|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 基于模式的股票价格走势预测方法研究 | | | | | | | | |
| (申请清华大学工程硕士专业学位论文) | | | | | | | | |
|  | | | | | | | | |
| 培 养 单 位 | | ： | 软件学院 | | | |
| 工程领域 | | ： | 软件工程 | | | |
| 申 请 人 | | ： | 林泉韬 | |  |
| 指导教师 | | ： | 邓仰东 | | 副 教 授 |
|  | |  |  | |  |
|  | | | | | | | | |
|  | | 二○一八年十月 | | |  | | | |

|  |  |  |
| --- | --- | --- |
| **Pattern Based Detection and Prediction of Stock Price With Dynamic Time Warping** | | |
| Thesis Submitted to  **Tsinghua University**  in partial fulfillment of the requirement  for the professional degree of  **Master of Software Engineering** | | |
|  | | |
| by  **Lin Quantao**  **(** **Software Engineering)** | | |
|  | | |
| Thesis Supervisor | : | Professor Dent Yangdong |
|  |  |  |
|  | | |
| **October, 2018** | | |

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遵守此规定）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 作者签名： |  |  | 导师签名： |  |
| 日 期： |  |  | 日 期： |  |

摘 要

股票市场是金融领域的一个重要组成部分，股票的价格涨跌既与经济领域的各项指标存在的错综复杂的关系，也与社会文化的特质存在着千丝万缕的联系。通过分析经济和非经济的相关数据对股票价格进行预测可以为股票交易参与者的交易决策提供依据，因此在股票发展的过程中，涌现了各种各样的数据分析手段。股票价格本身是一个非线性的时间序列，具有非结构化、非平稳等特点。在早期的研究中，就有很多研究者发现了有些股票价格的波动趋势具有重复性，即在某些股票价格波动模式出现后，往往会出现股票价格的上涨或者下跌，并因此形成了诸如波动操作理论，江恩理论等分析方法。本文试图通过趋势分析的方法将股票的历史数据分为上涨趋势和下跌趋势，并结合动态时间规划（DTW）和基于时间序列的shapelets算法分析出股票价格中具有区分股票上涨或者下跌趋势的股票价格序列片段。主要的工作如下：

研究了数据挖掘算法在股票预测方面的应用。通过对已有的研究成果的分析，了解数据挖掘在股票预测这个应用场景下可能遇到的问题和采用的改良方案。

研究了动态时间规划（DTW）和基于时间序列的shapelets算法。因为股票价格序列是一个长的连续序列，如果单纯使用DTW算法应用于股票价格预测，那么需要将近期的股票价格序列与全体股票的序列进行相似性度量，之后采用某种分类方法进行预测。这样不但会导致数据量巨大，而且因为股票价格序列中大量的噪音信息，对预测的准确性造成影响。因此引入基于时间序列的shapelets算法，通过前期挖掘出的股票序列特征子序列进行匹配，可以提高精度和计算速度。

本文结合股票实际操作的应用场景，结合动态时间规划（DTW）和基于时间序列的shapelets数据挖掘算法，提出了基于模式发现的股票价格走势预测方案，并通过实验对方案的预测结果进行了分析。

关键词：股票价格 股票预测 预测技术 动态时间规整 shaplet

**Abstract**

The stock market plays a vital part of the financial field. The rise and fall of stock prices are not merely intricately related to various indicators in the economic field, but closely related to social and cultural characteristics. Analyzing the relevant economic and non-economic data to predict stock prices can provide the basis for the trading decision-making of the participants in the stock trading. Therefore, various data analysis methods have emerged in the course of stock development. Stock price itself is a non-linear time series characterized by non-structure and non-stability. In the early studies, many researchers found that some stock price fluctuation trends are repetitive, that is, after some stock price fluctuation patterns appear, stock prices tend to rise or fall, thus forming analysis methods such as Fluctuation Operation Theory and Gann Theory. This paper attempts to divide the historical data of stocks into rising trend and falling trend by means of trend analysis, and analyzes the stock price sequence segments distinguishing the rising trend and falling trend in the stock prices by combining the dynamic time warping (DTW) and shapelets algorithm. The main work is as follows:

First, this work investigates the applicability of modern data mining algorithms in the field of stock forecasting. The analysis of existing research results reveals the problems that data mining may encounter in the application scenario of stock forecasting and the improvement scheme adopted.

Second, The Dynamic Time Warping (DTW) and shapelets algorithm are studied. Since the stock price sequence is a long continuous sequence, if only DTW algorithm is applied in stock price forecasting, it is necessary to measure the similarity between the recent stock price sequence and the sequence of all stocks, and then use some sort of classification method to forecast. This will not only lead to huge amount of data, but affect the accuracy of forecasting due to the large amount of noise information in the stock price sequence. Hence, the shapelets algorithm is introduced to match the feature subsequence of the stock sequence excavated in the early stage, which can enhance the accuracy and calculation speed.

In this paper, a prediction scheme of stock price trend based on model discovery is proposed based on the actual application scenario of stock operation, dynamic time warping (DTW) and shapelets data mining algorithm, and the forecasting results of the scheme are analyzed in detail.

**Keywords**: stock price; stock forecasting; forecasting technique; dynamic time warping; shaplet

目 录

[第1章 绪论 1](#_Toc24971_WPSOffice_Level1)

[1.1 研究背景及研究意义 1](#_Toc19017_WPSOffice_Level2)

[1.2 数据挖掘在股票预测的研究进展及现状 3](#_Toc16932_WPSOffice_Level2)

[1.3 本文的主要研究内容 6](#_Toc4308_WPSOffice_Level2)

[1.4 论文结构 6](#_Toc32251_WPSOffice_Level2)

[第2章 基于数据挖掘的股票分析方法的相关研究 8](#_Toc26817_WPSOffice_Level1)

[2.1 模糊时间序列模型在股票分析中的应用 8](#_Toc26537_WPSOffice_Level2)

[2.1.1 模糊时间序列模型的基本框架 8](#_Toc27509_WPSOffice_Level3)

[2.1.2 模糊时间序列模型的论域划分 10](#_Toc25594_WPSOffice_Level3)

[2.1.3 模糊时间序列的隶属度函数 11](#_Toc6449_WPSOffice_Level3)

[2.1.4 基于模糊时间序列模型的股票预测应用 12](#_Toc31773_WPSOffice_Level3)

[2.2 小波变换在股票分析中的应用 14](#_Toc6943_WPSOffice_Level2)

[2.2.1 小波变换 14](#_Toc13049_WPSOffice_Level3)

[2.2.2 小波神经网络 16](#_Toc26683_WPSOffice_Level3)

[2.2.3 小波变换在股票预测中的应用 18](#_Toc32693_WPSOffice_Level3)

[2.3 隐式马尔可夫模型（HMM）在股票分析中的应用 20](#_Toc19061_WPSOffice_Level2)

[2.4 本章小结 21](#_Toc20956_WPSOffice_Level2)

[第3章 时间序列特征提取算法的研究 23](#_Toc29470_WPSOffice_Level1)

[3.1 时间序列的定义 23](#_Toc23977_WPSOffice_Level2)

[3.2 时间序列相似性度量算法 23](#_Toc5384_WPSOffice_Level2)

[3.2.1 欧式距离 24](#_Toc9241_WPSOffice_Level3)

[3.2.2 动态时间归整(DTW) 25](#_Toc18637_WPSOffice_Level3)

[3.3 动态时间规整算法的优化 28](#_Toc20652_WPSOffice_Level2)

[3.3.1 全局约束 28](#_Toc21520_WPSOffice_Level3)

[3.3.2 快速动态时间规整（FastDTW） 30](#_Toc16914_WPSOffice_Level3)

[3.3.3 子序列动态时间规整 32](#_Toc29681_WPSOffice_Level3)

[3.4 时间序列特征提取（shapelets）算法 33](#_Toc32085_WPSOffice_Level2)

[3.4.1 shapelets算法介绍 33](#_Toc3092_WPSOffice_Level3)

[3.4.2 shapelets算法应用在股票价格预测上的可行性 35](#_Toc1857_WPSOffice_Level3)

[3.5 本章小结 36](#_Toc26344_WPSOffice_Level2)

[第4章 基于模式发现的股票价格走势预测方案 38](#_Toc15896_WPSOffice_Level1)

[4.1 整体股票价格走势预测方案的设计 38](#_Toc22798_WPSOffice_Level2)

[4.2 股票价格序列中长期走势状态的判定方案 41](#_Toc24016_WPSOffice_Level2)

[4.3 基于模式发现的股票价格走势特征序列提取方案 42](#_Toc26894_WPSOffice_Level2)

[4.4 基于股票价格走势特征序列的价格走势预测方案 42](#_Toc26723_WPSOffice_Level2)

[4.5 本章小结 42](#_Toc26940_WPSOffice_Level2)

[第5章 实验验证 44](#_Toc29807_WPSOffice_Level1)

[5.1 实验环境 44](#_Toc29327_WPSOffice_Level2)

[5.2 实验数据的选取 44](#_Toc27211_WPSOffice_Level2)

[5.3 代码实现 45](#_Toc27462_WPSOffice_Level2)

[5.4 实验结果分析 45](#_Toc14259_WPSOffice_Level2)

[5.4.1 预测性能 45](#_Toc10943_WPSOffice_Level3)

[5.4.2 模式挖掘结果 47](#_Toc13650_WPSOffice_Level3)

[5.5 本章小结 49](#_Toc25704_WPSOffice_Level2)

[第6章 总结与展望 50](#_Toc15593_WPSOffice_Level1)

[6.1 全文总结 50](#_Toc19338_WPSOffice_Level2)

[6.2 研究展望 50](#_Toc15081_WPSOffice_Level2)

[参考文献 52](#_Toc12282_WPSOffice_Level1)

[致 谢 57](#_Toc2154_WPSOffice_Level1)

[声 明 58](#_Toc5278_WPSOffice_Level1)

[个人简历、在学期间发表的学术论文与研究成果 59](#_Toc2815_WPSOffice_Level1)

