易通过这些渠道被很快地传播出去，这些情绪对人的决策和个人行为等可能有很大的影响。Matta 等人[22] 的论文研究了比特币的价格与推文的数量以及网络搜索之间的关系，以说明社会媒体活动与一些搜索信息对专业的比特币投资的影响。他们研究了比特币的价格趋势与谷歌趋势（Google Trends）以及推文（特别是那些带有积极情感的推文）数量的关系，发现了它们之间的互相关性较大。他们搜集了从 2015 年 1 月到 2015 年 3 月之间 60 天的 1,924,891 条推文，并且对这些推文进行了情感分析，以探究这些情感信息是否对比特币市场的预测有帮助。Stenqvist 等人[23] 通过分析 227 万比特币相关的推文的情感变化，并以此作为判断未来一段时间内比特币价格涨跌的依据。他们首先计算一定时间（从 5 分钟到 4 小时）内的推文的情感信息的累计和，这个累计值被用于推断 1 到 4 倍的时间间隔后比特币价格的涨跌。他们的实验结果表明利用 30 分钟内的推文的累计情感值来预测 4 倍时间间隔（2 小时）后的价格波动的准确率可以达到 79%。

在价格预测方面，Shah 等人[24] 讨论了如何利用贝叶斯回归方法预测比特币价格。同时他们设计了一个简单的比特币交易策略。在他们的交易策略下，投资在 60 天内可以获得一倍的收益。Madan 等人[25] 使用机器学习方法来预测比特币的价格。预测分为两阶段。第一阶段研究了比特币市场的每日趋势，并且探索影响比特币价格的最优特征。数据集包含了 5 年内的数据，每条数据有 25 个特征。这些特征对预测比特币价格的涨跌取得了 98.7% 的准确率。第二阶段研究了比特币在不同时间间隔内的价格数据与杠杆数据。价格预测问题被建模为一个二分类问题，并且使用了随机森林算法与广义线性模型进行了实验。在 10 分钟的时间间隔上，实验的准确率在 50% 到 55% 之间。Madan 等人[25] 提出了用基于 MLP 的非线性自回归模型来预测比特币的价格。模型使用了开盘价格、收盘价格、最低价格和最高价格以及这些价格的移动平均来作为输入，同时使用了离子群优化算法来优化模型的参数。Jang[26] 等人使用区块链上的信息以及宏观经济变量来预测比特币的价格以及价格波动程度。论文通过贝叶斯神经网络来分析比特币的时间序列信息。他们的实证研究表明贝叶斯神经网络在预测比特币价格时间序列方面表现很好，并且可以很好地解释比特币价格的大波动性。McNally 等人[27] 实现了基于贝叶斯优化的循环神经网络以及长短期记忆网络来预测比特币的价格涨跌。他们利用从 CoinDesk[28] 获得的比特币价格数据作为输入，对比了深度学习方法和 ARIMA 算法。实验结果表明 LSTM 获得了最优的准确率 52% 以及最优的 RMSE。同时该研究对比了在有 GPU 加速和没有 GPU 加速的情况下模型的效率，结果显

示在有 GPU 加速的情况下训练时间减少了 67.7%。Almeida 等人[29] 基于前几天的价格和交易量数据，使用人工神经网络（ANN）来预测后一天的价格趋势。所设计的表现最好的网络在两年的时间内获得了超过趋势跟随者的收益。他们还发现基于他们的模型，在加入交易量数据之后，预测准确率并没有显著的提升，这说明了交易量对价格的影响不大。

加密货币上的交易是记录在区块链上的，而区块链上的信息是公开的，在区块链上很容易跟踪每一个币的来源和目的地址。Greaves 等人[30] 使用区块链上提取出的特征以及比特币的价格来预测短期内比特币价格的涨跌，他们使用了机器学习的方法并且取得了 55% 的准确率。Akcora 等人[31] 利用区块链上的交易信息来预测比特币的价格。比特币的交易可以构成一个非常大的图，其中的结点为账户地址，边为交易额。在其研究中，一段时间内的交易构成了一个子图，这个子图经过处理之后被用来评估其在比特币价格形成中的作用。子图按照的拓扑结构被分为了许多类型，他们发现其中某些类型对比特币价格具有很大的预测作用。

1.2.3　其它

Jiang 等人[32] 使用深度强化学习研究了不同加密货币的投资组合。他们以不同的金融资产的历史价格数据作为卷积神经网络的输入，然后输出不同金融资产的投资组合权重。他们以强化学习的方式训练模型，以累计收益作为强化学习的激励值。在他们的回溯测试中，他们在不到两个月的交易间隔内，实现了 10 倍的回报。

Hayes[33] 研究了加密货币价格形成的可能决定性因素。研究通过具有代表性的经验数据检查了 66 种广泛使用的加密货币。他们的估计回归模型指出加密货币价值形成的三个主要的驱动因素为：生产者网络中的竞争水平、单位生产率和加密货币的挖矿难度。不同数字货币之间生产成本的差异表明了加密货币的相对消耗（以电力作为输入，加密货币为输出）。研究以此为起点，为加密货币建立无套利的情况，并将成本模型正式化，并且通过这个模型来确定加密货币的价值。

Sovbetov[34] 基于五种加密货币（比特币、以太币、达世币、莱特币和门罗币）在 2010 年到 2018 年间每周的数据分析它们的影响因素。该项研究采用了 ARDL 技术，发现了市场风险指数、交易量以及波动性对所研究的加密货币的价格在长期和短期内都是重要的决定因素。在长期来看，加密货币的吸引力对决定价格也很重要。另外，SP500 对比特币、以太币以及莱特币价格的长期正面影响较弱。

1.3　研究的主要内容及方法

1.3.1　论文结构

第一章介绍了研究的背景及意义，而后从加密货币市场以及市场预测两大方面调研了国内外的研究现状，并且对加密货币市场其它一些方面也进行了一些调研，从而引出了主要主要研究内容、研究框架、研究方法。

第二章分析了论文研究的关键研究点。本文的两大研究点一是市场的分析，二是基于市场的分析对市场进行预测。围绕这两大研究点，分解出了几个小的研究点。其中，市场分析主要分析市场影响因素以及对这些影响因素的影响力进行量化分析。市场预测主要研究：如何从不同类型的数据集中提取出有效的特征、不同预测模型的预测能力以及不同预测模型的收益状况。

第三章介绍我们提出的分析预测框架，介绍了框架中不同的模块，以及框架的工作流程。同时还对论文所使用的关键算法进行介绍。算法主要包含三个方面：第一是时间序列分析算法。时间序列分析是一种常用的数据分析技术，它对按时间顺序排列的数据进行分析，在市场预测方面常被用到。第二类是机器学习算法。算法主要用于分析新闻文本对市场价格涨跌的影响并且对市场价格涨跌进行预测。第三类算法为提取特征的相关算法。主要介绍了论文中使用的文本信息特征提取的方法。

第四章利用所提出的框架进行实验，给出在不同模型、算法下，对市场的分析、预测的结果，并且根据预测结果，进行模拟交易以分析不同算法、模型的收益。

第五章总结本论文的内容。

1.3.2　研究内容

本文主要以加密货币为例，研究可以交易数字资产市场。主要的研究点为：

- 数字资产市场分析与预测框架设计的研究；
- 数字资产市场的影响因素及市场价格波动特点的研究；
- 基于市场预测任务的新闻文本数据表征方法的研究；
- 不同算法、模型在数字资产市场的预测方面能力的研究。

1.4　论文创新点及不足

1.4.1　论文创新点

1. 研究设计了数字资产市场分析与预测框架；

2. 系统地量化分析了各种因素对加密货币市场的影响；
3. 对比分析了传统预测方法与基于机器学习方法的性能；
4. 提出了一种新的基于情感分析技术的新闻文章特征提取的框架；
5. 实现了本文中提出的预测系统，并进行了实际的线上交易实验。

1.4.2　论文不足

1. 一些影响数字资产市场的因素由于无法获取数据而无法包含在我们的实际实验中（比如加密货币在暗网中的具体使用[35-36] 数据）；
2. 本文的交易收益实验为模拟实验，模拟结果与实际结果存在一定的偏差。在实际环境中，市场对交易会产生一定的反馈，从而可能对价格产生一定的影响。

第 2 章　关键研究点分析

本节从最典型的数字资产---加密货币出发，对下面几个关键研究点进行分析：

1. 市场价格波动的特点；
2. 市场的影响因素；
3. 市场相关信息（如交易量、交易费、交易延时等等）与价格波动的相关性；
4. 如何从新闻文本信息中提取出对市场价格波动预测有利的特征。

2.1　市场价格波动的特点

新兴的加密货币的市场影响因素相较于传统的股票要多。它更加的开放，可以进行全天候不间断的交易，因此其价格波动应该比传统股票市场更大。这种市场价格的高波动性对于不同的投资者来说有着不同的影响，对价格波动进行分析具有十分重要的指导意义。市场价格的历史数据会对未来的市场有影响，目前已经有许多做量化交易的公司，利用历史价格信息来预测市场未来的走向，因此对市场价格的分析也具有十分重要的现实意义。

市场价格数据是一个典型的时间序列，以往的研究中已经有许多时间序列分析的方法。本文中采用了其中几种典型的分析模型来对价格序列进行分析。这些模型包括 AR、MA、ARMA、ARIMA、ARCH 以及 GRACH 模型。其中 AR、MA、ARMA 及 ARIMA 模型主要分析历史价格数据对未来价格的影响，而 ARCH 和 GARCH 模型主要对市场价格波动性进行建模，以得出未来市场价格波动的特点。

2.2　市场影响因素分析

加密货币作为一种可交易数字资产，它的价格受到许多因素的影响。第一，加密货币的交易是匿名的，它不便于监管、用途广泛，还被用于进行非法交易[37]，各种利用到加密货币进行的交易都有可能影响到市场价格。第二，各国对加密货币的态度不一，制定的政策也不一样。而政府的政策很大程度上影响了加密货币的发展，进而影响到市场。第三，随着时间的推移，加密货币的需求量也在不断的变大。加密货币价格的波动性较于传统的金融产品如：股票、债券等更大。以比特币为例，它的价格在 2013 年时才一百多美元，而在四年多之后它的价格在其顶峰阶段接近 20000 美元，投资回报率如此之高，吸引了一大批偏好高风险的投资者、套利者的关注。另一方面，加密货币由于其隐私保护的特性，在合理使用的情况下

极难被追踪溯源，因此它在暗网中逐步流行起来[35-36]，需求量也逐渐变大，在许多非法交易[37] 中也被使用着。需求量的不断增大，巨额的投资回报也催生了大批的、各种各样的加密货币，这些加密货币之间彼此竞争，在一定程度上形成了一种平衡。这些因素同样会影响到市场。第四，一些重要事件的发生也会对市场产生重大的影响，比如 MtGox 交易所比特币被盗事件就导致了比特币价格的大跌。公众的态度对市场也有着重要的影响。其它一些对股票有影响的因素也可能对加密货币的价格产生影响比如：微博[38]、谷歌搜索指数[39]、新闻[40] 等。

2.2.1　政府政策

加密货币管理调查报告[41] 对不同国家和地区的加密货币的政策进行了深入的调研。报告指出：加密货币在不同的国家的称呼不一样，这从不同程度上反映出各国、地区的态度。如在中国、加拿大等地区，加密货币被成为虚拟商品，在澳大利亚、泰国被称为数字货币，在德国被称为加密令牌，在墨西哥被称为虚拟资产等。政府经常发出对加密货币市场投资陷阱的警告。这些警告主要有中央银行发布，用于让人们了解法币与加密货币之间的区别。大多数警告都会关注加密货币市场价格的高波动性以及交易组织不受监管而导致的高风险性。并且人们在投资加密货币时，必须由个人承担其中的风险，无法从法律途径来获取受损利益的补偿。许多警告也会指出加密货币为洗钱以及恐怖主义等非法活动提供了机会。一些国家的政策不仅限于发布警告，同时也会颁布一些法律来规范加密货币的使用与监管。

一些地区（如：摩洛哥、尼泊尔等）对加密货币的投资进行了限制或者禁止涉及加密货币的任何活动。一些地区（如孟加拉、泰国等）还会禁止其境内金融机构进行加密货币交易。在一些地方，初始代币发行（ICO）也是被完全禁止的，但是大多数地区趋向于关注 ICO。一些地区对不同的 ICO 方式会有不同的对待方式。

并非所有的国家将区块链和加密货币技术视为一种威胁，它们注意到了加密货币背后的技术潜力，并且也在积极设置加密货币友好的监管机制，以作为吸引优秀的科技公司的手段。

在一些地区，加密货币还被视为支付手段。在瑞士的某些地方，甚至连政府机构都可以使用加密货币进行支付。墨西哥也允许使用加密货币作为支付手段和本国货币，安提瓜和巴布达政府允许通过政府支持的 ICO 为项目和慈善机构提供资金。

2.2.2　需求分析

吴[42] 从供求均衡方面以及网络经济学、货币经济学角度出发，利用对数平均分解法，分析了虚拟货币价格的形成机制。该研究将比特币的价格分解为多个因子，这些因子包含了经济发展、比特币功能因子、网络经济等相关元素。研究发现比特币功能因子对价格的影响最大，而比特币的热度因子的影响力最小，而且还在逐年减小。

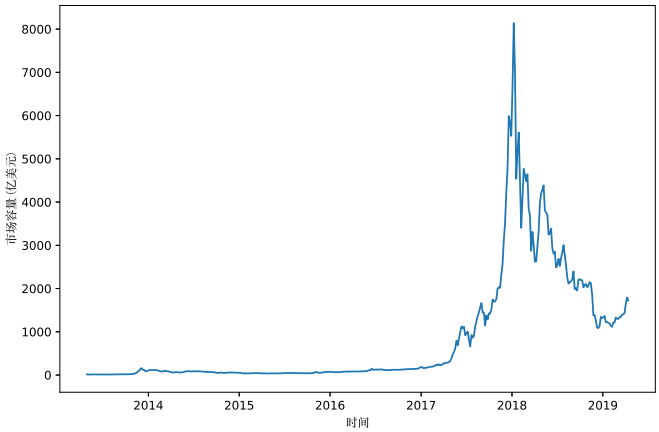

图 2.1　加密货币市场容量变化情况

加密货币的需求主要来自两大方面：

- **投资** 加密货币出现之后，形成了一个规模非常大的市场。许多 ICO 项目的成功吸引了大量的投资机构以及投资者的关注。交易所每天都在引入新的加密货币。现在加密货币的市场规模已经非常大，如图2.1所示。在加密货币市场最高峰，市场容量甚至超过了 8000 亿美元。Glaser 等人对比特币进行过一项研究，他们的研究发现新的用户更倾向于将比特币视为一种投机（投资）手段，而对其是否可以购买其它商品或者服务并不非常关心。
- **交易** 加密货币经过 10 年的发展，现在已初具规模，特性也逐渐被人们所了解，开始逐步走向成熟阶段。币价也由暴涨暴跌走向相对平稳，用途也逐渐发展到多个方面。人们可以通过加密货币进行捐赠活动、商品买卖、支付劳动报酬等等一些列活动。而且当前与愿意支持加密货币交易的公司企业、商家也越来越多，支持加密货币的国家也在逐年增加，加密货币的种类越来越多。各种改进版本的加密货币不断出现，提高了加密货币的质量，提升了用户服务水平。