第1章 绪论

本章首先介绍了本课题的研究背景及研究意义，然后对数据挖掘在股票预测中的研究现状作了简单介绍，最后阐述了本文的主要研究内容和论文结构

## 1.1 研究背景及研究意义

数据挖掘是在对历史数据统计分析基础上，通过统计算法提取隐含的有价值的数据特征，对未来预测提供支持的一种分析方法。数据挖掘算法经过多年的研究发展，形成了非常丰富的种类和分支，诸如KNN、贝叶斯、SVM、CNN、RNN等算法。并且在计算机技能能力的提高和互联网发展的支持下，数据挖掘技术从原先的小数据集向大数据集发展，从集中式计算向分布式计算发展，从静态数据计算向动态流式数据计算的方向发展。

时间序列是一种复杂的非结构化数据，广泛存在于各个领域，例如股票价格，商品销售情况，语音，天气情况，病人心跳等等。通过可视化的时间序列对事物的发展过程进行表述，有利于直观的表述事物的发展规律，并且能够对事物的发展规律进行量化。因此，通过对海量的时间序列数据进行分析，获取有价值的知识就有了非常重要的研究意义。

股票价格是一个二维时间序列，因此具备使用数据挖掘算法对股票价格进行挖掘的价值。而在股票交易领域存在量化交易分析这样一个分支，自从400多年前的1606年第一支现代意义上的股票在荷兰的东[印度](http://news.hexun.com/india/index.html)公司发行后，股票市场成为了金融领域的一个重要组成部分。在股票发展的过程中，投资界有各种各样的交易流派出现。而从上世纪80年代开始出现的量化交易成为了股票市场的一支后起之秀。而随着计算机技术的普及和发展以及现代金融理论的发展和交易成本的下降，量化交易系统得到了快速发展，量化交易系统经历了3个主要的发展阶段。

最初量化交易系统于20世纪80年代开始兴起，早期的量化交易系统主要以统计套利模型为主，比如1983年摩根士丹利大宗交易部门程序员格里·班伯格发现的“统计套利策略”，通过卖空价格高的股票，买入价格低的，在它们的价格恢复到历史均衡水平时平仓。还有同样诞生于80年代的“海龟交易法则”，这个法则诞生于著名的商品投机家理查德.丹尼斯与他的老友比尔.埃克哈特进行了一场辩论，即伟大的交易员是天生的还是后天培养的。海龟交易法则证明了对于一致性赚钱的交易，使用完全机械式的交易系统，抛弃人性中恐惧和贪婪的影响是最佳的交易策略。这个法则为后来的量化交易系统的发展提供了良好的理论支持。

进入20世纪90年代，量化投资进入了黄金发展的10年，在这期间标准金融理论形成，量化基金在成熟理论的武装下繁荣起来。期间产生了大量建立在不同数学模型基础上的程序化交易系统。在文献[1]中，作者介绍了金字塔交易决策系统，大智慧DTS程序化交易平台等国内比较知名的量化交易系统。在90年代，对于量化交易研究领域最重要的事件当属于James Simons创立的著名的量化基金公司Renaissance Technology，1989-2009的二十年间，其掌管的的大奖章（Medallion）对冲基金年均回报率高达35%[2]。虽然Renaissance Technology公司从来没有公布过他们使用的量化交易模型，但是市场通过对Renaissance Technology基金公司创立者和主要成员的学术背景中分析，认为该基金公司使用了隐式马尔科夫链这种数学模型。由此，隐式马尔科夫链模型在量化交易研究领域占据了重要的地位。

进入21世纪，随着计算机系统的快速发展和分布式系统的大规模应用，计算机系统的计算能力得到了大幅提升。使得应用基于大数据的数据挖掘算法的应用成为了可能。各种基于大数据的数据挖掘算法被广泛应用于各种领域，其中以应用于图像识别的人工神经网络的出色表现为标志，人工智能再次得到了广泛的关注。同样各种基于大数据的数据挖掘算法也被应用于量化交易的研究工作中，例如遗传算法[3]和人工神经网络[4]。

量化交易分为两种类型：一种是以统计学为基础的，通过对冲交易在不同的交易市场中寻找套利机会的套利模型。另一种是以数据挖掘和人工模型为基础的未来走势预测模型。最初的量化交易系统主要由熟悉金融交易市场的交易参与者构建，所以主要以统计套利模型为主，常见的对冲模型主要有：股票对冲(Equity Hedge)、事件驱动(Event Driven)、全球宏观(Macro)、相对价值套利(Relative Value)四种。后期随着数学家的加入，量化交易系统开始出现了以未来走势预测为主的模型，比如上面提到的隐式马尔科夫链。

国内由于股票市场出现的较晚，从1990年中国股票市场建立，到今天短短20多年的时间，中国的股票市场走过了西方股票市场上百年走过的路，国内的量化交易的相关研究在相当长的时机里也是非常落后的。因为股票量化系统对于股票交易决策有非常重要的指导作用，有非常好的经济效益，随着近年来国内互联网的大面积普及和计算机教育的推广，国内的量化交易研究出现了长足的发展。出现了类似京东量化平台（http://quant.jd.com/）、优矿量化平台（https://uqer.io/）、米筐量化平台（https://www.ricequant.com/）等量化交易平台。同时也出现了国内的开源量化软件，比如vnpy(https://github.com/vnpy/vnpy)和[DevilYuan](https://github.com/moyuanz/DevilYuan)（https://github.com/moyuanz/DevilYuan）。

## 1.2 数据挖掘在股票预测的研究进展及现状

早在1973年就有学者提出了关于量化交易系统的研究成果[5],该文章介绍了一种通过接受当前股票市场状态和投资者现有投资组合的信息，并将这些信息转化为合理和稳定回报率的投资决策，这个系统预测未来的价格行为，并估计风险，并将这些信息和个人股票组合被传递给决策模型，该决策模型被设计为满足连续投资组合管理的动态现实问题。机器人投资者包括预测者、决策模型、会计部分和自适应控制部分。这个早在1973年设计的系统已经具备了量化交易系统的全部要素。随着计算机技术的发展，量化交易领域的研究成为了计算机领域的一大热点，通过在web of science上搜索（TS=(stock AND market AND forecast) AND SU = computer science）发现，2014年有225篇相关论文期刊发表，2015年253篇，2016年285篇，2017年379篇，越来越多的数据挖掘算法被应用到股票预测领域，下面逐一介绍数据挖掘算法在股票预测方面的应用。

贝叶斯定理是一种根据不确定信息来做出对未来发生事件发生概率的推断方式，在18世纪，由英国学者[贝叶斯](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF/1405899)(1702～1763)提出。贝叶斯网络是一个基于贝叶斯定理的[概率推理](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E6%8E%A8%E7%90%86/9891299)的图形化网络，贝叶斯网络是为了解决不定性和不完整性问题而提出的，它对于解决复杂设备不确定性和关联性引起的故障有很大的优势，在多个领域中获得广泛应用。贝叶斯网络由一个有向无环图（DAG）和条件概率表（CPT）组成，有向无环图用来表示随机变量间的依赖关系，条件概率表用来表示互相依赖的参数。贝叶斯网络的优点在于允许学习变量间的因果关系和能够处理不完备的学习数据集，这对于股票预测具有非常重要的价值。在文章[6]中，作者通过贝叶斯网络验证了世界各国的股票指数对巴西圣保罗股票交易所主要交易指数iBOVESPA的影响情况。作者提出的模型给出了次日收盘方向和对iBOVESPA指数影响最大的一个因素，根据作者的实验数据，预测准确度可以达到71%。另外也有学者提出了基于贝叶斯推断的顺序蒙特卡罗方法对Lyapunov指数波动率进行推断的模型[7]，证明了在经济危机前后，存在嘈杂的混乱，但是当研究简单模型（不考虑共同收益和波动率之间的相互作用）时，混沌动力学的假设没有得到数据（“被忽略的混沌”）的大量支持。同时在原油期货市场，也有基于贝叶斯多链马尔可夫切换GARCH模型被提出[8]。

模糊时间序列的研究是发现不精确观测值时间序列内部的关联关系。虽然对于股票时间序列来说，观测值是准确的可量化数值。但是股票交易者并不关心股票价格数值本身，股票交易者更关心的是股票价格数值背后所代表的“暴涨”“暴跌”“中阳”“中阴”“横盘”这样的股票走势含义，因此模糊时间序列的研究对于股票走势的预测同样是有意义的。模糊时间序列是一种基于隶属函数的新的集合，这种集合被称为模糊集。通过隶属函数可以将某个元素映射到模糊集上，然后通过训练数据集建立模糊关系集合，由所有的模糊关系就得到了模糊关系矩阵，进而得到时间序列内部的关联关系的信息。模糊理论和模糊逻辑的概念最初是在1965年由美国自动控制专家Zadeh提出的，并建立了初步的语言处理模型[9].之后，有学者根据Zadeh提出的理论，建立了相应的预测模型[10-11]。近年来，诸多的研究人员提出了基于模糊时间序列的股票预测算法[12-14]，通过对模糊时间序列算法模型的改进，提高了对于股票走势预测的准确度，比如通过遗传算法和粒子群算法对隶属度函数中的参数值进行优化[15]，也有研究人员通过整合模糊集理论和其他时间序列挖掘算法预测方法，对股票价格进行预测取得了不错的效果[16-18]。另外，将k线图某些固定模式进行模糊化后的进行识别匹配相对于清晰版本的模式识别表现出了更好的稳定性[19]，能够比其他交易系统更小的风险获得更多的利润。

小波变换是基于小波基的时频域分析方法，小波变换最初是由J.Morlet在1974年首先提出的，之后数学家Y.Meyer偶然构造出一个真正的小波基，并与S.Mallat合作建立了构造小波基的统一方法加[多尺度分析](https://baike.baidu.com/item/%E5%A4%9A%E5%B0%BA%E5%BA%A6%E5%88%86%E6%9E%90)。相对于傅里叶变换只能分析频域信息，小波变换可以同时分析出信号的频率信息和频率信息出现的时间位置，因而小波变换更好的分析非稳态信号。股票价格时间序列就是一个非稳态信号，并且对于股票交易来说，价格信息的时域信号也同样重要，因此使用小波变换进行股票价格的预测有良好的效果。有文献[20]提出了一种基于小波分析的组合预测方法，以上证综合指数和深证综合指数日收盘价数据为样本的实验结果表明：新算法能较好地改善交易成本。因为小波变换能够提取股票价格序列的时频域的特征信息，所以有很多研究人员将小波变换和各种神经网络结合进行金融预测[21-23]，还有研究人员将上文提到的模糊时间序列理论与神经网络结合的模糊神经网络与小波分析技术进行结合，得到了不错的分析结果[24]。另外也有文章将小波分析与支持向量回归（SVR）结合进行股票预测分析[25-26]。从上面这些研究可以看出，因为小波变换对于股票价格这种非稳态的时间序列信号可以较好的提取时频域信息，所以将小波分析与各种不同的分类算法相结合都可以得到不错的分析结果。