2.2.3　关键事件

加密货币的市场前景目前仍然不明确，用户一般会通过他们一切可以接触到的渠道来获取加密货币相关信息，以判断加密货币未来的市场走向。信息的来源多是社交媒体、网络新闻、报纸、互联网社区等等。由于加密货币的价格缺乏一个固定的价值背书机构，这些信息来源对人们的影响将直接反映到加密货币市场，影响加密货币的价格。

比如如果某个加密货币的底层协议出现了安全性漏洞，这些信息将通过社交媒体、网络飞速传播，这必然会影响币持有者对该货币的信心。用户必然会重新评估该加密货币的价值，从而可能有大面积抛售该类加密货币的情况出现。目前加密货币的交易许多都是通过交易所进行的，用户可能对加密货币的安全性、匿名性以及可以很容易进行双边交易的特性并不感兴趣，从而导致关于交易所重大的积极或者消极的事件直接影响到加密货币的市场（如 Mt.Gox 事件）。Mt.Gox 交易所曾今是世界上最大的比特币交易所，每天处理了全球至少 70% 的比特币交易。2014 年 2 月 7 日，Mt.Gox 交易所停止了所有比特币的交易。2014 年 2 月 10 日，Mt.Gox 在所发表的申明中表示，它们的比特币交易系统出现了漏洞，这个漏洞让实际已经进行的交易在交易所看来却没有发生，导致了大量比特币的丢失。2014 年 2 月 28 日，Mt.Gox 提交了破产保护，Mt.Gox 的负债高达 65 亿日元。后来的调查结果显示，有 75 万枚用户的比特币与 10 万枚 Mt.Gox 自有比特币被盗了。该事件对比特币市场形成了巨大的冲击，导致了比特币价格的暴跌。

2.3　市场相关数据分析

加密货币的市场价格与其它一些数据（如链上数据）也存在着相关性。本文探究了从区块链上获取的可能与市场相关的几类数据与市场价格之间的相关性。这些数据类别见表4.1。

2.4　特征工程

在我们搜集的数据集中，既包含了时间序列的数据，也包含了文本的数据。这些数据都具有实际的、人类可以理解的含义。与此同时，这些数据可能对机器并不是友好的，比如说价格数据，价格数据是一个随着时间变化的时间序列。大多数时候将这些数据直接作为机器学习的模型的输入效果并不好。因此，我们需要对数据集通过各种方式进行特征增强。比如可以通过计算一定时间内的价格数据的方差来反映最近一段时间内的市场状况，通常情况下这种波动性对市场有着很重

要的影响。再比如我们搜集的的数据可能不再一个数量级上，有的数值很小（如在0~1之间），有的很大（如在0~100000之间）等。数据的这种差异性会增加后续模型学习的难度。文本数据通常情况下不能直接用于机器学习、深度学习模型的输入。一般情况下，我们会根据任务的特点，使用文本特征提取的方法，从纯文本信息中提取出可以代表原文本，并且可以反映学习目标特点、方便机器处理的特征。

在对数据集进行特征提取之后，所得到的特征集的维度可能会很大。有些情况下，特征的数量可能比样本自身的数量还要多得多。特征数量太多会导致学习器的搜索空间变大、学习难度增加，而且过拟合的风险也大大增加了。在此情况下，一般会通过一些方法降低特征的维度。有两种思路可以降低特征的维度：一、特征选择；二、降维。特征选择指从许多特征中，根据特征的重要性来选择那些比较重要的特征，而去除那些次重要的特征。降维主要指通过如 PCA、autoencoder 等方法，将高维空间的数据映射到低维空间去。

2.4.1　特征提取

2.4.1.1　数值数据

对于数值类型的数据，我们会根据数据的类型进行分析，利用特征增强技术，提取出对于分析预测作用更大的特征。

2.4.1.2　文本数据

文本数据中包含大量人类所能理解的复杂信息，这些复杂信息对人们的决策起着指导的作用。但是文本数据不能直接作为一些学习器（如：神经网络）的输入。常用的文本数据表示的模型为向量空间模型，该模型分为两个阶段：一、确定语义单元，二、确定语义单元的权重值。常用的语意单元可以是词语、短语或者 n-gram 单元等。权重值可以是语意单元的出现频率或者其它特征。该模型典型的例子如：n-gram、tf-idf。当前比较流行的方法还有基于文本嵌入 (text embedding) 的方法，比如 doc2vec[43] 方法等。

本文提出了一种新的文本表征方法，该方法利用了自然语言处理中的情感分析技术。我们将在算法介绍章节中详细介绍。

2.4.2　特征选择与降维

数据集经过特征提取之后，产生了大量的特征。这些特征有的对市场分析预测有重要的作用，有的作用微乎其微，甚至没有任何作用。因此，特征选择算法以

及降维算法被应用以降低特征维度，提升效率以及提高模型的范化能力。本文中主要使用 PCA 降维的方法对从新闻文本信息中提取出的特征进行降维。

第 3 章　分析与预测框架设计

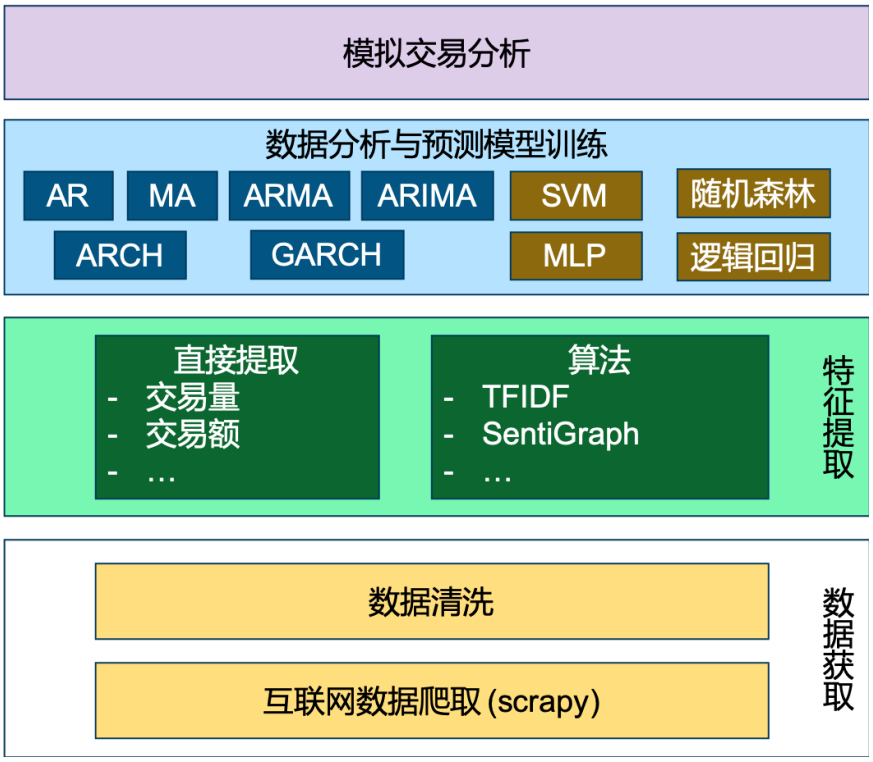

图 3.1　分析与预测框架整体结构图

框架主要分为 4 层，如图3.1所示。框架通过数据获取层从互联网上获取各类数据，并且对数据进行初步的清洗，以去除噪音数据。处理好的数据被传递给特征提取层，不同类型的数据会有不同的特征提取方法。提取出特征之后，数据分析与预测模型训练层对提取出的特征数据进行进一步的深入分析，以分析得到市场的特点。该层还通过机器学习模型的学习对市场进行预测。模拟交易层对于分析预测的结果进行模拟交易以分析对比各类算法、模型的实际性能。

具体处理流程见图3.2。

框架3.2中虚线的左半部分用于对数据进行分析、对机器学习模型进行训练，右半部分用于对市场进行预测。

数据采集层利用网络爬虫等技术从互联网上相关的站点爬取数据。在收集到数据之后，首先会进行数据清洗工作，对数据中的一些无用数据或者噪音数据进行去除。清洗过后的数据，一部分给数据加工模块通过算法进行分析。分析结果经过可视化处理之后，将结果传给结果输出模块，以呈现给用户。另一部分数据

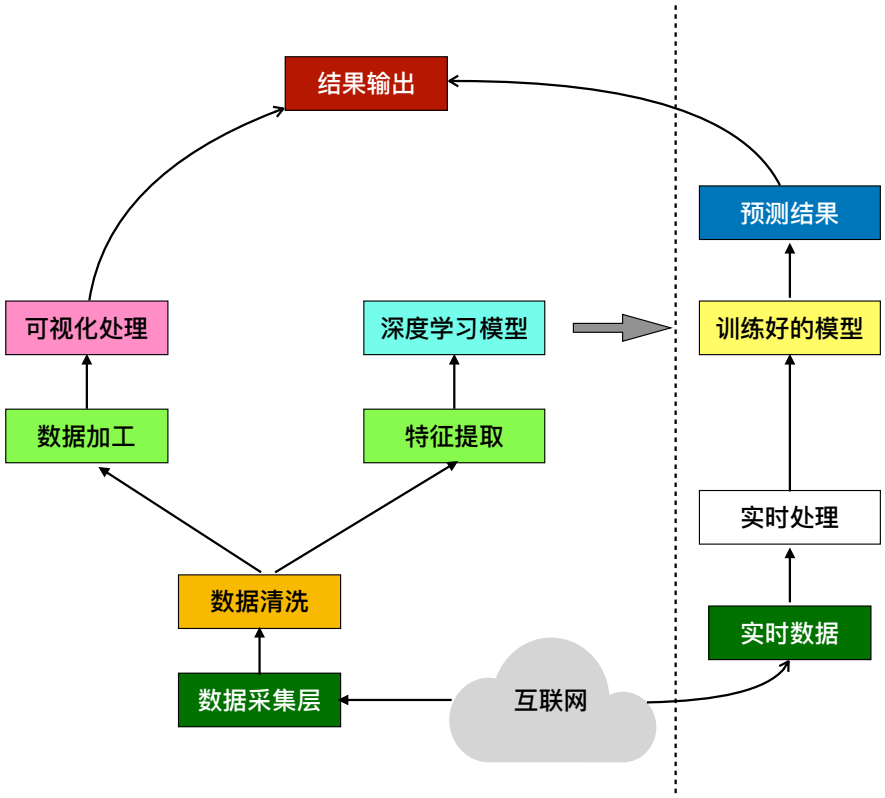

图 3.2　分析与预测框处理流程图

传给特征提取模块进行特征提取。特征提取模块进行特征工程相关的工作，包含特征提取、特征选择及降维等操作。特征提取模块输出的样本被作为机器学习模型的输入，以进行模型的训练。实时数据会经过处理之后传递给训练好的机器学习模型，从而得到预测结果。预测结果也会通过用户交互界面呈现给用户。

数据处理及模型训练的基本流程如图3.3所示。

数据集的标签有标签提取模块来生成，主要是生成价格的涨、跌或者基本不变等离散的标签。生成好的标签与特征向量一起组成了基本的数据集。数据集被分为训练集和验证集两部分，训练集用于训练各类预测的模型，而验证集用于验证训练出的模型的好坏，以便调整模型的超参，从而让模型具有较好的范化能力。

3.1　数据获取

该模块主要为了获取与市场相关的信息，主要包含三方面的数据，价格数据、链上数据以及新闻文本数据。

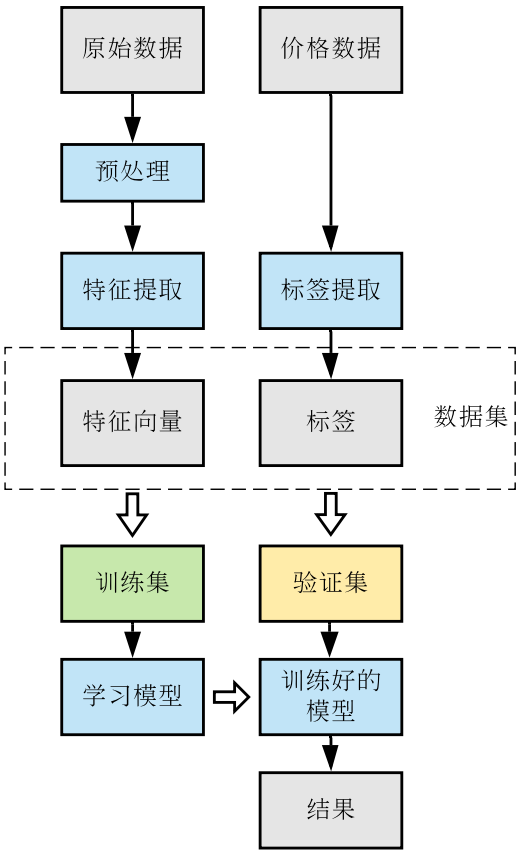

图 3.3　数据处理及模型训练流程

3.1.1　价格数据

价格数据可以从各大交易所获取，也可以从专业的价格提供网站获取。交易所和一些交易的网站会提供专用的 API 接口供用户使用，以方便用户及时获取相关数字资产的价格。

3.1.2　区块链上数据

区块链上的数据都是公开的，可以通过相应的区块链客户端同步整个区块链来获取数据。区块链的格式也是固定的，可以通过对区块链数据的分析获取有用的特征。另外，一些网站也会提供处理好的链上数据信息，在本地服务器功能有限的情况下，可以从这些提供数据的网站上通过标准的 API 接口获取。

3.1.3　新闻数据

数字资产发展迅速，目前已经有许多关于数字资产在线新闻网站，这些站点中包含了各类新闻信息，内容涵盖了各类数字资产的发展状况、各国政府的相关

政策、各种数字资产相关的重大事件、各类评论等等。与数字资产相关的重要新闻，都可以从这些网站通过爬虫来获取。

3.2　特征提取与降维

对于区块链上的数据，可以用直接提取的方法，论文提取并分析了九个和市场相关的特征。对于文本数据论文分析对比了几个常用的文本特征提取算法，同时论文还提出一种新的文本表征的方法。这种新的方法相比于以往的文本表征方法，在分析预测市场方面更具优势。下面我们简要介绍几种现有的文本特征提取的方法，并且仔细介绍新提出的文本表征方法。

3.2.1　文本特征提取

3.2.1.1　词袋模型

词袋模型提供了一种常用的文本表达模型，它不考虑文本间的顺序。在本文中，基于词袋模型对文本特征进行提取，主要分为以下三个步骤：

1. **分词**将文本按照字母或者单词等对文本进行划分。每一个划分出来的单元，在输出的向量中都占有一个位置；
2. **计数统计**每个词在文本中出现的次数，并将其放到文本表示向量的对应入口中去；
3. **标准化**将特征值（词频）映射到统一的范围内。

一般情况下，文本信息还会进行预处理操作，比如：去除标点符号、停用词等等。下面我们介绍论文中使用的几种典型、常用的文本特征提取方法。

3.2.1.2　TFIDF

TFIDF 是一种常用的文本表征方法，TF(term frequency) 表示词频，IDF(inverse document frequency) 表示反向文档频率。TFIDF 在统计词频的基础之上，引入了与文档集合相关的信息。

假设我们由 M 个文本文件，这 M 个文本文件中一共有 N 个不同的词。那么词频的计算公式如下：

$$tf(t,d) = \frac{freq(t,d)}{\sum_{t \in d} freq(t,d)} \tag{3-1}$$

上式3-1中，$freq(t,d)$ 表示词 t 在文档 d 中出现的频率，可以看出 TF 在词频的基础之上进行了标准化，以减少不同的文档长度不一致的问题。

IDF[51] 的计算公式如下：

$$idf(t, D) = \lg \frac{N}{|\{d \in D : t \in d\}|} \tag{3-2}$$

上式3-2中 D 表示文档集合，分子 N 为文档的个数，分母表示包含某个词 t 的文档的数量。

由上面两个式子我们可以得到 TFIDF 的计算公式：

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{3-3}$$

通过式3-3计算出某个词 t 在文档 $d \in D$ 中的 tfidf 值，如果词 t' 在文档中没有出现，那么在表示文档的向量中对应的位置则置零。

3.2.1.3　Doc2vec

Doc2vec[52] 由 Le 和 Mikolov 在 2014 年提出，它是一种非监督式的算法，是 word2vec[53] 的拓展。通过 doc2vec 学习出来的文本表征向量之间的距离可以反映文档之间的相似性。

Doc2vec 的思想和 word2vec 类似，在其向量训练的过程中，文本的 doc2vec 表示被用来预测段落中某个单词。具体来说，训练过程中不仅使用词向量来预测单词的上下文，整个段落或文档的表示也会被连接到词向量后面用于训练。这样在训练过程中，就可以学习到整个文档的表达。具体可参考 Le 和 Mikolov 的论文[52]。

3.2.1.4　情感图 (SentiGraph)

前面我们提到了几种通用的文本向量化的方法，有的是基于单个或多个词的，有的是基于文本嵌入（text embedding）的。这些方法常被用在各种文本分类的任务中。但是，这些方法是通用的方法，对于特定的任务没有做任何优化。

以市场预测任务来说，通过文本信息来进行市场预测，主要依赖于文本信息中：一、和市场息息相关的具体的实际的事件、数据等；二、一些带有情感信息的、可以影响到人们作出市场决策的信息，比如一些商业评论。以上基于词、词袋以及文本嵌入的方法并没有很好的反映这两点信息。

情感分析技术在近年来得到了非常广泛的研究，这项技术对于企业来说，有着重要的作用，它可以帮助企业分析消费者的态度[54]。对于投资者来说，它可以帮助分析当前市场是积极的或者消极的[55]。

在此背景下，我们提出了一种新的文本表征方法，该方法充分利用自然语言处理中情感分析技术，将一段长文本转化为一个图。而后我们可以利用这张图来

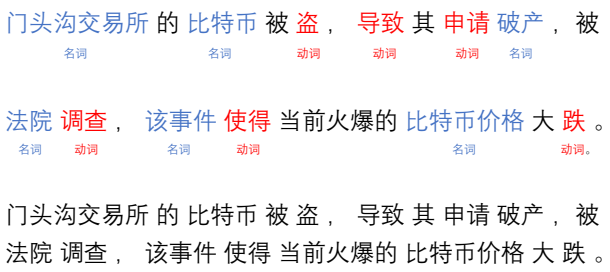

图 3.4　句子词性标注结果

分析该段文本是否会对市场产生影响。下面我们详细介绍我们的方法。

情感图的定义如下：

【定义】： 情感图 $SG(V,E)$ 由结点的集合 $V = \{v_1, v_2, \cdots, v_n\}$，以及边的集合 $E = \{(v_i, v_j, sv)\}$ 组成，其中 sv 表示边 (v_i, v_j) 的权重，它反映了结点 v_i 与 v_j 之间的情感信息（积极或消极）。sv 的求值取所有包含该词对 $\mathrm{A}(v_i, v_J)$ 的句子的情感值的均值。

情感图的构建　情感图的构建方法见算法1。情感图由结点以及边组成，我们对一段文本信息经过逐句遍历的方式来构建情感图。对于每一个句子，我们首先会对句子中的词进行词性标注 (POS tagging)，这样就可以知道每个句子中哪些是动词，哪些是名词。如：我们对句子”门头沟交易所的比特币被盗，导致其申请破产，被法院调查，该事件使得当前火爆的比特币价格大跌。“的标注结果如图所示3.4。那么该句子产生的词对有:(门头沟交易所, 破产)/(比特币, 破产)/(破产, 该事件)/(法院, 该事件)/(该事件, 比特币价格)，共五个。这些词对所在边的权重为该句子的情感分析值。通过情感分析技术，分析出这个句子的情感值为 0.12（情感值在 0 ~ 1 之间），也就是消极的。

在一篇文章中，某个词对可能出现多次，那么会有多个情感值，我们用这些情感值的均值作为词对最终的权重。特别的，当一个句子只有一个名词的时候那么在图中会产生一条指向该结点自己的边。

情感图举例　随着互联网的发展，纸制媒体正在逐渐被网络媒体所取代，网络新闻已经成为人们获取信息不可或缺的渠道。网络新闻中包含了大量的最新的事件

Algorithm 1 构建情感图

Input: 文本信息

Output: 情感图

```
1:  function BuildSentiGraph(text)
2:      for all sentence in text do
3:          value = sentiment_of(sentence)
4:          for all pair in sentence do
5:              if pair not in graph then
6:                  graph[pair].sentiment = emptylist
7:                  graph[pair].frequency = 0
8:              end if
9:              graph[pair].sentiment.append(value)
10:             graph[pair].frequency += 1
11:         end for
12:     end for
13:     for all pair in graph do
14:         value = mean(graph[pair].sentiment)
15:         graph[pair].sentiment = value
16:     end for
17:     return graph
18: end function
```

以及对这些事件的分析。这种分析可以反映出相关人员的态度（积极或者消极）。在一个句子中一个名词可能代表一个实体，而一个句子则可能表达了一个实体对另外一个实体的态度。下面我们通过网络上一篇实际的新闻文章来构建我们的情感图，并给出了一个构建好的情感图。该文章评述了 Mt.Gox 事件对比特币的影响。

文章的情感图表示见图3.5。下面我们简要分析一下该情感图。在图3.5中，边 (*bitcoin*, *mtgox*) 的权值是正的，而边 (*exchange*, *mtgox*) 和 (*exchange*, *bitcoin*) 的权重是负的。该条新闻描述的是 Mt. Gox 交易所有大量的比特币被盗事件。该事件导致了比特币市场的剧烈波动，引起了比特币价格的暴跌。直观来讲，如果在其它文章中，*mtgox* 和 *bitcoin* 一起出现了，那么很有可能作者想要表达一种消极的情感。然而我们可以看到在图中 (*bitcoin*, *mtgox*) 的权重为正的，因此这样单纯从词对的情感值分析并不全面，应该结合其它两个词对 (*exchange*, *mtgox*)/(*exchange*, *bitcoin*) 来综合分析。情感图中包含了所有词对之

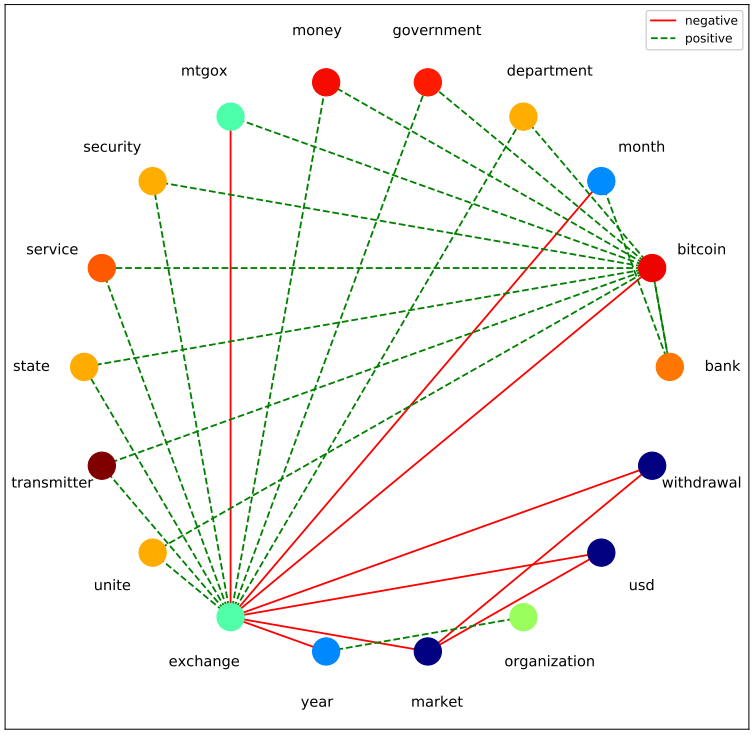

图 3.5　由实际的网络新闻文章构建而来的情感图。结点的颜色反映了该结点的平均的情感值，颜色越浅，积极情感成分越多，反之亦然

间的情感值，它相比于基于词、词袋的模型更具有表达能力。

3.2.2　降维

3.2.2.1　主成分分析 (PCA)

PCA 是常用的一种无监督式降维的方法，它将一个向量从一个坐标系中映射到另外一个低维的坐标系中。假设我们有一个样本 $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$，我们想要对其进行降维，使得转换后的向量长度为 l。PCA 的目标为学习转换矩阵。令转换后的向量为 $z_i = (z_{i1}, z_{i2}, \cdots, z_{il})$，那么有：

$$z_{ij} = w_j^T x_i \tag{3-4}$$

设样本集合 $S = \{x_1, x_2, \cdots, c_d\}$，其中样本已经经过了中心化处理。令样本所组成的矩阵为 $X = [x_1, x_2, \cdots, x_d]$，那么向量 w_j 可以由如下几个步骤给出：

1. 计算协方差矩阵 $C = XX^T$
2. 对 C 进行特征值分解：$C = Q\Lambda Q^T$
3. 取最大的 l 个特征值对应的特征向量作为 $W = [w_1, w_2, \cdots, w_l]$

3.3　数据分析与预测模型

数据分析模块对价格数据、链上数据、新闻文本数据进行分析，以发现市场相关的有用信息。价格数据主要使用时间序列分析的算法进行波动率分析，拟合等等。对于链上数据以及新闻数据，主要分析其和价格数据之间的相关性。对于链上数据，论文分析了其与价格数据之间的相关性。对于新闻文本，论文通过使用文本特征提取算法对其进行特征提取之后，使用机器学习方法分析了其与市场波动之间的关系。下面介绍论文使用的时间序列分析算法以及机器学习算法。

3.3.1　时间序列分析

3.3.1.1　线性时间序列模型

线性时间序列模型被设计用来对时间序列中的协方差结构进行建模。比较常用的线性模型包含两大类，一类是自回归模型，另一类是移动平均模型，这两者也可以结合起来形成自回归移动平均模型。

AR 模型　自回归模型可以从一系列过去的时间序列值中学习以预测下一个时间序列的值。它在统计学、经济学以及信号处理中有着非常广泛的应用。自回归模型 $AR(P)$ 的定义为：

$$X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \epsilon_t \tag{3-5}$$

其中 ϵ_t 是高斯白噪声，它与 X 相互独立。模型的系数 $a_i (1 \le i \le q)$ 是特征多项式：

$$a_1 z + a_2 z^2 + \cdots + a_p z^p = 1 \tag{3-6}$$

的根，所有解 z' 需要满足 $|z'| \ge 1$。特别的 $AR(1)$ 模型为马尔可夫过程，这表示 X_t 仅仅与 X_{t-1} 有关，而与再之前的值无关，这也就是说只需要通过 X_{t-1} 来预测 X_t 的值。

自回归建模主要包括模型判定、定阶、参数估计、模型检验四个步骤：

1. **模型判定**主要通过白噪声检验等方法，判断时间序列是否可以通过 AR 来进行建模。
2. **定阶**是为了确定公式3-5中 p 的值，也就是需要使用过去多少步的值来预测下一个值。
3. **参数估计**则是通过最大似然估计、最小二乘法、矩估计等方法来确定公式3-5中的参数 a_i。

4. **模型检验**的目的一是为了检验拟合的 AR 模型是否可以准确的反映时间序列的规律，通常可以通过对拟合的残差进行白噪声检验来进行；目的二是为了确定拟合的优劣程度。

MA 模型　移动平均模型又叫移动平均过程，是对单变量时间序列建模的一种常用方法。移动平均模型假定输出值与当前的随机误差以及过去的随机误差线性相关。q 阶移动平均模型 $MA(q)$ 的定义为：

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} = \mu + \epsilon_t + \sum_{j=1}^{q}\theta_j\epsilon_{t-j} \tag{3-7}$$

其中 μ 是时间序列的均值，ϵ_i 为随机误差，q 被称为移动平均模型的阶。

移动平均模型的工作方式类似于作用于白噪声的有限脉冲相应滤波器。它与 AR 模型不同的是，在 MA 模型中 ϵ_{t-1} 直接作用于 X_t，而在 AR 模型中 ϵ_{t-1} 不会直接影响 X_t。但在 AR 模型中 ϵ_{t-1} 会直接作用于 X_{t-1}，而 X_{t-1} 又会直接影响到 X_t，因此 ϵ_{t-1} 对 X_t 的影响是间接的。此外在 MA(q) 模型中，当前的误差或者白噪声只会影响到当前的预测值以及未来 q 步的预测值，而 AR(q) 模型中，当前的误差会无限向后传递 $X_t \to X_{t+1} \to \cdots \to X_{t+j} \to \cdots$

移动平均模型可以使用自相关函数来定阶，通过最大似然估计、矩估计、逆相关函数等方法进行模型参数估计。

ARMA 模型　ARMA (autoregressive moving average) 模型是自回归模型和移动平均模型的结合体。一个时间序列 $\{X_t; t = 0, \pm1, \pm2, \cdots\}$ 为 $ARMA(p, q)$，需满足[44]：1. 时间序列是平稳的；2.

$$X_t = a_1X_{t-1} + a_2X_{t-2} + \cdots + a_pX_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \tag{3-8}$$

其中 $a_q \neq 0, \theta_q \neq 0$，$p, q$ 分别为自回归和移动平均的阶。

如果 X_t 均值为非零值 μ，那么可设

$$\alpha = \mu(1 - a_1 - \cdots - a_q) \tag{3-9}$$

此时，模型变为：

$$X_t = \alpha + a_1X_{t-1} + a_2X_{t-2} + \cdots + a_pX_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \tag{3-10}$$

显然当 $q = 0$ 时，$ARMA(p, q)$ 退化为 p 阶自回归模型 $AR(p)$，当 $p = 0$ 时，$ARMA(p, q)$ 退化为 q 阶移动平均模型 $MA(q)$。