马尔可夫模型是计算一个时间序列上发生事件间的转移概率的模型，进而可以通过当前的某个事件推算下个时间点各种事件发生的概率，该模型广泛应用于语音识别，词性自动标注，音字转换等领域，也可以被用于金融领域[27-32]。隐式马尔可夫模型包含两条随机过程链。一条是可以观测到的输出值，一条是不可见的事物内部状态链。在事物内部状态见具备相应的转移函数，在内部状态和输出值之间存在输出概率函数。对于股票市场来说，第一条链可观测输出值就是股票价格，成交量，换手率等等可以观测到的股票交易数据，第二条链就是股票市场内在的状态（牛市上涨、牛市下跌、熊市上涨、熊市下跌、震荡市上涨和震荡市下跌）[33]。有文献[34]采用了ANN算法来处理HMM的输入值，用POS方法来寻找更好的HMM初始参数，该文通过对美股市场价格的检验可以得到较好的收益效果。有文献[35]进一步融入了Fuzzy来处理股票价格时间序列的非平稳性。

通过上面介绍可以发现，随着数据挖掘技术的发展，各种各样的数据挖掘算法都被应用于股票预测中来。虽然各种不同算法的预测准确度有所差别，但是大部分算法的预测精度都在60%到80%之间，这主要是由于股票价格的波动受到了非常复杂的因素，比如政策变化、经济发展状况、人口数量、社会突发事件的影响，所以想要通过单一的预测技术很难非常准确的预测股票价格的走向。另外，由于股票价格具有反身性的特点[36-37]，所以当一个预测算法被广泛应用的时候，这个预测算法预测准确程度就会下降，最后会变成看谁的交易网络快的比赛。所以一个量化交易系统是由多个由不同算法构建的预测模块构成的，综合利用不同算法各自的优势来提高交易的收益率，并且随着时间的推移要更新算法，来领先于市场。

在研究现有算法应用于股票预测的过程中，我发现动态时间归整（DTW）较少被应用于股票价格预测，并且使用这种算法的预测准确度也不高[38-40]，这主要是由3方面原因造成的：

1. 股票价格走势拐点只出现在相对较短的一段时间内

股票走势大部分时间是平稳的。例如，在股票上涨趋势中，股票价格会一直上涨，区别仅仅是短期内上涨幅度大小的区别。而股票走势发生逆转仅仅于某一小段时间的走势有关，例如开头提到的头肩顶和w底走势。但是如果单纯使用DTW算法，需要将近期走势与过去全部的价格走势进行匹配，这无疑会引入大量的噪音数据，导致预测准确度的下降。在论文38中，作者通过金融领域对于股票走势的一些研究，引入了预先约定的一些规则来寻找股票走势中的特征，虽然这样做屏蔽了大量的数据噪音，但是同时也把潜在的股票价格序列中的特征片段忽略掉了。

1. 股票价格时间序列的数据量相对较大

股票的价格是从股票上市就一直连续的，如果对整个股票价格时间序列进行匹配的话，计算量会非常大，在计算时间上不现实。而股票交易是强调实时性的，过慢的计算结果会是预测的实际价值大大折扣

1. DTW算法算出的两个序列距离与挖掘结果的转换关系不同

在量化交易系统中，需要输出的是买入点和卖出点，而为了保证盈利能力，需要对匹配结果和输出做出一定的转化。DTW算法只是计算两段时间序列距离的一种算法，而在股票预测这个实际应用场景中，取哪段序列进行匹配，而匹配的结果和后续的走势又有一个什么样的关系是需要进行判断的。

在一些研究员的研究中，将DTW算法与在股票预测中常用的特征抽取算法结合，将DTW算法应用于特征与当前数据间距离的计算有不错的预测准确度[41]。

## 1.3 本文的主要研究内容

本文围绕模式发现算法应用于股票价格趋势预测这一课题进行了研究和实践。

本文首先对现有应用于股票预测的主要数据挖掘算法进行了研究。通过研究现有挖掘算法应用于股票预测这种实际应用场景，探讨了股票时间序列本身的特点，不同数据挖掘算法应用于股票预测时的优缺点，以及数据挖掘算法应用于股票预测的一般过程。研究过程中特别关注了与DTW算法能够结合并达到不错预测准确度的小波网络算法。

之后本文对于一般意义上的时间序列以及时间序列的数据挖掘算法进行了探讨。在这一部分首先研究了时间序列相似性度量的主要算法，并着重对动态时间归整（DTW）算法进行了研究，并对DTW算法的一些优化算法进行了进一步的研究。之后研究了时间序列特征的提取算法，主要讨论了基于时间序列的shapelets算法，并对该算法在股票预测的可行性上进行了研究。

最后本文提出了一个基于模式发现的股票价格走势预测方案，使用python语言实现了该方案，并使用上证50成分股的历史数据对该方案的预测准确度进行了实验验证。

## 1.4 论文结构

本文主要分为以下3个部分：

1. 现有股票市场模型及分析理论的研究。在第二章主要对现有的市场量化分析模型进行介绍。从模糊时间序列模型，到后来的小波网络，包括因为量化基金公司Renaissance Technology而在量化交易模型中被广泛使用的的隐式马尔科夫模型。
2. 动态时间规划（DTW）和基于时间序列的shapelets算法的研究。在第三章中，对本文将要用到的两种算法进行了介绍。动态时间规划算法是用于语音识别的一种序列匹配算法，因为其良好的鲁棒性而被广泛使用，并且发展出了多种优化算法。基于时间序列的shapelets算法最初是被提出用于图像识别的一种算法，根据信息熵对形状片段的区分度进行分析。
3. 在第四章中，提出本文应用于股票价格分析预测的基于模式发现的股票价格走势预测方案。并在第五章中通过上证50的成分股指数的历史数据对该算法的应用效果进行检测。

第2章 基于数据挖掘的股票分析方法的相关研究

随着近年来数据挖掘技术在不同领域的出色表现，将数据挖掘技术应用于股票走势的分析已经成为了一个重要的研究方向。在股票走势分析中，通过分析股票价格序列进行股票走势的预测是一个重要的研究方向，这主要是因为股票价格序列获取容易并且与预测内容高度相关。时间序列数据大量存在于各行各业，因此对于时间序列数据的挖掘已经有很多成熟的算法，大量的研究表明：通过对股票价格的时机序列进行建模并通过分析历史数据中股票价格序列中的特征与股票走势的关系可以较好的预测股票的价格走势。本章将介绍模糊时间序列模型、小波网络模型、隐式马尔科夫模型（HMM）在股票分析中的应用，并在此基础上进行评述。

## 2.1 模糊时间序列模型在股票分析中的应用

证券市场非常复杂，股票价格的走势受到了社会经济发展状况、政策变化、文化背景、行业技术发展、社会突发事件等等因素的影响。因此，在股票价格时间序列建模中很难在精确的股票价格记录和股票走势间挖掘关系。而通过将股票价格的精确值与模糊的对股票价格走势判断的语言变量建立模糊逻辑联系，可以更好的描述股票价格时间序列的特征，最后再将模糊的预测结果转换成对股票价格值的预测。这一方法可以更好的描述股票价格序列的特征，可以一定程度的提高预测精度。

### 2.1.1 模糊时间序列模型的基本框架

模糊时间序列模型研究的基础是由Song[10-11]提出的模糊时间序列模型。模糊时间序列模型在精确的数值与模糊的语言变量间建立关系，之后通过训练数据集挖掘模糊预测关系矩阵的参数。这里先对这种模型的定义做简单的介绍，以便后续描述模糊时间序列在股票预测中的应用。

定义2.1 在论域U中，定义一个U的顺序分割子区间,对于任一定义在论域U中的模糊集A（即语音变量集）表示如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-1） |

这里的加号表示的是一种集合关系而非计算意义上的加。即是子区间对于模糊集合A的隶属度函数，表示了子区间多大程度上属于模糊集A，。

定义2.2 对于时间序列，定义为实数域的论域，设为定义在上的模糊集合，则为定义在上的模糊时间序列。这里就是可能的语言变量，则是关于t的函数。

定义2.3 定义一阶模糊关系为R，则为定义在到的模糊关系，满足，也就是可以仅仅由得到，可以用模糊关系来表示。这里“\*”代表合成运算，被称为模糊关系的前件，被称为模糊关系的后件。

定义2.4 由定义2.3可以推论，n阶模糊关系可以表示为，也就是说的状态可由它的n个前置状态得到。

定义2.5 假设有n个论域。对于某一论域的模糊值可由n个论域的t-1的模糊值（）取得，则具有n个前件的一阶模糊关系可以表示为

定义2.6 综合定义2.5和定义2.6，可以得到具有n个前件的m阶模糊关系可以表示为：对于为定义在各自论域上的n个模糊时间序列，

根据上文描述模糊时间序列的基本定义，模糊时间序列的构建和基础预测步骤一般分为六步：



1. 根据训练集和其他相关设定对论域进行模糊划分。
2. 定义模糊集合A，并通过训练数据集计算模糊关系的隶属度函数。
3. 定义模糊关系矩阵，并通过训练数据集计算得到模糊关系矩阵。
4. 将测试数据集的数据根据步骤1和2获得的模糊划分和隶属度函数求得模糊集。
5. 根据步骤3获得的模糊关系矩阵计算步骤4中获得的模糊集，获得预测的模糊集
6. 根据步骤1和2获得的模糊划分和隶属度函数将步骤5求得的模糊集还原成精确值区间。

### 2.1.2 模糊时间序列模型的论域划分

模糊时间序列构建的第一步是论域划分，这一步对于预测模型的结果影响很大。在下面介绍的这些论域划分的方法中，可以发现，根据目标数据在论域上的分布特点，选择合适的论域划分方法在模糊时间序列模型的构建过程中十分重要。

1. 等间隔论域划分方法

在Song等人提出的模糊序列模型[10-11]中，采用的就是等间隔论域划分方法。

对于论域U，先取目标数据的最大值和最小值，然后根据模糊集（例如语言变量）定义分割区间个数k，可以计算出分割间隔：

|  |  |  |
| --- | --- | --- |
|  |  | （2-2） |

根据分割间隔L可以得到分割子区间，其中，对于每一个区间：。

等间隔论域划分方法的优点是模型简单，易于理解。但是因为等间隔划分方法的间隔是主观规定的，与研究数据没有任何关系，也就不能正确反映目标数据的变化。

1. 基于比率的论域划分方法

在基于比率的论域划分方法[42]中，使用相对误差和比率来反映时间序列的波动。在这种论域划分方法中，比率体现了模糊时间序列对原有序列值的拟合程度，比率越小，模糊时间序列与原有时间序列越相似。所以如果比率太大，则模糊时间序列无法反映原有时间序列的波动状况，如果比率太小，则模糊时间序列与原有时间序列过于相近，丧失了模糊的意义。一般来说比率选择为50%。具体的算法如下：

1. 计算相邻数据的相对误差值：

|  |  |  |
| --- | --- | --- |
|  |  | （2-3） |

1. 从所有相对误差中取最小值，查询文献[42]中的基数表获取分布图的步长，绘制分布图
2. 确定比率（ratio），这个值一般来说取分布图中50%对应的值
3. 计算初值：，计算间隔：。可以得到论域U的划分，其中

基于比率的论域划分方法考虑到了目标数据的变化，能够一定程度的提高模糊时间序列模型的预测精度，但是对于分布不均匀的数据，可能会导致模糊后的数据某些区间没有数据，在这种情况下这种论域划分方法不能正确反映数据的分布特征。

1. 多尺度比例论域划分方法

多尺度比例论域划分方法是基于FCM聚类划分方法[44],FCM是一种使得被划分到同一簇的对象之间相似度最大，不同簇之间相似度最小的分类算法。通过使用FCM算法对论域进行划分，对每类数据使用不同的划分比率，根据分布的疏密不同进行划分，充分考虑不同数据的分布情况差异，可以有效提高划分论域的精度。具体划分方法如下[43]：

1. 对样本数据集进行从小到大排序，排序后的序为：

（是最小值，是最大值）

1. 使用FCM算法对样本数据进行分类，将样本数据集分成m个数据集:



1. 分别计算每个数据集的相邻数据的相对误差：

|  |  |  |
| --- | --- | --- |
|  |  | （2-3） |

1. 计算每个数据集的平均相对误差：

|  |  |  |
| --- | --- | --- |
|  |  | （2-4） |

1. 根据数据集，定义每个数据集对应的论域：，和是使论域向下向上取整的适当整数
2. 定义初值：,定义间隔：。可以得到论域U的划分，其中

### 2.1.3 模糊时间序列的隶属度函数

隶属度函数是模糊时间序列模型中的基础，是应用模糊时间序列模型的关键。隶属度函数表达的是精确值对模糊集从属程度的高低。本质上讲，隶属度函数应该是客观的。但是实际应用中，每个人对于模糊集的理解是不同的，因此对于同一个模糊集不同的研究者建立的隶属度函数模型可能是不同的，所以隶属度函数的确定过程带有一定的主观性。确定隶属度函数主要有如下几种方法。

1. 模糊统计法：通过设计实验，统计同一个精确值属于不同模糊集的概率来计算隶属度函数。

|  |  |  |
| --- | --- | --- |
|  |  | （2-5） |

1. 专家经验：利用专家的实际经验确定隶属度函数。因为人的经验往往具有局限性，所以实际应用中往往先确定大概的隶属度函数，然后根据实验效果调整隶属度函数的参数。
2. 例证法：对于某一个论域，先确定几个语言真值，然后分别用数字参数来对应这些语言真值。最后对某一个确定值与模糊集的关系用语言真值进行询问，得到离散表示的隶属度函数
3. 二元对比法：通过对不同事务间两两对比来确定某个精确值与模糊集的特征顺序，由此来决定隶属度函数。
4. 模糊分布：论域是实属域时，隶属函数称为模糊分布。常见的模糊分布有：正态分布，三角分布，梯形分布，矩形分布或半矩形分布，岭形分布。

### 2.1.4 基于模糊时间序列模型的股票预测应用

有学者[13]在B-L模型中使用模糊视图构造预测模型，并重新定义了视图的期望收益和不确定性矩阵。该模型使用中国证券市场的CSI指数作为样本数据集，使用2005年1月4日到2012年6月11日的数据作为训练数据集，使用2012年6月12日到2014年6月11日数据作为测试数据集，实验表明模糊模型可以更好的表达时间序列中的不确定信息。

同时，也有研究表明，直接将模糊时间序列模型应用到股票趋势的预测中也有不错的效果[16].有文献[17]采用了聚类C-均值的算法对样本数据进行论域划分，并将模糊集合定义为划分后的子区间。某个精确值对应的模糊值既为其落在的子区间，这里可以发现实际是一个聚类算法使得分散的精确值转换为几个聚类核心。之后作者建立了一个4前件的3阶模糊关系模型，通过训练数据训练构造一个预测模型。该文章的实验过程中，选用了上证和深成指数中的每日收益率，最高/最低价、成交量、收盘价作为样本数据集。使用的股票数据时间为2005年12月14日到2009年4月15日。作者用上述模型对这些数据中前750交易日的数据作为训练数据集，后50个交易日的数据作为测试数据集，并将预测结果与经典的模糊序列模型[45]的预测准确度进行了比较，如表2-1所示。

表2-1 杨一文文献中模糊序列模型预测结果[17]

|  |  |  |  |
| --- | --- | --- | --- |
| **指数** | **模型** | **平均相对误差/(%)** | **趋势预测正确率/(%)** |
| 上证指数 | Chen的模型[45] | 2.47 | 54 |
| 新的的模型[17] | 1.18 | 72 |
| 深证指数 | Chen的模型[45] | 2.68 | 59 |
| 新的的模型[17] | 1.34 | 68 |

## 2.2 小波变换在股票分析中的应用

小波变换因为能够在时频域更好的分析时间序列的特征，所以对于股票走势预测这种非常强调事件发生顺序的应用场景下，能够表现出良好的分析效果。这就使得小波变换被广泛应用于股票预测中。

### 2.2.1 小波变换

小波变换的概念是为了解决傅里叶变换中无法展示时域信息的问题提出的，最初是由J.Morlet在1974年首先提出的，之后数学家Y.Meyer偶然构造出一个真正的小波基，并通过与S.Mallat合作建立了构造小波基的统一方法加[多尺度分析](https://baike.baidu.com/item/%E5%A4%9A%E5%B0%BA%E5%BA%A6%E5%88%86%E6%9E%90)，完善了小波变换的概念，使得小波变换成为了一种可以实际应用的概念。

傅里叶变换可以清晰的反映时间序列包含的频率信息，但是并不能反映各个频率信息出现的时间。因此傅里叶变换适合分析相对平稳的信号，但是在现实应用中，时间序列往往是非平稳信号。如图2-2所示，在股票走势信息中不但需要知道包含了哪些频域信息，还需要知道这些频率信息出现的位置，这对于股票走势的预测至关重要。

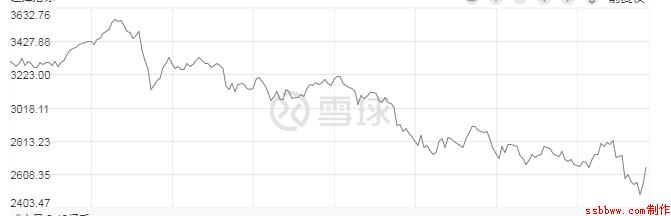


图2-2 上证指数一段时间的日k线图

对于这个问题，傅里叶变化的一个变种“短时傅里叶变换”可以在一定程度上了解到频率信息在时间序列中出现的位置。短时傅里叶变换简单说就是在时间序列上加窗，然后在每个窗内进行傅里叶变换，这样就可以获得频域信息在时间序列上的位置。短时傅里叶变换的主要缺陷在于窗的宽度选择非常困难。如果窗选择的太大，那么分析结果中的时域信息就不够精确，如果选择的窗太小，那么分析结果中的频域信息就不够精确。而且由于实际的时间序列中的平稳程度是变化的，而窗的大小是固定不变的，所以使用短时傅里叶变换分析时间序列时，窗大小的选择非常困难，并且分析的结果很难在频域信息和时域信息的精确度上同时达到比较精确的程度。

小波变换正是为了解决时间序列分析时频域分析而提出的算法。小波变换的公式如下

|  |  |  |
| --- | --- | --- |
|  |  | （2-6） |

其中指代的就是小波基的伸缩尺度、指代的是小波函数的平移量，也就是小波基在时域上的信息。小波基如图2-4所示。通过小波变换可以将时间序列分解成在不同平移量上的不同伸缩度的小波基的组合，也就是得到了时间序列的时频域信息，如图2-5所示。另外，小波变换对于突变信号的拟合也要好于傅里叶变换。

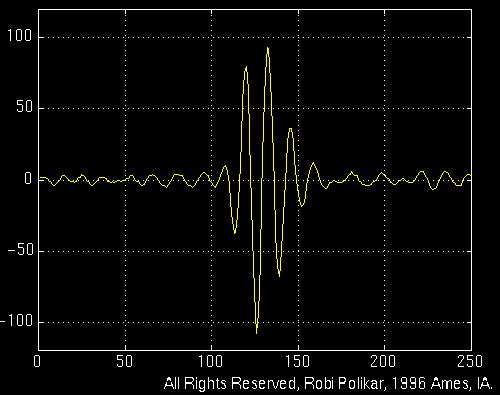


图2-4 小波基[67]

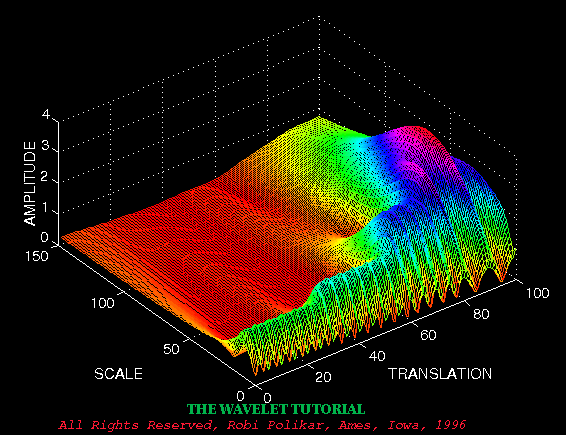


图2-5 小波变换的时频域信息[67]