ARMA 模型的定阶可以由小到大尝试，保证 $p+q$ 越小越好，可以通过 AIC 准则来选择参数，用最大似然估计法进行参数估计。

ARIMA 模型　ARIMA(autoregressive integrated moving average) 是 ARMA 模型的改进版本。它们的目的都是为了深入分析序列数据或预测未来数据。ARIMA 模型在时间序列为非平稳时使用。模型名字中的"I"表示"集成"，表示数据值已经被替换为它们的值和之前值的差值，计算完差值之后，只需要对差分序列建立 ARMA 模型就行了。

3.3.1.2　波动率模型

ARCH 模型　Engle[45] 提出了 ARCH（autoregressive conditional heteroscedasticity）模型，中文译为"自回归条件异方差"模型，它第一次对波动率提出了理论模型，它将波动率定义成了条件标准差。ARCH 模型假定时间序列的条件方差为一个常数，在此种情况下，它对过程的条件均值进行建模。

$ARCH(m)$ 模型的定义如下：

$$\begin{aligned} a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \cdots + \alpha_m a_{t-m}^2 \end{aligned} \tag{3-11}$$

其中 ϵ_t 为白噪声过程，其均值为 0，并且独立同分布。

$ARCH$ 模型的定阶可以通过序列 $\{a_t^2\}$ 的偏自相关函数的截尾性来估计。$ARCH$ 模型的参数可以通过最大化条件对数似然函数得到。对于建立好的 $ARCH$ 模型可以通过计算标准化残差

$$\tilde{a}_t = \frac{a_t}{\sigma_t} \tag{3-12}$$

而后对 $\{\tilde{a}_t{}^2\}$ 以及 $\{\tilde{a}_t\}$ 进行白噪声检验以考察波动率方程以及均值方程的充分性。

GARCH 模型　GARCH 模型是 ARCH 模型的一种重要的推广，它有 Bollerslev[46] 于 1986 年提出。在 GARCH 中，ARMA 模型被用来对误差方差进行建模，$GARCH(p,q)$ 模型有如下公式定义：

$$\begin{aligned} a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \end{aligned} \tag{3-13}$$

$\{\epsilon_t\}$ 为白噪声序列，并且系数 β_i, α_i 均非负且 α_0 为正。GARCH 模型的定阶常使用试错法来尝试较低阶的模型（如 GARCH(1, 1)）。

3.3.2　机器学习算法

近几十年，机器学习飞速发展，现在在各行各业都有应用，已成为信息技术的支柱[47]。机器学习算法常被用来解决分类、回归、聚类等问题。在本论文中，我们对比了几种典型的算法的性能，包括：线性回归、支撑向量机 (SVM)、逻辑回归 (Logistic Regression)、随机森林和多层感知机 (MLP) 算法。下面我们对这几个算法进行简要的介绍。

3.3.2.1　线性回归

给定 m 个样本样本 $x = x_1, x_2, \cdots, x_m$，$x_i (1 \le i \le m)$ 为 d 维向量，x_i 对应的标签为 y_i。回归任务需要找到一个函数 $f : R^d \mapsto R$，使得：

$$\sum_{i=1}^{m} L(f(x_i), y_i) \tag{3-14}$$

最小，L 为损失函数。

线性回归中

$$f(x) = \omega^T x + b \tag{3-15}$$

ω, x 为 d 维列向量，b 为标量。学习器学得 ω, b 的值之后，我们就可以通过输入 x 来预测相应的 y 值。ω 中的值反应了输入的特征向量中，各个特征的重要程度，因此线性回归模型具有很好的解释性[48]。

以均方误差为例，线性回归的学习目标如下：

$$\begin{aligned}(w^*, b^*) &= \frac{1}{2} \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \\ &= \frac{1}{2} \arg\min_{(w,b)} \sum_{i=1}^{m} (\omega^T x_i + b - y_i)^2\end{aligned} \tag{3-16}$$

基于均方误差来求上述最小值的方法被称为最小二乘法。计上式为 $L_{(w,b)}$，上式对 ω 和 b 求导，有：

$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{m} x_i^T (\omega^T x_i + b - y_i) \tag{3-17}$$

$$\frac{\partial L}{\partial b} = mb - \sum_{i=1}^{m}(y_i - \omega^T x_i) \tag{3-18}$$

令上两式为 0 就可以求得最优闭式解 (ω^*, b^*):

$$\omega = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2} \tag{3-19}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - \omega^T x_i) \tag{3-20}$$

当方程不存在闭式解的时候，可能存在多个最优解，此时通常可以通过引入正则项，来根据归纳偏好来选择最优解。

3.3.2.2　支撑向量机 (SVM)

对于样本集合 $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，假设这些样本来自于类别 ω_1, ω_2 且两类样本是线性可分的，那么可以找到一个超平面:

$$g(x) = \omega^T x + b = 0 \tag{3-21}$$

将所有的样本 $x \in S$ 正确的分类。SVM 的的基本思想是找到一个超平面不但可以正确的对所有样本进行分类，而且可以是所有点到该平面的最小距离尽可能大。点到平面的距离为:

$$l = \frac{|g(x)|}{||\omega||} \tag{3-22}$$

我们假设到超平面最近的距离为 $\frac{1}{||\omega||}$，我们令类别 w_1, w_2 对应的类别 y 值为 +1, -1，并且有:

$$\begin{aligned} g(x) &\geq 1, \quad \forall x \in \omega_1 \\ g(x) &\leq -1, \quad \forall x \in \omega_2 \end{aligned} \tag{3-23}$$

那么 SVM 的求解可以转化为优化问题:

$$\min L_{(w,b)} = \frac{1}{2}||\omega||^2 \tag{3-24}$$

满足条件:

$$y_i(\omega^T x_i + b) \geq 1, \forall i \in [1, m] \tag{3-25}$$

根据 KKT 条件，上两式的问题必须满足：

$$\frac{\partial \mathcal{L}_{(\omega,b,\lambda)}}{\partial \omega} = 0 \tag{3-26}$$

$$\frac{\partial \mathcal{L}_{(\omega,b,\lambda)}}{\partial b} = 0 \tag{3-27}$$

$$\lambda_i \geq 0, \forall i \in [1, m] \tag{3-28}$$

$$\lambda_i(y_i(\omega^T x_i + b) - 1) = 0, \forall i \in [1, m] \tag{3-29}$$

上式中 $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ 为拉格朗日乘子。$\partial \mathcal{L}_{(\omega,b,\lambda)}$ 为拉格朗日函数，表示为：

$$\mathcal{L}_(\omega, b, \lambda) = \frac{1}{2}\omega^T\omega - \sum_{i=1}^{m} \lambda_i[y_i(\omega^T x - I + b) - 1] \tag{3-30}$$

可以通过求解上述问题的对偶问题来求解最优的拉格朗日乘子，之后再利用式3-26和3-27来求解得到参数 ω, b。

3.3.2.3　逻辑回归

逻辑回归是常用的解决分类问题的算法。我们将线性回归的输出经过 Sigmoid 函数：

$$\sigma(x) = \frac{1}{1 + exp(-x)} \tag{3-31}$$

处理之后再输出，则可以得到：

$$f(x) = \frac{1}{1 + exp(-(\omega^T x + b))} \tag{3-32}$$

将式3-32变换之后得到：

$$\ln \frac{f(x)}{1 - f(x)} = \omega^T x + b \tag{3-33}$$

这样可以将 $f(x)$ 看作 x 为类别 w_1 的可能性，$1 - f(x)$ 为 x 属于 w_2 的可能性。这样我们可以求得：

$$p(w_1|x) = \frac{exp(\omega^T + b)}{1 + exp(\omega^T + b)} \tag{3-34}$$

$$p(w_2|x) = 1 - p(w_1|x) \tag{3-35}$$

得到3-34和3-35之后，我们就可以通过优化最大似然估计函数来求得最优解。

3.3.2.4　随机森林 (RF)

集成学习可以分为两大类：boosting 和 bagging。Boosting 可以用来降低学习器的偏差，而 bagging 用来降低学习器的方差。随机森林 (RF)[49] 是一种非常典型的 bagging 类集成学习方法。Bagging 方法的本质是通过大量的模型来将噪音平均掉，这就降低了方差。决策树分类器非常适合作为 bagging 方法的弱分类器。树可以一直生长，直到每个结点都 z 只包含一个样本，这样可以相对减少偏差。随机森林中的树是同分布的，这意味着随机森林（树的集合）的偏差和单个树的偏差是一致的。这样 RF 在维持偏差不变的情况下降低了方差，从而提高了性能。它和 boosting 方法不一样，boosting 方法通过不断的调整来降低偏差，boosting 方法中的基分类器之间不是同分布的。

我们假设有 N 个独立同分布随机变量（方差为 σ^2），那么它们的均值的方差为 $\frac{1}{N}\sigma^2$。如果随机变量不是独立的，并且相关系数为 ρ，那么均值的方差为：

$$\rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \tag{3-36}$$

由式3-36可知，当我们增加基学习器的数量 N 的时候，方差会减小，但是方差的减小的收益会越来越小，极限是式3-36的左边部分。

3.3.2.5　多层感知机

多层（单隐层）感知机 (MLP) 是一种常用的神经网络，它包含三层：一个输入层、一个输出层以及一个输出层。除了输出层之外，每个神经元的输出都会使用一个非线性的激活函数处理之后再输出。MLP 使用反向传播算法进行模型的训练。它和线性模型的主要区别在于它是多层的而且包含非线性的激活函数。

单隐层的 MLP 可以用如下公式表示：

$$h(x) = \sigma(W_1 x + b_1) \tag{3-37}$$

$$f(x) = \phi(W_2(h(x)) + b_2) \tag{3-38}$$

其中 W_1 和 W_2 为权重矩阵，b_1, b_2 为偏差矩阵，ϕ, σ 为激活函数。公式3-37$h(x)$ 为隐层的输出。

$$relu(z) = \begin{cases} z, & \text{if } z \geq 0 \\ 0, & \text{if } n \leq 0 \end{cases} \tag{3-39}$$

$$tanh(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3-40}$$

神经网络常用的激活函数有 Sigmoid(公式3-31)，Relu(公式3-39), Tanh(公式3-40) 等等。

3.4　模拟交易

该层是为了验证各类模型预测的实际效果而设计，本层利用模型的预测结果，通过模拟交易策略来对收益结果进行仿真。

仿真采取最简单的交易策略，即如果预测涨并且有剩余资金，我们就买入，如果预测跌并且有剩余数字资产，那么就卖出；否则，不做任何操作。

第 4 章　实验设计与验证

本章基于第3章提出的框架进行实际的实验。实验基于一种典型的可交易数字资产---比特币进行。比特币是一种典型的加密货币，当前它占据了加密货币市场一半以上的市场份额，对其研究具有代表意义。下面我们对实验的各个阶段进行逐一详述。

4.1　数据采集

4.1.1　区块链上的数据

区块链上的数据可以从比特币的区块链上直接获取，也可以从提供这些数据的服务的网站上用爬虫爬取。在本实验中，数据直接从在线服务网站[57] 上获取，通过实时监控，我们还可以获取交易平均确认时间，获取的主要数据见表格4.1。

表 4.1　链上数据类型

数据名称	数据描述
num-of-tx	比特币的交易的次数
total-bitcoin-volume	总比特币交易量
unique-addr-num	仅出现过一次的比特币地址数量
total-tx-fee	总交易费用
fee-ratio	交易费占比
difficulty	难度
block-size	区块大小
num-tx-per-block	平均区块交易次数
tx-confirmation-time	交易确认时间

4.1.2　价格数据

有许多网站提供加密货币的实时价格信息，我们从 blockchain.com[57] 获取。

4.1.3　新闻数据

我们搜集了关于加密货币的从 2013 年起的 10 家主要提供比特币资讯的网站，数据源网站列表见表格4.2。

搜集到的新闻总数量为 108,433，总大小超过 200MB。

表 4.2　新闻数据来源

编号	站点
1	bitcoin.com
2	bitcoinist.com
3	bitcoinmagazine.com
4	ccn.com
5	cnbc.com
6	coindesk.com
7	coingeek.com
8	cointelegraph.com
9	investing.com
10	themerkle.com

4.2　数据特征提取与分析

本节对我们采集到的数据进行分析。针对不同类型的数据，我们用了不同的分析手段：

1. 对于从区块链上提取出的特征，我们对其进行相关系数分析；
2. 对于价格数据，我们进行了常用的时间序列分析；
3. 对于新闻数据，我们利用第3章中描述的文本特征提取的方法，对文本进行处理，并进行相关分析。

4.2.1　区块链数据特征提取与分析

本节通过分析各类数据与市场价格之间的相关系数来判断它们与市场价格之间的相关性。

图4.1显示了区块链上各个特征与价格之间的相关系数，特征根据相关系数的绝对值由小到大排列。相关系数的绝对值越大表示特征与价格之间的相关性越大，反之越小。相关系数为正表示正相关，相关系数为负表示负相关。

从图4.1中可以看出，每日比特币的总交易量与价格几乎不相关，总交易量与价格图见图4.2。

每个交易的交易费（以比特币为单位）是唯一一个与价格成反相关的，这意味着，如果交易费用上涨，那么比特币的价格就很有可能下跌，具体见图4.3。一般情况下如果交易费用上涨，那么说明人们对某一个交易愿意付出更多的比特币来支付。大多数情况下，人们愿意为一笔交易支付的实际金额（用法币结算）基本不变，因此在支付费用（比特币结算）上涨时，只有价格（以法币结算）降低才能