### 2.2.2 小波神经网络

小波网络因为具有良好的鲁棒性和自学习性，所以成为了小波变换在股票预测中经常用到的一种数据挖掘技术组合。神经网络是模拟动物神经系统的算法模型。神经网络的基本单位是神经元，神经元是一个sigmoid函数，该函数通过接收n个输入并产生一个输出，这就类似动物的神经元细胞一样。神经网络通过多层的神经元组成，包含了输入层、隐藏层、输出层，如图2-6所示。神经网络的学习过程就是通过梯度下降算法求得在样本数据中最拟合的sigmoid函数。

输出

输入

图2-6 神经网络模型

小波神经网络即是将小波变换算法与神经网络算法结合的一种数据挖掘算法，主要的结合方式有两种：

1. 将时间序列数据进行小波变换，将小波变换后的时频域特征值作为前馈神经网络的输入进行训练和预测分类，这种形式的小波网络被称为松散的结构，如图2-7所示。

小波变换

前馈神经网络

输入

输出

图2-7 松散的小波神经网络

1. 神经元的sigmoid函数使用小波变换函数，这种形式的小波网络因为将小波变换和神经网络深度的结合在了一起，因此被称为紧凑的结构，如图2-8所示。

输出

输入

小波变换函数

图2-8 紧凑的小波神经网络

因为紧凑的小波神经网络能更好的避免普通神经网络训练过程的盲目性问题，让学习过程更快的收敛，所以目前研究中更多被采用的是紧凑的小波神经网络。

### 2.2.3 小波变换在股票预测中的应用

小波变换因为能够提取时间序列的时频域信息，所以在股票时间序列预测中得到了广泛的应用，例如文献[63]中作者将小波变换的特征提取值作为AR预测模型的输入来进行预测的模型应用到上证指数的预测中，得到了不错的预测效果，如表2-2所示。

表2-2 使用小波预测上证指数的均方根误差[63]

|  |  |  |
| --- | --- | --- |
| **预测方法** | **小波分析预测方法** | **AR模型预测方法** |
| 预测误差均方根值 | 0.3880 | 2.0912 |

如前文所述，小波网络因为其鲁棒性好，更快收敛的特性在股票预测上也有不错的表现，有文献[64]将小波网络模型应用于华鲁恒生这支股票2006年7月1日倒2006年7月31日数据作为训练集，用该股票2006年8月8日到2006年8月20日的股票作为测试集，获得了不错的预测效果，如图2-9所示。

图2-9 小波网络在华鲁恒生股票上的预测效果[64]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **预测效果评价指标** | | **SSE** | **MAE** | **MSE** | **MAPE** | **MSPE** |
| 个体预测 | 二次指数平滑法 | 29.6553 | 0.9685 | 0.272283 | 0.094585 | 0.0271855 |
| 最小平方法 | 83.4823 | 1.7885 | 0.456843 | 0.17421 | 0.04468334 |
| 组合预测 | 梯度算法 | 6.1573 | 0.12407 | 0.05263 | 0.008354 | 0.018525 |
| 基于自构造算法的小波网络 | 0.016639 | 0.18428 | 0.002377 | 0.003416 | 0.003878 |

有学者将基于小波变换基础上提出的小波包与神经网络结合，提出了基于小波包和神经网络的股票预测模型，将预测模型应用到上证指数的预测中，并与神经网络和小波网络的预测结果进行了比较，显示出了较大的改进，如图2-10所示。

图2-10 基于小波包的神经网络的预测精度比较[65]

|  |  |
| --- | --- |
| **模型名称** | **预测精度MSE** |
| NN | 1360.1 |
| WNN | 906.5 |
| WPNN | 653.6 |

有文献[25]将小波变换与经典的支持向量机（SVM）算法相结合应用到了国内外的股票价格预测中也同样得到了不错的效果，如图2-11和图2-12所示。

图2-11 基于小波变换和支持向量机对大盘指数的预测效果[25]

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **股票指数** | **高斯核心** | |  | **Morlet小波核** | |  | **Marr小波核** | |  | **DOG小波核** | |
| **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |
| 道琼斯 | 7.81e-5 | 99.98% |  | 0.0016 | 99.76% |  | 0.0015 | 99.79% |  | 0.0044 | 99.78% |
| 纳斯达克 | 0.0062 | 98.53% |  | 0.0034 | 99.35% |  | 0.0049 | 99.37% |  | 0.0163 | 99.35% |
| 上证180 | 0.0382 | 91.23% |  | 0.0094 | 99.05% |  | 0.0086 | 99.29% |  | 0.01 | 99.31% |

图2-12 基于小波变换和支持向量机对个股指数的预测效果[25]

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **股票指数** | **高斯核心** | |  | **Morlet小波核** | |  | **Marr小波核** | |  | **DOG小波核** | |
| **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |  | **均方误差** | **相关系数** |
| 万科A | 0.0081 | 98.96% |  | 0.0027 | 99.68% |  | 0.0022 | 99.68% |  | 0.0054 | 99.38% |
| 美的电器 | 0.0578 | 77.57% |  | 0.0099 | 99.26% |  | 0.0089 | 99.43% |  | 0.0086 | 99.14% |
| 苏宁电器 | 0.0161 | 97.51% |  | 0.0099 | 98.87% |  | 0.0092 | 98.90% |  | 0.0108 | 98.79% |
| 浦发银行 | 0.0935 | 60.83% |  | 0.0346 | 97.82% |  | 0.0377 | 97.86% |  | 0.0335 | 97.77% |
| 首创股份 | 0.0315 | 83.42% |  | 0.0173 | 99.03% |  | 0.0094 | 99.08% |  | 0.0146 | 98.51% |
| 中国联通 | 0.1302 | 94.42% |  | 0.1003 | 97.03% |  | 0.0931 | 97.75% |  | 0.0520 | 98.17% |
| 葛洲坝 | 0.0271 | 94.02% |  | 0.0257 | 99.27% |  | 0.0795 | 99.34% |  | 0.0832 | 97.68% |
| 中青旅 | 0.2875 | 13.17% |  | 0.0028 | 99.34% |  | 0.0028 | 99.37% |  | 0.0076 | 99.05% |
| 北大荒 | 0.0344 | 90.73% |  | 0.0235 | 98.69% |  | 0.0209 | 98.45% |  | 0.0183 | 98.37% |
| 中国中铁 | 0.0060 | 96.68% |  | 0.0061 | 96.9% |  | 0.0049 | 97.04% |  | 0.0052 | 96.67% |
| 招商轮船 | 0.0063 | 98.43% |  | 0.0088 | 99.25% |  | 0.0178 | 98.77% |  | 0.0135 | 98.58% |
| 紫金矿业 | 0.0057 | 99.53% |  | 0.0058 | 99.52% |  | 0.00412 | 99.43% |  | 0.0027 | 99.53% |
| 广电传媒 | 0.0195 | 96.58% |  | 0.0161 | 98.53% |  | 0.0092 | 98.20% |  | 0.0237 | 97.56% |

## 2.3 隐式马尔可夫模型（HMM）在股票分析中的应用

隐式马尔可夫模式是通过观察到到参数链来分析一个由隐含参数组成的马尔可夫过程的统计模型。隐式马尔可夫模型假设现实中的观测值序列是由隐藏在背后的一套隐含状态决定的，例如对于股票的走势，假设股价的走势是由背后“大势上涨”、“大势下跌”、“大势震荡”的大盘状态来决定的。隐式马尔可夫模型有两条链组成：隐含状态链和可见状态链。隐含状态链中的隐含参数具有一定的转换概率，例如“大势上涨”转为“大势下跌”的概率是20%，“大势上涨”转为“大势震荡”的概率是30%，“大势上涨”转为“大势上涨”的概率是50%。每个可见状态链中的观测值是由隐含参数决定的，因此每个隐含参数对应一个观测值都有一个输出概率，例如“大势上涨”对应股价上涨5%－10%的概率是30%、股价上涨0%－5%是40%，股价下跌0%－5%的概率是20%，股价下跌5%－10%的概率是10%。每个隐式马尔可夫模型都是由隐含参数、隐含参数间的转换概率、隐含参数对可见状态的输出概率、可见状态链这四部分组成。在实际应用中，总是有部分组成部分是不知道的，隐式马尔可夫模型解决的问题就是通过训练数据求得不知道的部分参数来对未来进行预测。

对于股票趋势预测应用，隐含参数可以通过假设获得，可见状态链也有历史股票数据，需要求得的是隐含参数间的转换概率。隐式马尔可夫模型在各种编程语言下都有成熟的包可供调用，如图2-13是用隐式马尔可夫模型对上证指数2004年到2017年指数的建模，其中可以看到模式1就是“大势震荡”、模式2是“大势下跌”、模式3是“大势上涨”。

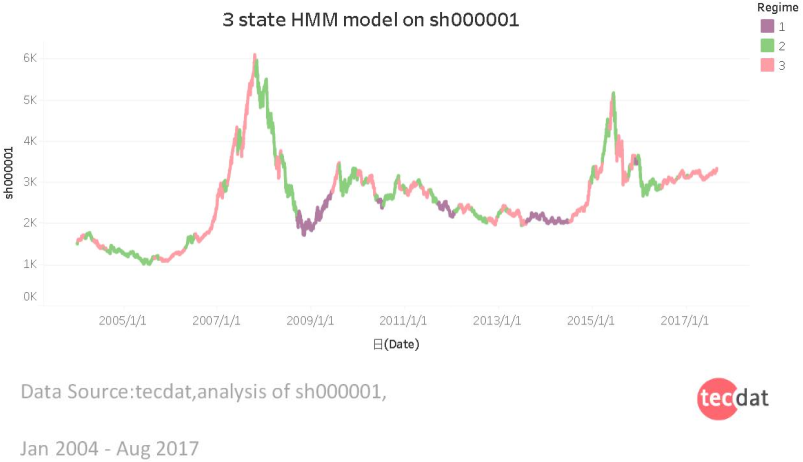


图2-13 隐士马尔可夫模型对上证指数的预测[66]

## 2.4 时间序列特征提取（shapelets）算法

对于股票走势来说，单一股票的各种指数的时间序列非常长，但是只有其中的局部的特征序列会反复出现。因此在股票走势预测中，需要在长时间序列中挖掘出有良好预测效果的局部特征序列。本节将介绍一种时间序列特征子序列shapelets提取算法。

### 2.4.1 shapelets算法介绍

基于时间序列的shapelets的分类算法最初是由Lexiang Ye[59,60]等人提出的，该算法最初用于图像识别领域，通过计算任意一个子序列对于样本数据分类的信息熵增益来判断该子序列是否是特征子序列shapelets。最终将筛选出的特征子序列shapelets构建成一棵决策树来实现图像分类。实验表明[59]该算法对于图像分类中经典的Gun/NoGun分类问题有较好的分类效果，相对于基于欧式距离的最近邻分类算法的分类精度91.3%、无约束动态时间规整算法的分类精度90.7%、约束动态时间规整算法的分类精度91.3%，使用基于时间序列shapelets的决策树分类算法分类精度可以达到93.3%，并且该分类算法要比基于欧式距离的最近邻分类算法计算速度也要快4倍。

基于时间序列的shapelets算法是一种使用决策树进行分类的分类算法。如图3-11所示，构建yes／no的决策树，通过shapelets特征子序列与被分类数据的相似性来对数据进行分类。对于多类型的数据集，通过构建多层决策树来进行分类，如图3-12所示。

Leaf Decison Tree

I

0

1

yes

no

图3-11 shaplet算法决策树[59]

Arrowhead Decison Tree

I

1

II

0

2

图3-12 多类型数据多shapelets决策树[59]

基于时间序列的shapelets算法是通过信息熵增益指标来判断特征子序列的。信息熵是1948年由香农提出用于度量事物信息量的指标。信息熵越高表明事物的不确定性越高，对应到样本数据上的表现就是每个分类的数据分布越平均。信息熵的计算公式是：

|  |  |  |
| --- | --- | --- |
|  |  | （3-7） |

其中指代分类集中某个事物发生的概率。对于基于时间序列的shapelets算法，该算法中决策树只有是或者不是两个分类，因此该算法中的信息熵是：，其中A和B分别指代决策树中是或者不是的分类数据。

信息熵增益是指的通过分类之后信息熵减少的值，减少的越多说明事物的确定性增加的越多，也就代表分类越精确。这里假设对于某个截取的子序列S，根据每个时间序列与子序列间的相似性差异将原有的数据集分为和两个数据集，那么用和分别表示和所占总样本数的百分比，即

|  |  |  |
| --- | --- | --- |
|  |  | （3-8） |

那么对于分类后的两个数据集来说信息熵，对于整个分类过程的信息增益.

shapelets特征子序列就是所有可能子序列中信息熵增益最大的那个子序列。所以筛选子序列的过程就是在数据集中每个时间序列中选取的可能的子序列计算其分类的信息增益，最终选择其中信息增益最大的那个子序列作为决策树的特征子序列。整个基于时间序列的shapelets算法可以总结如下：

1. 计算数据量的信息熵
2. 顺序选取某个时间序列的一个子序列s
3. 计算每个时间序列与该子序列的相似度，并根据预先设定的阀值将数据集中的所有时间序列分成两个数据集和
4. 计算分类后的数据集和的整体信息熵
5. 计算信息增益
6. 如果当前的信息增益是所有已计算子序列中最大的，则选择当前子序列为特征子序列
7. 判断是否所有可能子序列都被计算过，如果没有则重复至步骤（2）

### 2.4.2 shapelets算法应用在股票价格预测上的可行性

基于时间序列的shapelets算法是一种能够筛选出时间序列中对数据分类具有最大区分度的一个特征子序列的算法。对于股票市场，我们可以发现在价格时间序列中某些相似的价格走势总是重复出现，如图3-13所示。因此通过使用基于时间序列的shapelets算法具有将股票价格时间序列中具有区分一定走势（上涨、下跌、横盘）的特征时间子序列挖掘出来的可能性。



图3-13 股票k线图

使用基于时间序列的shapelets算法挖掘股票价格时间序列的问题在于这种算法需要计算每个可能子序列的信息增益，而计算信息增益过程中需要计算每个子序列与所有时间序列的相似度，因此计算的时间复杂度相对较高。而国内A股市场有3000多支股票，自1990年建立后，积攒了大量的股票价格时间序列，所以使用该算法需要的计算时间可能非常长。对于该问题，首先已经有一些学者提出了对于基于时间序列的shapelets算法的改进算法来提升其计算速度。例如通过剪枝等策略降低该算法的时间复杂度[61]，还有通过SAX表示的快速算法[62]。另外虽然股市中股票数量众多，但是某些股票更具有代表性，并因此组成了例如上证50、沪深300这样的指数，这样的股票往往对于所有股票的走势更具有代表性，所以通过缩小股票的选择范围可以降低数据量同样可以提高运行速度。

## 2.5 动态时间规整(DTW)算法

在基于时间序列的shapelets算法中使用了欧式距离作为相似性度量算法，在图像分类问题上，因为可以通过调整图像的像素来使得图像的边缘长度一致。但是在股票特征序列匹配上，因为股票的价格走势在时间维度上存在相似的走势有拉伸和压缩的情况，所以需要采用动态时间规整（DTW）来提高识别率。

### 2.5.1 动态时间规整(DTW)

由于欧式距离的计算方法是对两个时间序列按照时间顺序计算距离然后计算总和。因此对于有位移和扭曲的时间序列，使用欧式距离的计算方式往往会得到一个非常大的距离，而实际上这两个时间序列是非常相似的。例如图3-1所示，可以看到两个时间序列实际非常相似，但是上面的序列在时间上要比下个序列晚一个单位时间，但是如果使用欧式距离计算会得到一个比较大的差异。

图3-1 时间相差1的两个时间序列

在实际应用场景中，因为各种外力的干扰，对于同一事件不同次的取样时间序列往往会有轻微的位移和扭曲。例如在语音数据中，因为朗读者语速、语调的变化，对于同一单词发音的声音时间序列会有一定的变形和位移，但是实际表示的同一个单词，这时如果使用欧式距离来计算相似度会导致错误的分类结果。为了应对这种问题，产生了动态时间规整算法（Dynamic Time Warping, DTW）。

在动态时间规整算法中，在保证任意两个事件发生顺序不变的前提下，允许两个时间序列在时间轴上弹性的移动和变形，来计算两个时间序列一条最短的对齐距离，这种方法可以更好的测量两个时间序列的相似性。动态时间规整算法最初被成功应用于语音识别领域[50,51]，但是因为动态时间规整算法的时间复杂度是O(MN)，所以在处理长时间序列的相似性度量问题时，会收到计算能力的限制而影响应用效果。但是随着计算机硬件技术的发展，计算机的计算能力得到了大幅提高，动态时间规整算法也广泛应用于其他领域[52-54]，同时也有学者提出动态规整的改进算法以提高预测精度[55]。下面简单介绍一下动态规整的经典算法。

假设有两个时间序列和。首先构建构建一个m\*n的矩阵D，对于矩阵任意一点的值，该矩阵称为时间序列的距离矩阵。如图3-2展示的是一个m=6,n=6的时间序列的距离矩阵。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x |  |  |  |  |  |  |  |
| 3 | 2 | 2 | 2 | 2 | 2 | 2 |  |
| 4 | 1 | 3 | 1 | 3 | 1 | 3 |  |
| 1 | 4 | 0 | 4 | 0 | 4 | 0 |  |
| 3 | 2 | 2 | 2 | 2 | 2 | 2 |  |
| 1 | 4 | 0 | 4 | 0 | 4 | 0 |  |
| 4 | 1 | 3 | 1 | 3 | 1 | 3 |  |
|  | 5 | 1 | 5 | 1 | 5 | 1 | y |

图3-2 距离矩阵

通过构建距离矩阵，求两个时间序列最短距离的问题就转换成了对一个有权矩阵求从(1,1)点到(m,n)点的最短路径问题。这里首先要有三个约束条件。

定义3.3 对于时间序列和的一条扭曲路径必须满足如下三个条件：

1. 边界约束：路径的起始点，路径的结束点
2. 单调性：对于扭曲路径中的任意两个点和，如果，则
3. 步长约束：对于扭曲路径中任意两个连续的点和，并且

在定义3.3的约束下，可以通过动态规划的方法，将求解的问题分解为一个局部问题。这里假设对于距离矩阵中点的最短路径值是，那么就有如下定义：

定义3.4 定义距离矩阵中点的值是，那么对于距离矩阵中到任意一点的最短距离是

|  |  |  |
| --- | --- | --- |
|  |  | （3-3） |

同时，。另外根据约束条件也可以得出

|  |  |  |
| --- | --- | --- |
|  |  | （3-4） |

根据公式3-3，可以从距离矩阵的点开始计算每个点到点的最短路径，这个矩阵就是积累代价矩阵（Cost Matrix），以图3-2为例可以得到对应的积累代价矩阵如图3-3所示。在计算过程中可以得到最短路径，最终距离是6。这里与欧式距离计算的结果相同，实际上欧式距离就可以看作是动态时间规整算法的一个特例：当m=n时，限制了路径中任意一点中。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x |  |  |  |  |  |  |  |
| 3 | 14 | 8 | 6 | 6 | 6 | 6 |  |
| 4 | 12 | 6 | 4 | 6 | 4 | 7 |  |
| 1 | 11 | 3 | 7 | 3 | 7 | 7 |  |
| 3 | 7 | 3 | 3 | 5 | 7 | 9 |  |
| 1 | 5 | 1 | 5 | 5 | 9 | 9 |  |
| 4 | 1 | 4 | 5 | 8 | 9 | 12 |  |
|  | 5 | 1 | 5 | 1 | 5 | 1 | y |

图3-3 积累代价矩阵

在更一般的情况下，通过动态时间规整算法计算的两个时间序列的距离是要大大小于欧式距离的计算结果的，对应路径也存在一定的扭曲。通过动态时间规整算法计算结果能更精确的反应两个时间序列的相似性。图3-4表示了两个长度不同的时间序列的扭曲对应路径。