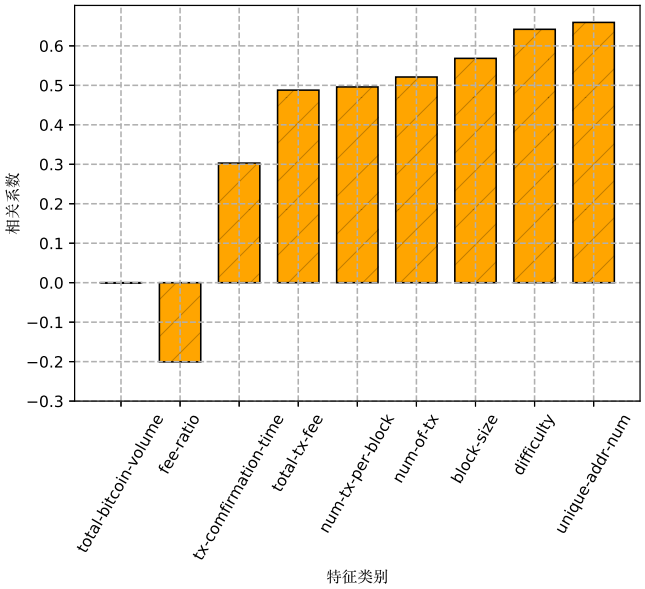

图 4.1　区块链提取的特征值相关性分析

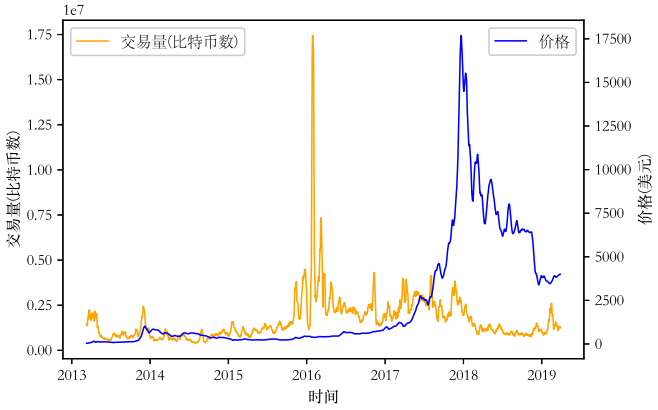

图 4.2　总比特币交易量与价格关系

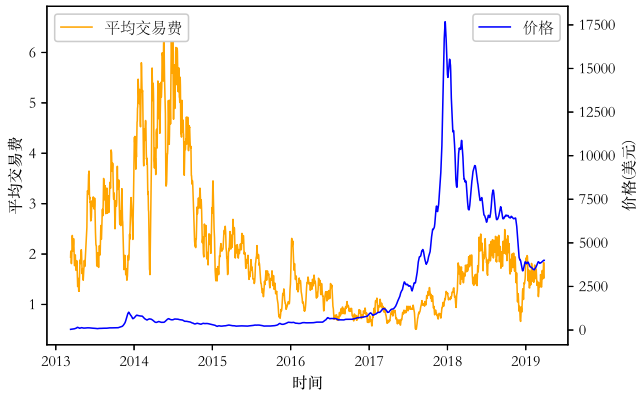

图 4.3　单笔交易交易费（以比特币为单位）与价格关系

保持实际支付的交易费用（法币结算）基本保持不变。因此，交易费（以比特币结算）与价格成负相关符合基本常识。挖矿难度以及唯一地址的数量与比特币价格

图 4.4　崭新地址的数量与价格关系

的相关系数超过了 0.6。崭新的地址的数量与价格4.4的相关系数最高，对于此现象的一种可能解释是：价格上涨时，愿意加入到比特币网络中的人会变多，新用户加入时会创建一个崭新的账号地址，因此在区块链上仅出现过一次的地址数量会增加；相反的，如果价格下跌，那么愿意在此时加入比特币网络的人会减少，那么新的地址就会相应的减少。挖矿难度与比特币价格4.5相关性大反映了价格对挖矿

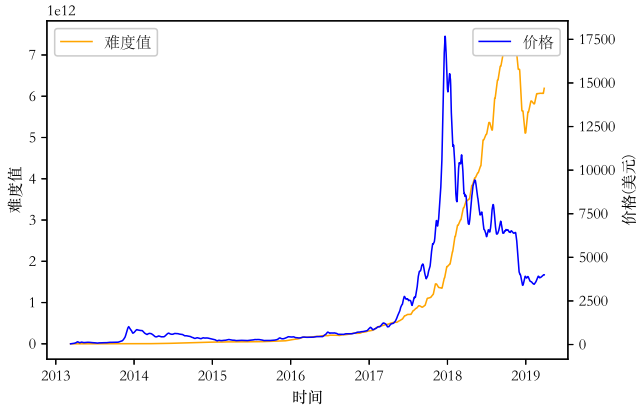

图 4.5　挖矿难度与价格关系

的影响。比特币价格上升的时候，矿工愿意投入更多的资源去挖矿，而价格下降的时候，矿工愿意投入的跨矿资源会减少，因此挖矿难度也会相应的下降。其余四个特征与价格的相关系数均在 0.5 左右。

以上分析的是区块链上的特征的总体情况。实际上，这些特征与价格之间的

关系是不断变化的，每一年都不一样。图4.6分析了最近 5 年每一年各个特征值与价格的相关系数的变化情况。

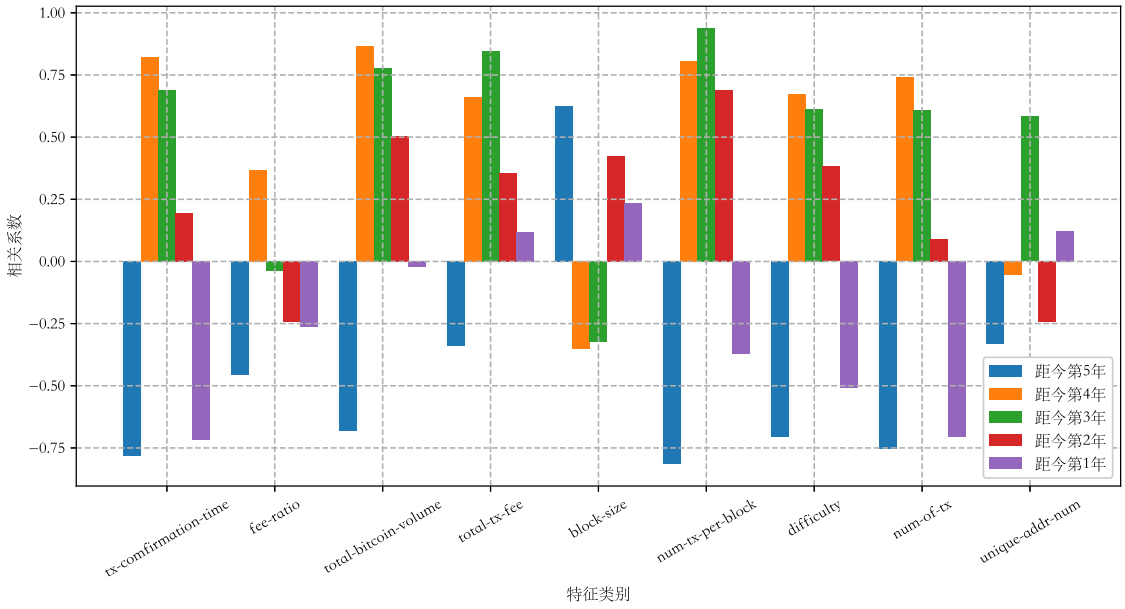

图 4.6　相关系数对比及其随年份变化情况

由图4.6中我们可以看出最近一年比特币的价格与前几年相比比较反常。之前几年正相关的几个特征，比如：tx-confirmation-time, fee-ratio, total-bitcoin-volume 等等，在最近一年与价格的相关系数都变成了负值。这说明在预测的时候，一定要考虑到各个特征的重要性、与价格的相关性是随时间变化的。

4.2.2　价格数据分析

4.2.2.1　原始价格数据

比特币的价格数据如图4.7所示。在 2017 年以前，比特币的价格相对平稳，波动相对较小，而在 2017 年之后，比特币的价格波动剧烈，整体大趋势呈现先上升后下降。

4.2.2.2　价格波动

图4.8给出了每 30 天的价格的标准差，从中我们可以直接看出市场波动状况。比特币价格在 2017 年末到达了顶峰，此时的价格波动巨大，而后波动逐渐变小。在 2018 年第四季度，价格波动又上升了一段时间。2019 年的价格波动状况虽然变小了，但是仍然比 2017 年以前的波动要大。

图 4.7　比特币价格数年变化情况

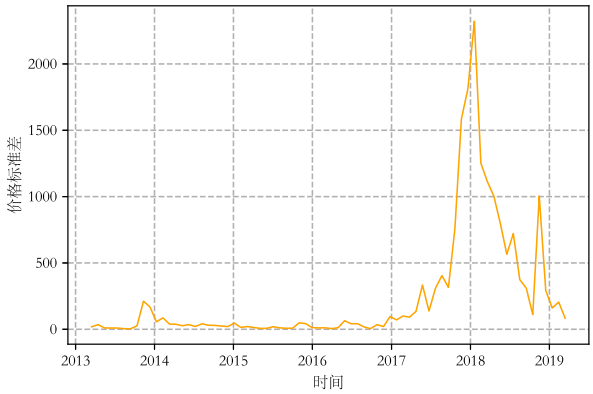

图 4.8　比特币价格波动情况

4.2.2.3　收益率分析

本小节，我们分析比特币的收益率。对数收益率的公式4-1如下：

$$logreturn = \ln\ (1 + \frac{price_{i+1} - price_i}{price_i}) \tag{4-1}$$

图4.9展示了比特币的每日的对数收益率逐年的变化以及对数收益率的概率密度。我们将其和与其均值方差一样的正太分布对比，见图4.10。在左图中我们将正态分布和对数收益作于同一图中，我们很明显可以观察出尖峰现象。为了便于观察肥尾现象，我们将对数收益率小于 10 的数值去掉，重新对概率密度曲线进行了拟合，见图4.10右图，我们可以很明显地观察到肥尾现象。比特币对数收益中出现了明显的尖峰肥尾，这说明投资比特币的风险较大，因此投资需谨慎。

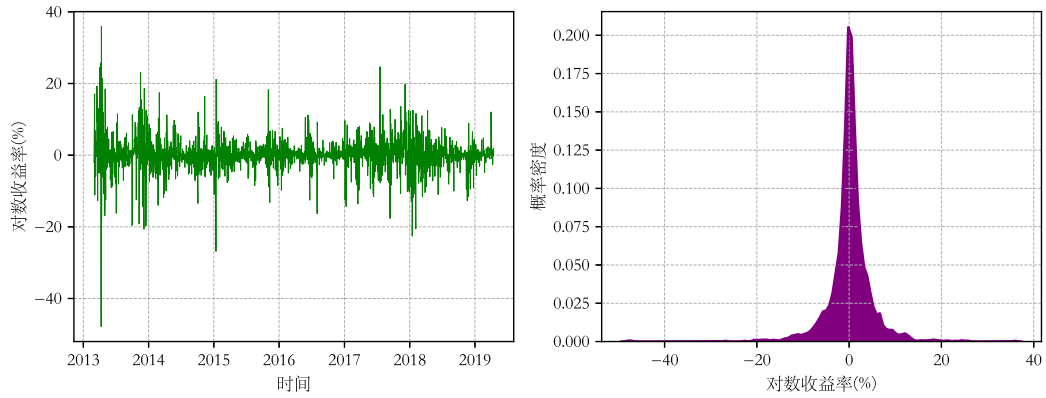

图 4.9　比特币收益率情况：每日对数收益率（左）；对数收益率概率密度（右）

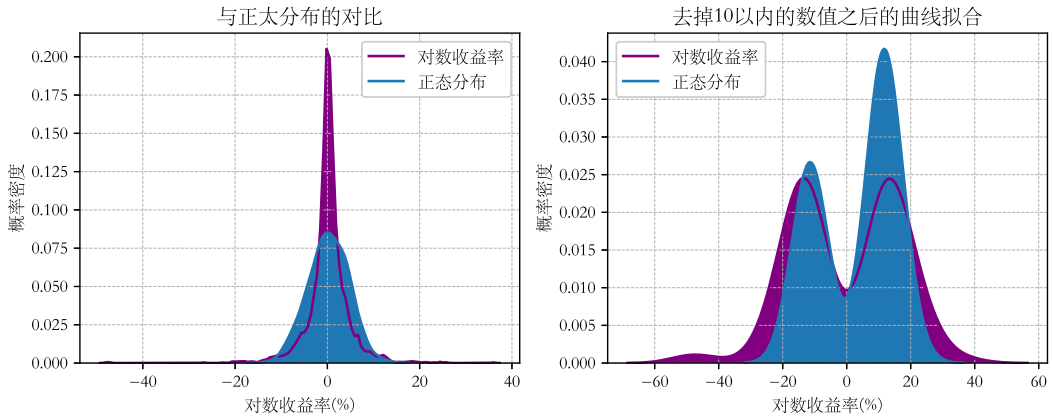

图 4.10　比特币对数收益的尖峰（左）、肥尾（右）现象

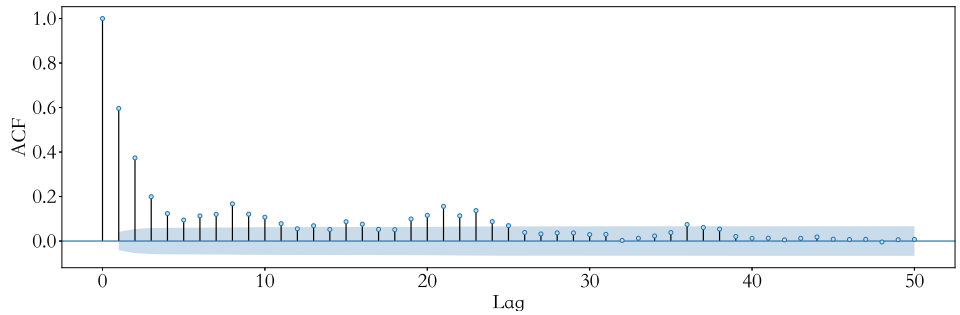

图 4.11　收益率平方的 ACF

4.2.2.4　(G)ARCH 建模分析

对市场的波动性进行建模与预测对于投资者的决策来说，具有非常重要的指导意义。本节将使用两种常用的模型对时间序列的方差进行建模，即自回归条件异方差模型（ARCH）与广义自回归条件异方差模型（GARCH）。

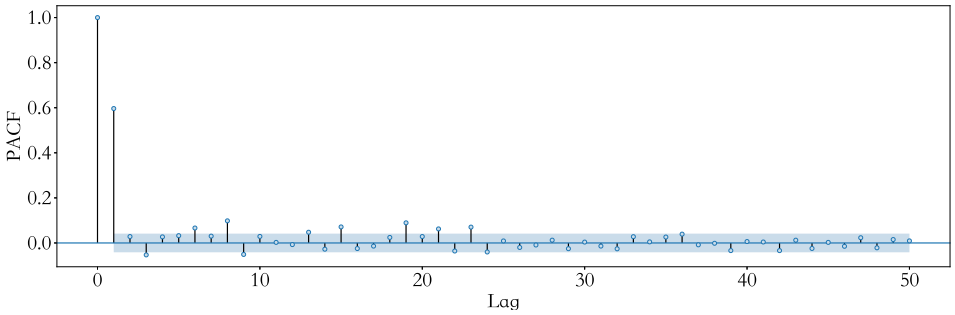

图 4.12　收益率平方的 PACF

对随时收益率序列进行 Engle 的拉格朗日乘子法 ARCH 效应检测之后，得到 p 值小于 3.6×10^{-169}，因此可以拒绝原假设，即序列具有相关性，因此具有 ARCH 效应。

图4.11展示了比特币对数收益率的平方的 ACF 图，图中显示出了较长期的低相关性。而后实验使用 $\{a_t^2\}$ 的偏自相关函数 (PACF) 来定阶。PACF 见图4.12。图4.12显示出滞后一直持续到 23。由此，我们可以根据 PACF，定阶为 23，从而建立 $ARCH(23)$ 模型较为合适，建模后，可以得到如图4.13所示的波动率拟合效果。从图中可以看出拟合的条件异方差序列很好的反映出了波动率。

$ARCH(23)$ 模型的参数较多，考虑用 $GARCH$ 模型进行改进。从图4.9(左)，可以看出对数收益率具有明显的波动聚集特性，而 $GARCH$ 模型就是基于此特性对时间序列进行建模的。波动聚集指的是大的波动之后有很大可能性仍然跟随着大的波动，而小的波动后面则很有可能跟随小的波动。