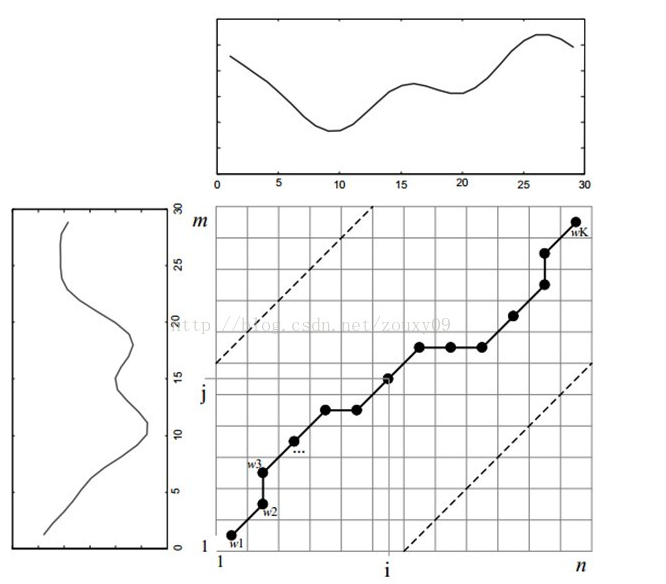


图3-4 DTW计算的两个时间序列的扭曲路径[49]

### 2.5.2 子序列动态时间规整

在一些应用场景下，需要比较的不是两个完整的时间序列的相似度，而是在一个长序列中寻找与短序列最接近的最优子序列。如图3-9所示，在一个长时间序列中识别了一个与查询短序列最接近的子序列片段。这个寻找最优子序列的问题可以通过经典动态时间规整算法的一个优化算法来实现，下面简单介绍这种子序列动态时间规整算法。

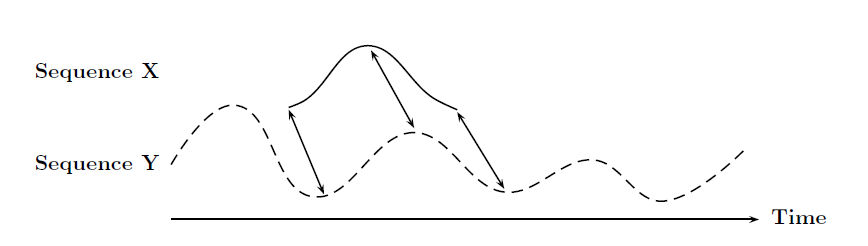


图3-9 子序列识别[49]

假设有两个时间序列和，其中M远大于N。目标是选择一个Y的子序列，使得

|  |  |  |
| --- | --- | --- |
|  |  | （3-6） |

在经典动态时间规整算法中，对于两个序列的对齐不允许漏掉任何点，对于任何漏掉某个序列中点的路径都会增加一个正无穷的惩罚使得这个路径计算的距离值大于其他路径。通过对经典动态时间规整算法做一个修改，可以使得长序列在对齐中可以被略掉开头和结尾的点。通过这个修改，可以算出长序列与短序列最相似的子序列。对于定义3.4具体的修改如下：

1. （相对于定义3.4是）
2. （相对于定义3.4是）

通过修改后的定义，我们可以计算出新的积累代价矩阵。其中N行对应的值即是所有终点为该点的Y的子序列与X序列的DTW距离，因此。根据该处的值也就可以计算出X序列和Y序列中子序列的最相似距离，并且可以得到该子序列的终点。因为在本文的应用中，得到长序列的对齐终点即可（该点为买入或卖出点），所以如何计算得到略过，具体内容可以参考文献[49]，计算过程如图3-10所示。

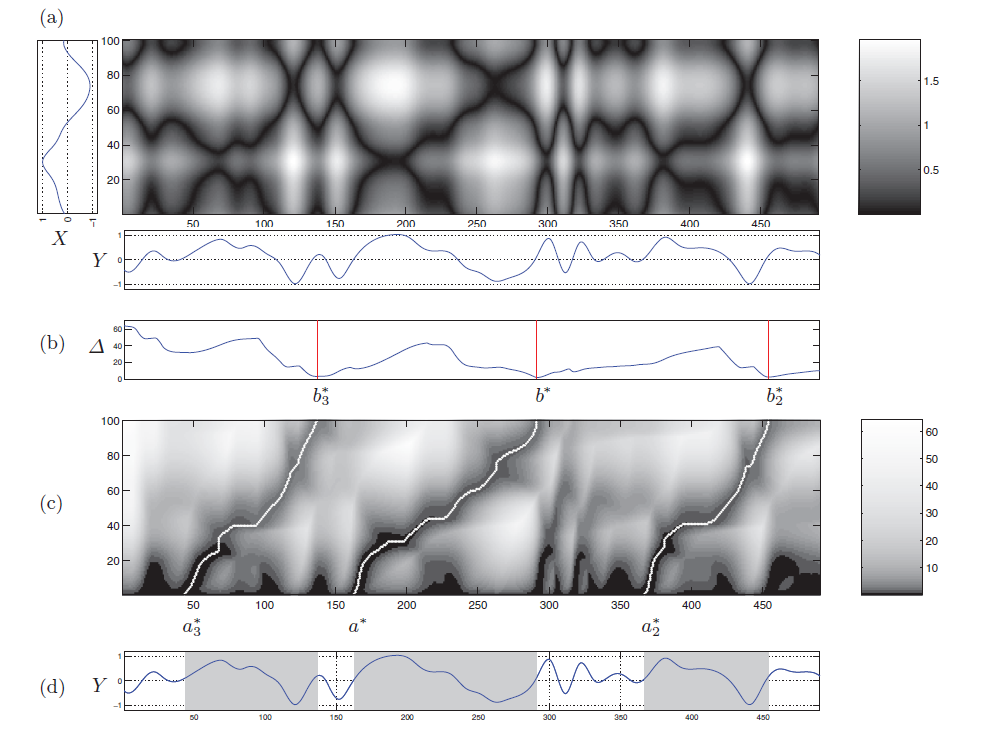


图3-10 子序列动态时间规整计算过程图示[49]

## 2.6 本章小结

本章主要研究了现有被应用与股票价格预测领域的一些算法。主要研究了模糊时间序列算法、小波变换、隐式马尔科夫模型在股票分析中的应用。

模糊时间序列是通过将时间序列中的精确值与预测目标论域的模糊值建立函数联系，通过样本数据的学习得到关系矩阵，进而进行股票走势的预测。小波变换是通过提取时间序列中的时频域信息，并将这种提取算法与神经网络或者支持向量机等等分类算法相结合来进行预测。隐式马尔可夫模型是假设观测值背后可能的马尔可夫链存在某种联系，通过训练数据的训练得到这种联系的一种预测算法。

通过对现有股票走势预测技术的研究可以发现，预测方式有两种方向：价格预测和趋势预测。小波变换模型是价格预测，通过预测价格来判断后续走势。模糊时间序列模型和隐式马尔可夫模型都是趋势预测，预测后续的涨跌趋势。

在小波变换模型的预测中可以看到，预测的均方误差都很小，并且相关系数也比较高。但这跟国内股市有涨跌幅限制的市场设计有关。根据国内A股市场的涨跌停限制，最极端的情况，预测的误差也不会超过20%（涨停10%，跌停10%）。而实际应用场景中，股票的走势往往更加平缓，体现在股票预测统计结果上就是均方误差很小。并且，在实际股票操作中，更关注的是股票价格的涨跌，这决定了后续的买入／卖出操作。例如对于实际走势上涨2%的情况，预测上涨10%和预测下跌1%，虽然预测下跌1%的均方误差更低，但是对于股票操作的决策来说，预测上涨10%是更贴近实际情况的。所以对趋势进行预测的实际应用效果要好于对实际价格进行预测。

在模糊时间序列模型和隐式马尔可夫模型中，对于趋势的预测都是短期的预测。由于中国股市有涨跌停的限制，在涨停阶段很难买入，而在跌停阶段很难卖出。另外就交易成本来说中国股市有印花税，这也是大多数股票市场所没有的。因为中国股市所拥有的这些特点，采用短期趋势预测的方法，例如隐式马尔可夫模型如图 2-13所示，可以看到在上涨的阶段的回调中，判断状态会变为下跌，如果此时采取了卖出操作，当后续股票重新上涨时，因为涨停中无法买入的特点，可能导致丧失后续上涨的大量收益。同样的问题在下跌阶段因为跌停无法卖出的特点，也会导致更大的损失。同样在横盘震荡阶段，因为频繁上涨下跌切换，会导致频繁的买入／卖出交易，造成交易成本的上升。

因为以上的问题，本文希望建立一个能够预测股票中长期走势的预测模型。在第三章中，将围绕这个问题，对时间序列的分析手段进行研究，以期能够在现有的时间序列挖掘算法中寻找能够预测股票中长期走势

第3章 基于模式的股票价格走势预测方案

模式是对于某一个观测对象中反复出现的一个具有一定区分度的事件。如2.4.2节所述，在股票价格走势中存在某些反复出现的价格走势，本文希望通过基于时间序列的shapelets的分类算法提取股票价格走势中的模式。下文将介绍本文的提出的基于模式发现的股票价格走势预测方案。

## 3.1 整体股票价格走势预测方案的设计

本文提出的基于模式的股票价格走势预测方法流程如下，如图4-3所示：

1. 对训练数据集中的股票价格走势进行中长期走势状态的判断，生成已经分割好并完成上涨/下跌等状态打标的价格走势序列
2. 通过基于模式的股票价格走势特征序列提取方案（具体方案在4.3节中）提取步骤1中的价格走势序列中的所有价格走势特征子序列
3. 使用所有走势特征子序列和需要预测的股票近期价格走势进行子序列动态时间规整，所有距离小于阈值的匹配终点都被标记为相应的预测点，预测方向与走势特征子序列代表的走势方向相同