首先对时间序列进行 $GARCH$ 建模并调参，选择使用正态条件分布建立 $GARCH(1,2)$ 模型时，效果较好作为最终 $GARCH$ 模型：

$$\begin{aligned} r_t &= 0.00050 + a_t, \\ a_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= 0.000043 + 0.1\times a_{t-1}^2 + 0.39\times\sigma_{t-1}^2 + 0.39\times\sigma_{t-2}^2 \end{aligned} \tag{4-2}$$

其中 ϵ_t 独立同分布与标准正态分布。$GARCH(1,2)$ 模型的拟合速度远快于 $ARCH(23)$ 模型，而且拟合效果也非常好，效果如图4.14所示。

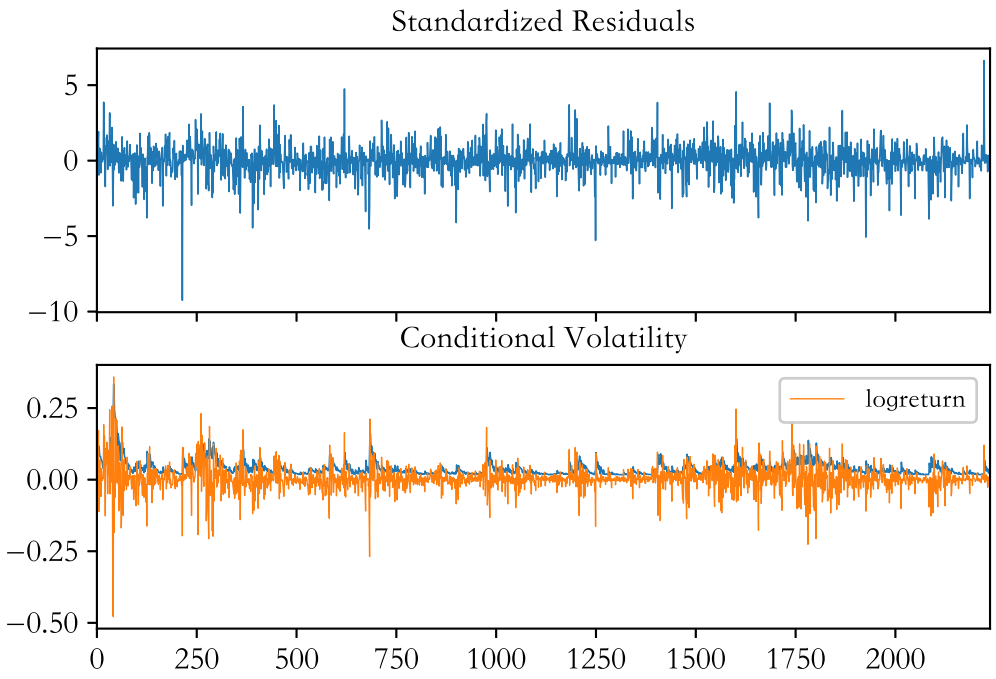

图 4.13　*ARCH*(23) 模型波动率拟合效果，第一张图展示了标准化残差；第二张图中橙色为对数收益率，蓝色为条件异方差序列

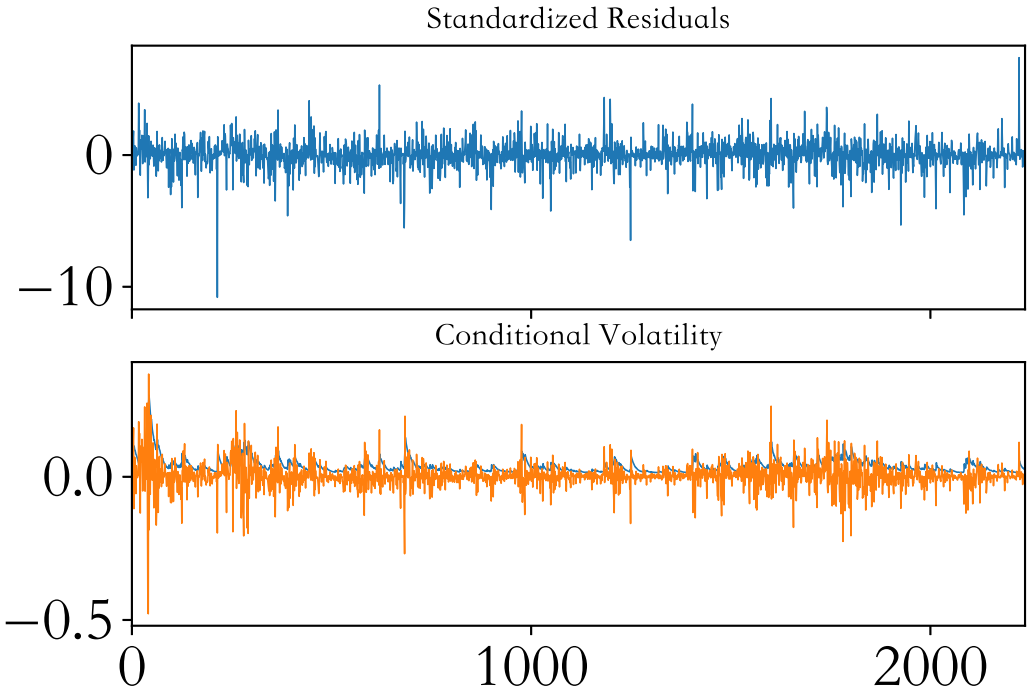

图 4.14　*GARCH*(1, 2) 模型波动率拟合效果，第一张图展示了标准化残差；第二张图中橙色为对数收益率，蓝色为条件异方差序列

4.2.3　文本数据特征提取与分析

文本特征提取使用了第3章介绍的几种文本提取方法，即基于词袋模型的方法、TFIDF、Doc2vec 和情感图。基于词袋模型的方法、TFIDF 以及 Doc2vec 的特征提取我们使用了已有的开源库。本节重点介绍新提出的文本表征方法，即情感图 (SentiGraph)。

情感图的构建主要分为三个步骤：

1. **预处理**对从网络中爬取的新闻文本信息进行清洗。主要任务是对文本中的非 ASCII 码字符、其它正常字符文本的 html 标签等进行去除，从而提取出新闻正文文本。
2. **情感分析**阶段，我们对句子进行情感分析。首先我们将新闻文本分割成一个个的句子。然后对每一个句子进行情感分析，标记每个句子的情感值。在实验中，我们使用了 python 的自然语言处理库 *nltk*[58] 中的 *Vader*[59] 模块。*Vader* 模块的情感分析对每个句子的积极情感、消极情感、自然情感以及综合情感都会给出一个值。在本实验中我们使用其中的综合情感值来作为句子的情感值。
3. **词性标注与图构建**阶段，对每个句子中的词进行词性标注并且根据词性标注的结果得到词对，以此来进行情感图的构建。词性标注使用了 *nltk* 中的 *pos_tag* 函数。在标注完词性后，我们使用算法1来构建情感图。

以上三个步骤介绍了基本情感图构建方法。下面介绍从情感图如何构建出机器学习可以使用的特征向量，即对图进行向量化。

为了研究新闻对比特币的影响能力，实验通过从每日新闻中提取出的特征来学习比特币价格的涨跌，并以学习模型的性能来作为新闻有效性的评价标准。第3章提到的常用的文本表征方法，此处不再赘述，重点放在新的文本表征方法，即情感图上。

情感图有一个个的结点和边组成，结点为名词实体，边为一对词所在句子的情感值。图4.15展示了所有词对情感值的分布状况。从图中可以看出，情感值为积极的要多于情感值为负的。在数据集中，所有词对的情感值的均值为 0.19，积极的词对大概占总词对的 3/4，而消极词对仅占 1/4。图4.15还显示出，情感值为 0 也就是不带情感的词对处的概率密度明显突出，这部分词对对于预测来讲无用，在处理特征时，这些词对会被去掉。在为 0 的词对被去掉之后，词对的情感值的均值呈现出正太分布。

每一个词对可能出现在许多句子中，因此一个词对在不同的上下文中的情感值很有可能是不同的，实验对词的在不同句子中情感值的序列的标准差进行了计

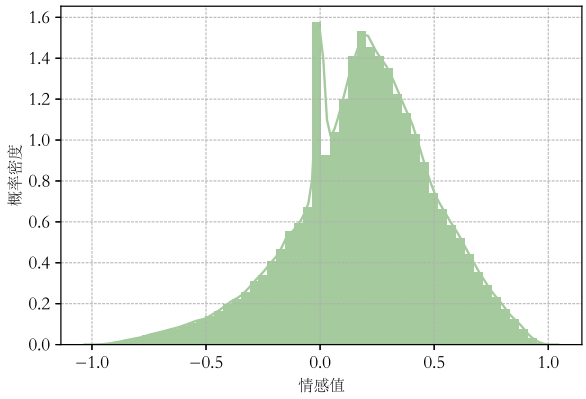

图 4.15　所有词对情感值分布

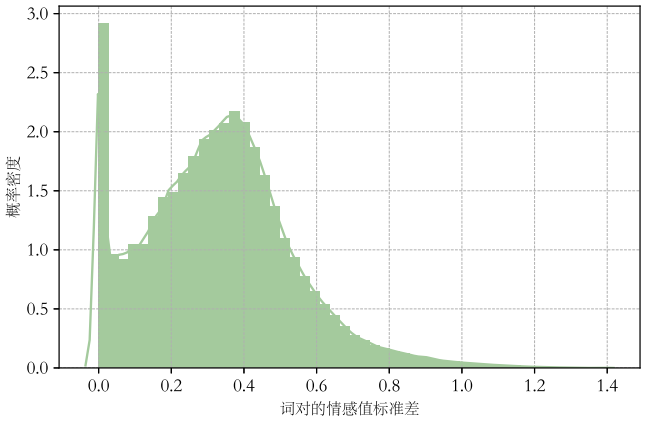

图 4.16　所有词对情感值的标准差概率分布

算，得到了其概率分布，见图4.16。图显示方差为 0 附近的词对的概率密度较高，这主要是由于，所提取出的词对中包含大量只出现过一次的词对，这些词对对于预测来讲无用，因此需要被过滤掉。

为了利用一天以内的新闻来预测价格的涨跌，实验首先必须将一天内新闻所构造出的情感图合并起来。情感图的合并需要注意的是不能直接对一个词的在不同图中的情感值进行简单求均值，而是要根据其在不同文本（新闻）中出现的频率来进行加权求和，这样得到的情感值，更具有区分度，更有利于预测。图4.17展示了使用加权求平均与直接求平均值的情感值的分布状况。其中横坐标 $Meanofsv$ 为均值，纵坐标 $StandardDeviationofsv$ 为方差，蓝色的点 $weightedmean$ 表示加权之后求均值均值，橙色的点 $simplemean$ 表示简单的求均值。从图中可以看出，根据词对出现的频率进行加权求和之后，不同词的情感值的分散程度远大与简单求均值。通过简单求均值合并情感图之后，可以发现情感值的均值集中在 0 左右，

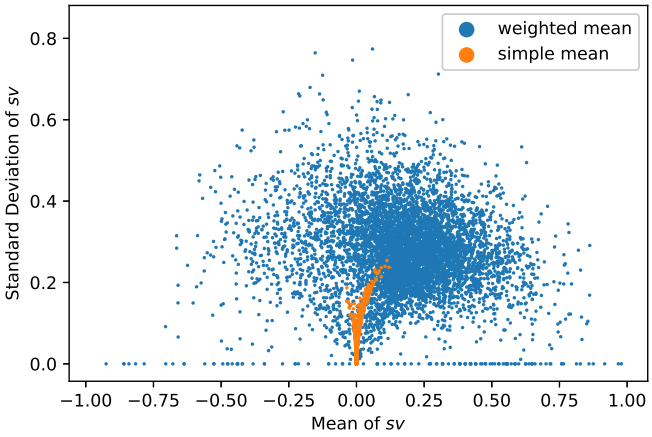

图 4.17　合并情感图时，使用加权求均值与直接求均值后，情感值的均值与方差的分布状况

而且标准差也接近 0，这很难用于预测。

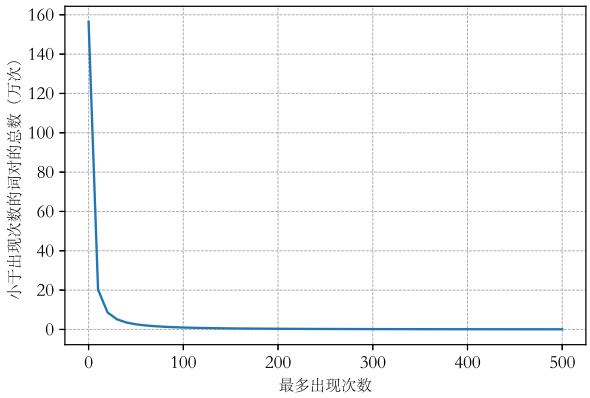

图 4.18　小于出现频率的词对的数量分布

构建情感图另外需要注意的一点就是，大量的词对可能仅出现几次 < 10，具体每个词对出现次数的分布如图4.18所示。可以看出绝大部分的词出现次数小于 10 次，对于这些预测能力有限的词，在构建情感图的时候需要被过滤掉。

对词对进行过滤之后，剩余的词对组成一个列表，列表中每个位置对应一个词对，这样通过将情感图中词对对应的情感值放入向量的相应位置，就可以得到一个情感图的向量表示。在对数据集进行 PCA 降维之后，我们得到了一个文本信息的低维数值表示。

4.2.3.1　各类文本表征方法对比

本节我们通过实验对比之前的一些文本表征方法与新的情感图，在对比特币日价格趋势预测方面的性能。

实验对比了 SVM、逻辑回归、随机森林以及多层感知机在使用不同表征方法在预测比特币价格涨跌方面的性能。

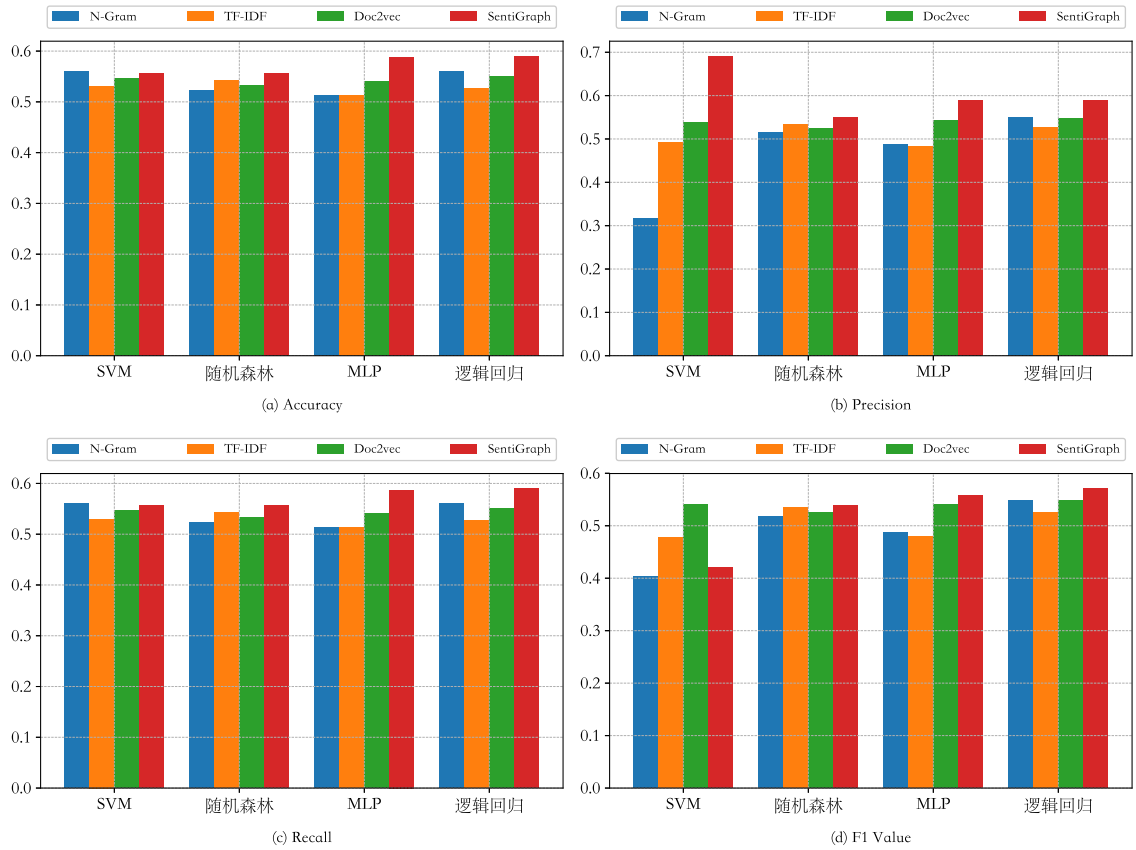

图 4.19　不同文本表征方法在各类算法下的性能对比

在我们的数据集中，一共包含了两千多天新闻的样本，我们用最后 300 天的数据作为测试集，其余的作为训练集。图4.19展示了我们的实验结果。

从实验结果我们可以看出，各种文本表征方法在四种机器学习方法下的准确率均不高于 60%，分类准确率最高的是使用情感图 + 逻辑回归的组合，准确率达到了 59%。总体看来使用情感图来表示新闻文本在价格预测方面的性能要显著高于其它的文本表征方法，它取得了最优的准确率 (Accuracy)0.59，最优的精度 (Precision)0.691，最优的召回率 (Recall)0.59 以及最优的 F1 值 0.571。在机器学习算法方面逻辑回归算法的分类效果总体高于其它三个机器学习算法。

4.3　预测模型实验对比

4.3.1　基于时间序列分析的方法

4.3.1.1　AR 模型

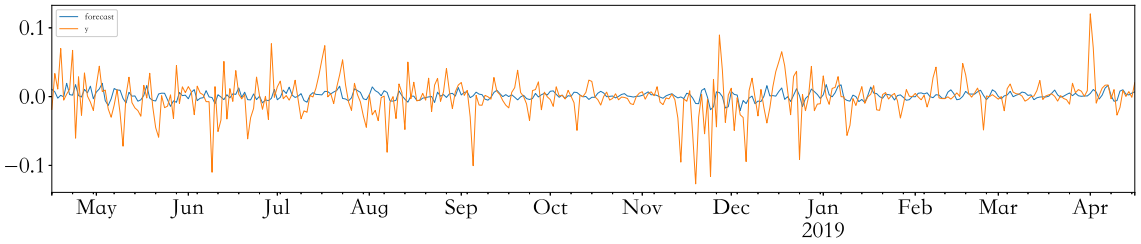

图 4.20　AR(24) 模型对数收益率预测效果

首先我们使用 ADF 函数对时间序列进行定阶，得到阶数为 24，然后我们对于对数收益率序列建立 $AR(24)$ 模型，并进行参数调整。图4.20展示了 AR 模型的部分预测结果，可以看出，给出的结果差强人意。AR 模型预测结果的计算出均方差为 $2.04 \times 10^{-3}\%$。同时，论文对结果进行了二值化处理，将收益率分为正负两类，得到了分类的结果见表4.3，同时计算得到了精度 (Accuracy) 为 54.65%。从表格4.3中我们可以看出预测“跌”的效果较差，$F1$ 值只有 0.45，而预测“涨”的准确率较高，召回率也达到了 0.66。

表 4.3　$AR(24)$ 模型对对数收益率预测二值化后的分类结果

类别	Precition	Recall	F1	Support
跌	0.50	0.40	0.45	1010
涨	0.58	0.66	0.62	1228

4.3.1.2　MA 模型

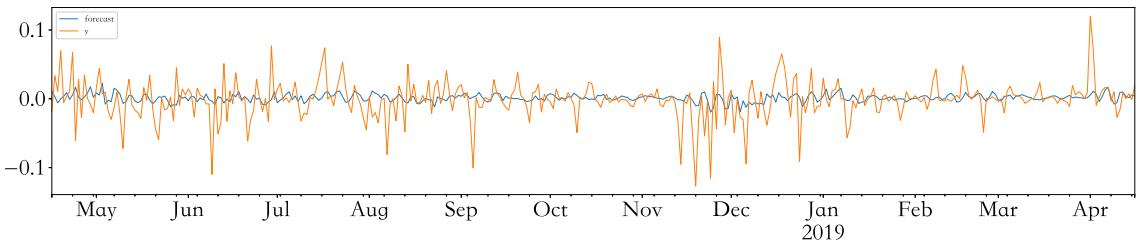

图 4.21　MA(23) 模型对数收益率部分预测效果

本节使用 MA 模型对比特币的对数收益率进行建模。实验首先使用了 $PACF$ 对对数时间序列进行定阶，得到阶数为 23。而后我们对序列进行 MA(23) 建模。图4.21展示了 $MA(23)$ 模型的部分预测结果，MA 模型的预测效果与 AR 模型差不多，$MA(23)$ 模型预测的均方差为 $2.06\times10^{-3}\%$，相较于 AR 模型稍微差一点。在对实验结果进行二值化之后，可以得到分类的效果，见表4.4，MA 分类的 $Accuracy$ 为 53.44%，效果略差于 AR 模型，但是准确率也超过了 50%。从表格4.4中我们可以看出，MA 模型的预测效果整体都不如 AR 模型。

表 4.4　$MA(23)$ 模型对对数收益率预测二值化后的分类结果

类别	Precition	Recall	F1	Support
跌	0.48	0.38	0.43	1010
涨	0.56	0.66	0.61	1228

4.3.1.3　ARMA 模型

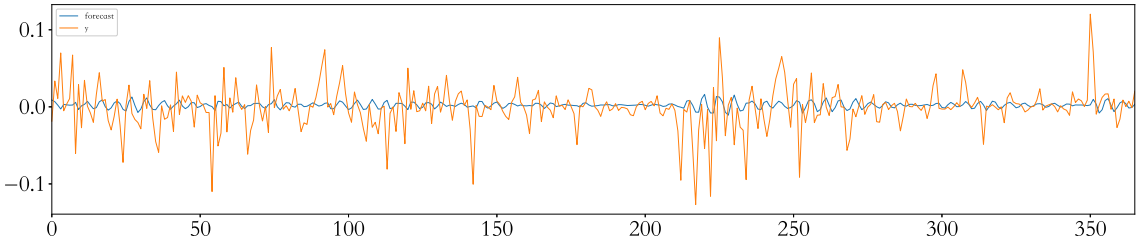

图 4.22　ARMA(2, 2) 模型对数收益率部分预测效果

ARMA 模型综合了 AR 模型与 MA 模型的特点。实验使用 AIC 值确定了 $ARMA(2,2)$ 模型。实验模型的部分预测效果如图4.22所示。$ARMA(2,2)$ 模型的预测的 $MSE = 0.0021$，预测值二值化之后的分类效果见表格4.5，从表格可以看出 $ARMA(2,2)$ 的效果不如 AR 模型和 MA 模型。

表 4.5　$ARMA(2,2)$ 模型对对数收益率预测二值化后的分类结果

类别	Precition	Recall	F1	Support
跌	0.45	0.28	0.35	1008
涨	0.55	0.72	0.62	1228

4.3.1.4　ARIMA 模型

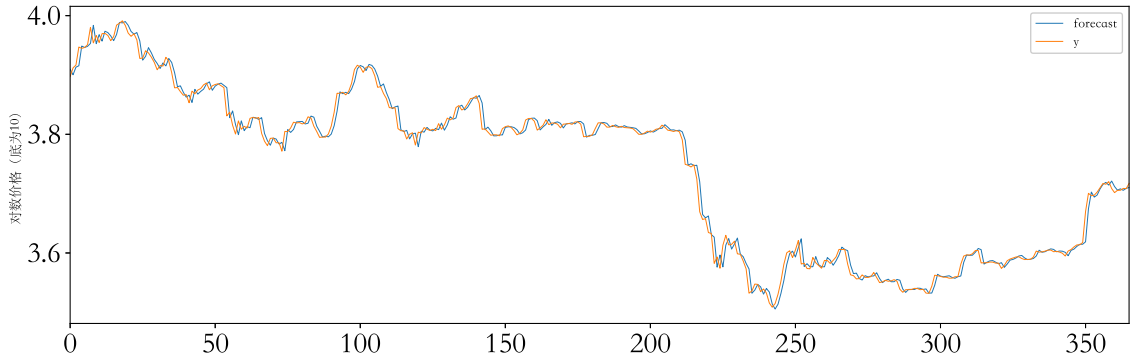

图 4.23　ARIMA(4, 1, 3) 模型对数收益率部分预测效果

实验对对数价格序列 $\{\log_{10}(price_i)\}$ 进行了 ARIMA 建模。经过参数选择得到参数 $ARIMA(4,1,3)$，也就是对对数价格序列进行一阶差分，自回归阶数为 4，移动平均阶数为 1。使用该参数对最近一年的对数价格向前一步预测的结果如图4.23所示。*ARIMA* 模型对于价格的拟合效果略微优于趋势跟随者，*ARIMA* 模型得到的均方误差为 0.00039，而趋势跟随者的均方误差为 0.00041。

4.3.2　基于机器学习/深度学习的方法

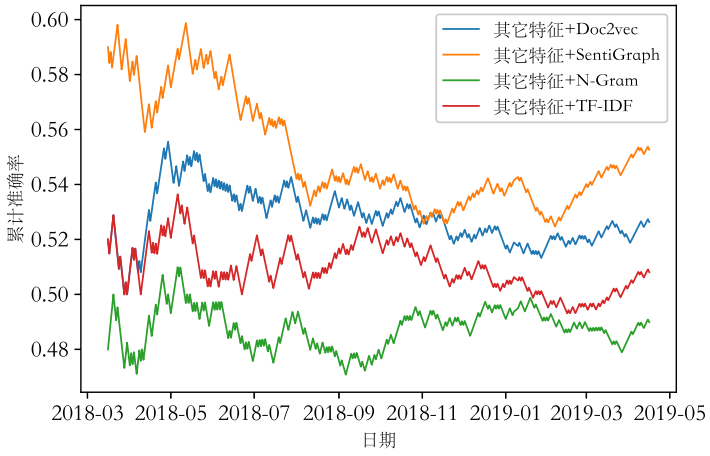

图 4.24　逻辑回归算法预测的累计准确率

本节将从区块链上提取的特征、价格数据与新闻文本信息中提取的特征连接到一起，并且利用在上文分类实验中效果较好的逻辑回归算法对输入数据进行训

练、预测。实验对比了新闻文本信息在不同的表征方法下与其它特征信息联合到一起之后的预测结果。从 2017 年末后的累计准确率如图4.24所示。从图中可以看出，我们提出的 *SentiGraph* 算法所给出的预测准确率为 55.3%，优于其它文本表征方法。*N − Gram* 算法的准确率最低，为 49.0%。另外从图中还可以发现，从 2018 年 5 月开始，预测的准确率开始逐渐下降，这从一定程度上说明了套利的空间减小了。

4.4　模拟在线交易

基于以上的预测结果，我们进行了模拟在线测试。实验使用了最简单的策略进行交易，本金为 1 万元，策略为：

1. 如果预测涨，并且手中拥有法币，那么使用所有法币购买虚拟货币；
2. 如果预测跌，并且手中拥有虚拟资产，那么将所有的虚拟资产卖出；
3. 否则，什么也不做。

4.4.1　基于时间序列预测的收益

本节基于时间序列预测的结果，进行模拟交易。

在无手续费的情况下，实验得到了如图4.25所示的实验结果。

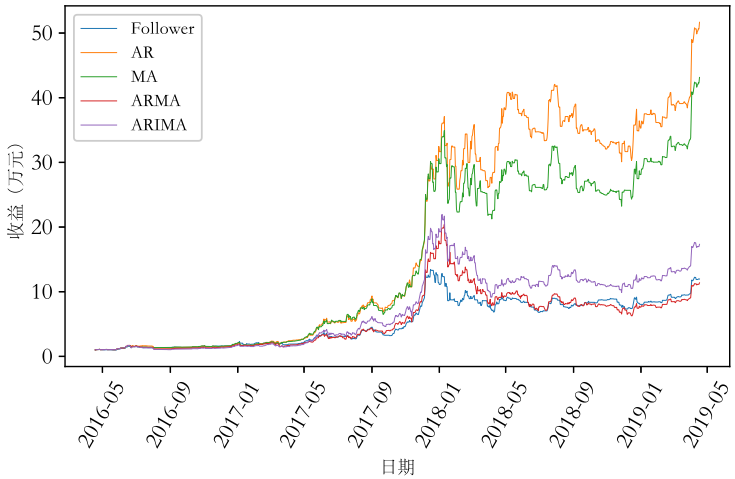

图 4.25　基于时间序列预测结果的模拟交易收益（无手续费）

基于时间序列的预测结果，在每次卖出资产情况下，如果收取 0.1% 手续费，则得到了如图4.26所示的实验结果。

其中 *Follower* 为趋势跟随者，从实验结果我们可以看出，AR 模型的三年间

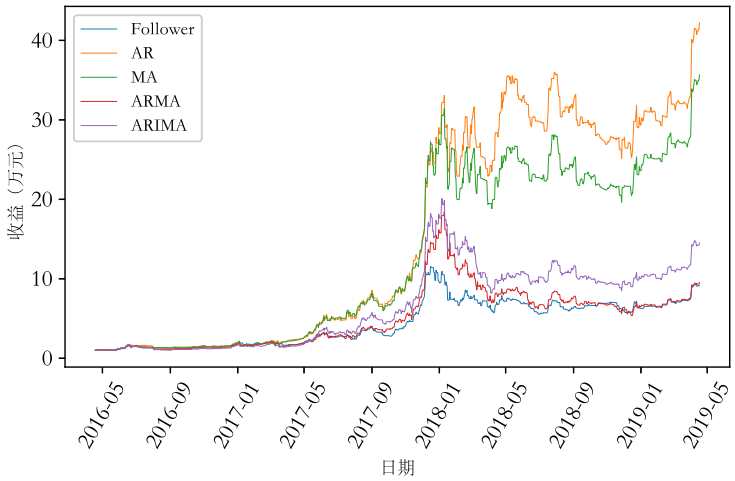

图 4.26　基于时间序列预测结果的模拟交易收益（0.1% 的手续费）

的收益最高，在无手续费的情况下，最终收益近 50 万，在 0.1% 手续费的情况下收益近 40 万。*ARMA* 模型的收益最低，甚至低于基准的趋势跟随者。

4.4.2　基于机器学习算法的收益

图4.27和图4.28分别给出了基于机器学习算法预测结果在无手续费与 0.1% 手续费时的累计收益状况。 从上图可以看出，无手续费和 0.1% 手续费的最终收益

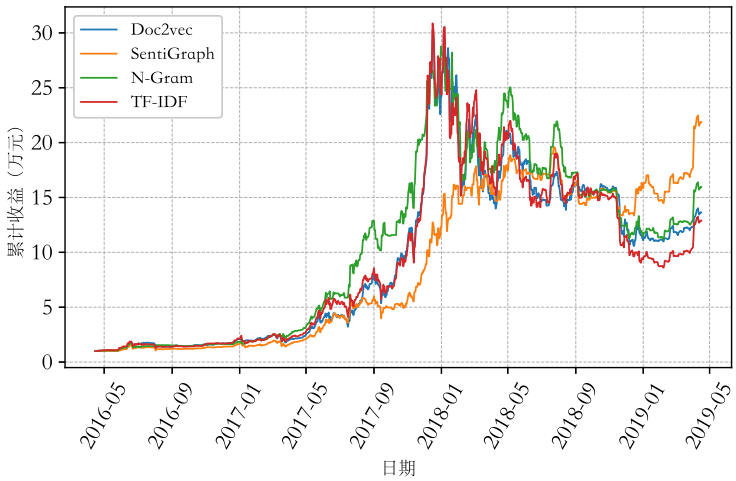

图 4.27　基于机器学习算法预测结果的模拟交易收益（无手续费）

相差较大，以使用 *SentiGraph* 特征为例，最终无手续费的收益多了约 4.27 万。同时 *SentiGraph* 方法在近期的收益明显优于其它方法。除 *SentiGraph* 以外的其它方法，在 2018 年之后，整体处于亏损状态。

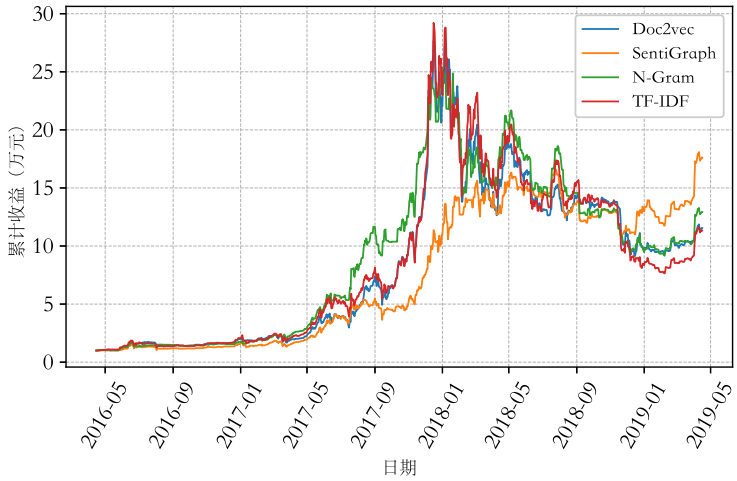

图 4.28　基于机器学习算法预测结果的模拟交易收益（0.1% 的手续费）

经过分析还可以发现，基于机器学习算法的预测结果所模拟出的收益要略差于基于时间序列分析的方法。而基于机器学习方法给出的预测的准确率却与时间序列分析方法差别不大，由此可见，预测准确率的高低并不一定意味着实际收益的高低。

第 5 章　结论

本文通过所设计的市场分析与预测框架，以比特币为研究对象对数字资产市场进行了分析。框架通过网络爬虫从网络获取市场相关的数据，并对其进行分析，主要分为两大部分内容：

- 第一部分针对市场价格进行分析建模。在对市场收益率的分析中，发现了数字资产市场的对数收益率具有很典型的“尖峰肥尾”现象，这表明数字资产市场的投资风险较高。实验还利用 (G)ARCH 模型对市场波动进行了建模分析，发现市场收益率具有波动集聚性。同时，框架利用时间序列分析法即 AR、MA、ARMA、和 ARIMA 对市场的收益率以及价格进行建模，研究市场的可预测性。模拟交易实验的结果表明使用 AR 对市场波动率进行建模可以在三年内获得数十倍的收益。
- 第二大部分内容为对市场的影响因素进行分析，本文分析了两大类数据：1. 区块链上的交易数据；2. 新闻信息。研究结果表明区块链上提取出的特征和市场价格的相关性是随时间变化的，而新闻信息对于市场价格波动具有一定的影响力。同时针对新闻数据我们提出了一种新的文本表征方法，该方法在分析新闻对市场价格波动的影响时，相对于其它传统的文本表征方法显示出了优势。最后我们利用从区块链、历史价格数据以及新闻信息中提取出的特征对市场价格波动进行了预测并且进行了模拟交易，发现最优的特征组合在三年的时间内可以获得近 20 倍的收益。实验结果还发现，随着近年来市场的有效性的逐步增加，套利的空间正在逐步减小，同时该新兴市场的回报率的波动大，投资的风险巨大。

插图索引

表格索引

公式索引

参考文献

[1] Nakamoto S, et al. Bitcoin: A peer-to-peer electronic cash system[Z]. [S.l.]: Working Paper, 2008.

[2] Wood G. Ethereum: A secure decentralised generalised transaction ledger[J]. Ethereum project yellow paper, 2014, 151: 1-32.

[3] Ripple. Ripple - One Frictionless Experience To Send Money Globally | Ripple[EB/OL]. 2019. https://ripple.com/.

[4] LiteCoin. Money for the Internet Age.[EB/OL]. 2019. https://www.litecoin.com/en/.

[5] Dash. What is Dash?[EB/OL]. 2019. https://docs.dash.org/en/stable/introduction/about.html.

[6] CoinMarketCap. Cryptocurrency Market Capitalizations | CoinMarketCap[EB/OL]. 2019. https://coinmarketcap.com.

[7] Microsoft. How to use Bitcoin to add money to your Microsoft account[EB/OL]. 2019. https://support.microsoft.com/en-us/help/13942/microsoft-account-how-to-use-bitcoin-to-add-money-to-your-account.

[8] Dyhrberg A H. Bitcoin, gold and the dollar–A GARCH volatility analysis[J]. Finance Research Letters, 2016, 16: 85-92.

[9] Gandal N, Halaburda H. Competition in the cryptocurrency market[Z]. [S.l.]: CEPR Discussion Paper No. DP10157, 2014.

[10] ElBahrawy A, Alessandretti L, Kandler A, et al. Evolutionary dynamics of the cryptocurrency market[J]. Royal Society open science, 2017, 4(11): 170623.

[11] Cocco L, Concas G, Marchesi M. Using an artificial financial market for studying a cryptocurrency market[J]. Journal of Economic Interaction and Coordination, 2017, 12(2): 345-365.

[12] Caporale G M, Gil-Alana L, Plastun A. Persistence in the cryptocurrency market[J]. Research in International Business and Finance, 2018, 46: 141-148.

[13] Khuntia S, Pattanayak J. Adaptive market hypothesis and evolving predictability of bitcoin[J]. Economics Letters, 2018, 167: 26-28.

[14] Chun Wei W. Liquidity and market efficiency in cryptocurrencies[J]. Economics Letters, 2018, 168: 21-24.

[15] Krafft P M, Della Penna N, Pentland A S. An experimental study of cryptocurrency market dynamics[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. [S.l.]: ACM, 2018: 605.

[16] Bouri E, Lau C K M, Lucey B, et al. Trading volume and the predictability of return and volatility in the cryptocurrency market[J]. Finance Research Letters, 2018.

[17] Caporale G M, Plastun A. The day of the week effect in the cryptocurrency market[J]. Finance Research Letters, 2018.

[18] Bouri E, Shahzad S J H, Roubaud D. Co-explosivity in the cryptocurrency market[J]. Finance Research Letters, 2018.

[19] Alessandretti L, ElBahrawy A, Aiello L M, et al. Anticipating cryptocurrency prices using machine learning[J]. Complexity, 2018, 2018.

[20] Catania L, Grassi S, Ravazzolo F. Predicting the volatility of cryptocurrency time-series[M]// Mathematical and Statistical Methods for Actuarial Sciences and Finance. [S.l.]: Springer, 2018: 203-207

[21] Bukovina J, Marticek M, et al. Sentiment and bitcoin volatility[R]. [S.l.]: Mendel University in Brno, Faculty of Business and Economics, 2016.

[22] Matta M, Lunesu I, Marchesi M. Bitcoin spread prediction using social and web search media. [C]//UMAP Workshops. [S.l.: s.n.], 2015: 1-10.

[23] Stenqvist E, Lönnö J. Predicting bitcoin price fluctuation with twitter sentiment analysis[Z]. [S.l.: s.n.], 2017.

[24] Shah D, Zhang K. Bayesian regression and bitcoin[C]//2014 52nd annual Allerton conference on communication, control, and computing (Allerton). [S.l.]: IEEE, 2014: 409-414.

[25] Madan I, Saluja S, Zhao A. Automated bitcoin trading via machine learning algorithms[J]. URL: http://cs229. stanford. edu/proj2014/Isaac Madan, 2015, 20.

[26] Jang H, Lee J. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information[J]. IEEE Access, 2018, 6: 5427-5437.

[27] McNally S, Roche J, Caton S. Predicting the price of Bitcoin using Machine Learning[C]//2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). [S.l.]: IEEE, 2018: 339-343.

[28] CoinDesk. Bitcoin Price Index[EB/OL]. 2019. https://www.coindesk.com/price/bitcoin.

[29] Almeida J, Tata S, Moser A, et al. Bitcoin prediciton using ann[J]. Neural networks, 2015: 1-12.

[30] Greaves A, Au B. Using the bitcoin transaction graph to predict the price of bitcoin[J]. No Data, 2015.

[31] Akcora C G, Dey A K, Gel Y R, et al. Forecasting bitcoin price with graph chainlets[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. [S.l.]: Springer, 2018: 765-776.

[32] Jiang Z, Liang J. Cryptocurrency portfolio management with deep reinforcement learning[C]// 2017 Intelligent Systems Conference (IntelliSys). [S.l.]: IEEE, 2017: 905-913.

[33] Hayes A S. Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin[J]. Telematics and Informatics, 2017, 34(7): 1308-1321.

[34] Sovbetov Y. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero[J]. Journal of Economics and Financial Analysis, 2018, 2(2): 1-27.

[35] Hurlburt G F. Shining light on the dark web.[J]. IEEE Computer, 2017, 50(4): 100-105.

[36] Piazza F. Bitcoin in the dark web: a shadow over banking secrecy and a call for global response [J]. S. Cal. Interdisc. LJ, 2016, 26: 521.

[37] Brown S D. Cryptocurrency and criminality: The bitcoin opportunity[J]. The Police Journal, 2016, 89(4): 327-339.

[38] Bing L, Chan K C, Ou C. Public sentiment analysis in twitter data for prediction of a company's stock price movements[C]//2014 IEEE 11th International Conference on e-Business Engineering. [S.l.]: IEEE, 2014: 232-239.

[39] Loughlin C, Harnisch E. The viability of stocktwits and google trends to predict the stock market[J]. Retrieved from stocktwits: http://stocktwits. com/research/Viabilityof-StockTwits-and-Google-Trends-Loughlin_Harnisch. pdf, 2014.

[40] Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news: The azfin text system[J]. ACM Transactions on Information Systems (TOIS), 2009, 27(2): 12.

[41] of Congress T L L. Regulation of cryptocurrency around the world[EB/OL]. 2019. https://www.loc.gov/law/help/cryptocurrency/world-survey.php.

[42] 吴静. 互联网经济背景下虚拟货币价格形成机制研究——以比特币为例[博士学位论文]. [出版地不详]: 中国海洋大学, 2015.

[43] Lau J H, Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation[J]. arXiv preprint arXiv:1607.05368, 2016.

[44] Shumway R H, Stoffer D S. Time series analysis and its applications: with r examples[M]. [S.l.]: Springer, 2017

[45] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation[J]. Econometrica: Journal of the Econometric Society, 1982: 987-1007.

[46] Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of econometrics, 1986, 31(3): 307-327.

[47] Smola A, Vishwanathan S. Introduction to machine learning[J]. Cambridge University, UK, 2008, 32: 34.

[48] 周志华. 机器学习[M]. [出版地不详]: Qing hua da xue chu ban she, 2016

[49] Ho T K. Random decision forests[C]//Proceedings of 3rd international conference on document analysis and recognition: volume 1. [S.l.]: IEEE, 1995: 278-282.

[50] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[51] Robertson S. Understanding inverse document frequency: on theoretical arguments for idf[J]. Journal of documentation, 2004, 60(5): 503-520.

[52] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International conference on machine learning. [S.l.: s.n.], 2014: 1188-1196.

[53] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[M/OL]//Burges C J C, Bottou L, Welling M, et al. Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013: 3111-3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[54] Hu G, Bhargava P, Fuhrmann S, et al. Analyzing users'sentiment towards popular consumer industries and brands on twitter[C]//2017 IEEE International Conference on Data Mining Workshops (ICDMW). [S.l.]: IEEE, 2017: 381-388.

[55] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of computational science, 2011, 2(1): 1-8.

[56] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm [C]//Aaai: volume 2. [S.l.: s.n.], 1992: 129-134.

[57] blockchain.com. 比特币图表[EB/OL]. 2019. https://www.blockchain.com/zh-cn/charts.

[58] Loper E, Bird S. Nltk: the natural language toolkit[J]. arXiv preprint cs/0205028, 2002.

[59] Hutto C J, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text[C]//Eighth international AAAI conference on weblogs and social media. [S.l.: s.n.], 2014.

致　谢

衷心感谢导师徐恪教授和李琦副教授对本人的精心指导。他们的言传身教将使我终生受益。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1991 年 2 月 11 日出生于江苏省射阳县。

2009 年 9 月考入河海大学计算机与信息学院通信工程专业，2013 年 7 月本科毕业并获得工程学士学位。

2016 年 9 月进入清华大学计算机科学与技术系攻读工程硕士学位至今。

发表的学术论文

[1] 徐恪, **姚文兵.** 赛博智能经济与区块链 [J]. 广东工业大学学报, 2018, v.35; No.134(03):5-13.

[2] **W. Yao**, K. Xu and Q. Li, "Exploring the Influence of News Articles on Bitcoin Price with Machine Learning," 2019 IEEE Symposium on Computers and Communications (ISCC) (To be published).