股票价格时间序列分隔打标

基于模式的股票价格走势特征序列提取

股票价格走势预测

训练数据

已分割打标的训练数据

特征序列

待预测股票的近期价格走势

预测结果

图4-3 基于模式发现的股票价格走势预测整体方案

本文提出的基于模式的股票价格走势预测方案能够对股票当前的走势状态（上涨／下跌／横盘）以及未来中期的走势趋势提供预测，之所以选择这样一个预测方式而不是直接预测下一天的股票价格是由国内股市的特点所决定的。国内股票交易市场由于起步晚，散户众多等特点，所以造成了股票价格易受政策影响，交易行为非理性成分大。体现在市场的股票价格走势上就是牛市短暂而熊市漫长，股票价格剧烈上涨和剧烈下跌的时间都比较短，股票价格横盘震荡下跌的时间比较长，如图4-1所示上证指数2004年1月到2018年11月的走势，从图中可以看到两个明显的剧烈上涨阶段（分别是2006年7月到2007年10月和2014年7月到2015年5月）和两个明显的剧烈下跌阶段（分别是2007年10月到2008年10月和2015年5月到2015年9月），其他时间都处于横盘震荡下跌阶段。相对于图4-2所示的道琼斯指数从2004年1月到2018年11月走势，可以看到道琼斯指数走势是处于总体上涨的状态。



图4-1 上证指数2004年1月到2018年11月的走势



图4-2 道琼斯指数从2004年1月到2018年11月走势

这样的指数走势体现在个股的股价波动上就是国内股票市场在急速上涨阶段会出现整个股票市场的股票全部上涨，而在急速下跌阶段出现整个股票市场的股票全部下跌，也就是所谓的“千股涨停”和“千股跌停”的现象。并且由于中国股市有涨跌停的限制，在涨停阶段很难买入，而在跌停阶段很难卖出。另外就交易成本来说中国股市有印花税，这也是大多数股票市场所没有的。因为中国股市所拥有的这些特点，采用直接预测股票下一天的价格会有两个问题：

1. 直接预测股票下一天的价格的模型对于股票价格走势趋势的预测有一定的滞后性。通过第二章的讨论可以看到不管是模糊时间序列模型还是隐式马尔可夫模型都是在股票价格走势发生变化后才会对下一天的股票价格走势趋势的预测发生变化。由于国内涨跌停的限制，可能此时股票价格已经涨／跌停了，会导致股票无法买入／卖出，带来可能的收益损失。
2. 在股市震荡盘整阶段，可以看到直接预测股票下一天的价格的模型对于股市的判断频繁在涨跌间切换，此时虽然不会有长时间涨停、跌停的情况出现，但是会导致频繁的交易。而由于国内股市的交易成本相对较高，会有较大的交易手续费支出。

为解决这一问题，本文提出了上文描述的可以对股市的中长期走势做出预测的模型。通过第二章的讨论可知，基于时间序列的shapelets算法能够提取不同分类的时间序列中的特征子序列。在本文的应用场景下，可以将股票走势序列从中长期走势人为的分为大幅上涨、大幅下跌、震荡盘整这样的分类，然后就可以将基于时间序列的shapelets算法应用到股价序列的趋势分类中。与文献[59]中应用场景所不同的是以下两点：

1. 在图像识别中图片边缘的序列可以通过调整分辨率等方法让时间序列的长度相等，所以在原文中选择了欧式距离作为时间序列的相似度度量。而在股票价格走势的时间序列相似性度量上，因为股市交易的随机性，选择欧式距离显然是不合适的，所以本文选择了动态时间规整(DTW)作为算法中的相似性度量算法。动态时间规整算法在这里作为基于时间序列的shapelets算法中使用的相似性度量算法，提升了原算法的适应性。
2. 在基于时间序列的shapelets算法中，选择了信息熵增益最大的子序列作为特征子序列。而在股票走势预测这样的应用场景下，目标是希望更多的匹配到上涨或者下跌的预测点并且不同的特征子序列可能是由不同的市场或政策原因导致的，所以本文中提出的算法将设定信息熵增益阀值，所有高于阀值的特征子序列都被认为是具有区分度的特征子序列。

基于时间序列的shapelets算法通过信息熵增益的方法可以较好的识别在不同分类的时间序列中具备区分能力的特征子序列。本文通过将识别出的股票价格走势特征子序列与待预测的股票近期价格走势通过动态时间规整进行相似性度量来进行预测。如2.4.2节所述，因为国内股票市场具备一些重复出现的股票价格走势，所以通过本文提出的预测方案是可以在一定程度上预测股票的后续走势的。

## 3.2 股票价格时间序列分割打标方案

基于模式发现的股票价格走势特征序列提取需要提供有标的训练数据集，因此提取特征序列前首相需要对训练数据进行分割打标。本文提出的预测方案是对股票价格走势的中长期趋势进行预测，所以在股票价格时间序列分割打标时需要获取符合预先定义的上涨比率或者下跌比率的股票价格时间序列中最高点最高和最低点最低点的序列。基于此，本文中所采取的股票价格时间序列分割打标方案如下所述：

首先定义上涨比率a、下跌比率b、最短序列长度c，对于任意一点x其价格为f(x)，则区分上涨区间、下跌区间依赖如下定义：

定义4.1：对于任意两个点m、n，如果m-n>c，(f(m)-f(n))/(m-n)>a，且不存在两个点x、y，x<=n，y>=m，f(x)<=f(n)，f(y)>=f(m)，(f(y)-f(x))/(y-x)>a，则m和n之间的序列是上涨序列

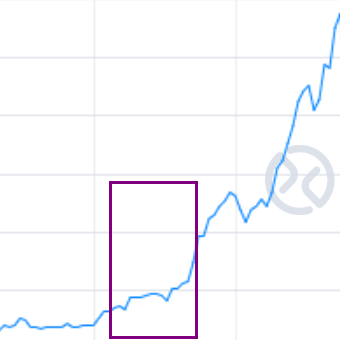
定义4.2：对于任意两个点m、n，如果n-m>c，(f(m)-f(n))/(n-m)>b，且不存在两个点x、y，x>=n，y<=m，f(x)<=f(n)，f(y)>=f(m)，(f(y)-f(x))/(x-y)>b，则m和n之间的序列是下跌序列

定义4.3：股票价格序列的分割和打标方案如下：

* + 1. 对整个序列的所有点对应的值按照大小排序
    2. 取序列中的最大值点p和最小值点q
    3. 判断此两个点之间距离是否小于c，如果小于c，则移除点p与点q间点所有点（包括点p与点q）
    4. 然后依据最大值和最小值的相对顺序决定是进行上涨还是下跌序列判定。
    5. 如果是上涨判断则计算两点间斜率是否大于a，如果是下跌判断则计算两点间斜率是否小于-b
    6. 如果步骤e中符合，则此序列为上涨/下跌序列，并将点p与点q间的所有点移除（包括点p与点q）
    7. 如果步骤e中不符合，则如果是进行上涨判断则在最大和最小两个点分别做一个斜率为a的二元一次函数直线；如果是进行下跌判断则在最大和最小两个点分别做一个斜率为-b的二元一次函数直线。然后做如下操作：
       1. 从最大值点p做的直线与点p和点q间曲线相交距离点p最远点为m，点p与点m间距离点p大于c的所有点中的最小值点n。则点p与点n间曲线为上涨/下跌序列，并将点p与点n间的所有点移除（包括点p与点n）
       2. 从最小值点q做的直线与点p和点q间曲线相交距离点q最远点为x，点q与点x间距离点q大于c的所有点中的最大值点y。则点q与点y间曲线为上涨/下跌序列，并将点q与点y间的所有点移除（包括点q与点y）
       3. 如果点p与点q所做直线都没有与点p和点q间曲线相交，则将点p与点q间的所有点移除（包括点p与点q）
    8. 如果还有点没被移除，则重复步骤b

## 3.3 基于模式的股票价格走势特征序列提取方案

依据3.4节对基于时间序列的shapelets算法的描述和4.1节对应用于股票价格预测场景的差异，本文的基于模式发现的股票价格走势特征序列提取方案如图4-2所示。



计算训练数据集信息熵

依据相似度阈值分类

计算分类后数据集信息熵

该子序列为特征序列

特征序列归集

否

是

信息熵增益是否大于阈值

图4-2 基于模式的股票价格走势特征序列提取流程

具体流程描述如下：

1. 计算已经分割打标的训练数据集的信息熵GAIN
2. 定义子序列窗口长度W、上涨序列距离阈值、下跌序列距离阈值、信息熵增益阈值G、特征子序列最小步长阈值M
3. 通过在上涨/下跌序列中移动子序列窗口，获取上涨/下跌序列中所有可能的子序列（本文中实验设定的子序列窗口长度W为20，原因是一个月是股票走势形成的一个周期[38]）
4. 根据3.4.1节所述计算每个子序列对应的信息熵增益，采用3.3.3节所述的子序列动态时间规整作为相似性度量算法，分类方法如图4-2所示。



子序列动态时间规整

距离大于阈值

距离小于阈值

图4-2 子序列动态时间规整分类

1. 根据设定的信息熵增益阀值G，所有信息熵增益大于阀值的子序列都是特征子序列
2. 因为有部分特征子序列是前后步长相差非常少的子序列，他们所代表的特征是相同的如图4-2所示。所以需要对所有特征子序列进行归集，所有步长差距在设定阀值M内的子序列作为一组，选择信息熵增益最大的子序列作为特征子序列，移除其他子序列

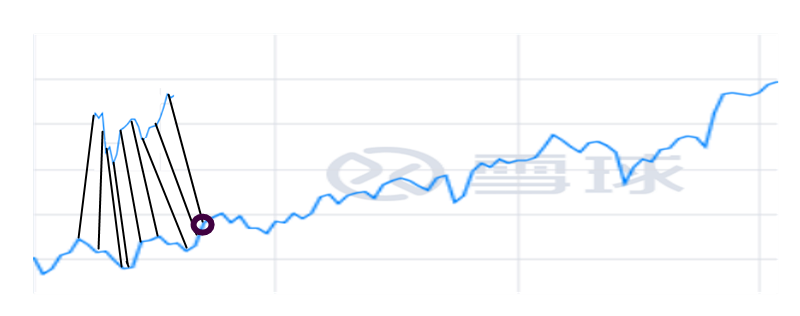
图4-2 步长相差非常少的特征子序列



## 3.4 基于股票价格走势特征序列的价格走势预测方案

股票价格走势的预测是基于子序列动态时间规整算法，预测方案如下：

1. 截取待预测股票近期的股票价格走势（时间序列的长度要大于3.3步骤中的子序列窗口长度W）
2. 计算所有特征子序列与待预测的股票近期（选择窗口为4.3节中的子序列窗口长度W的2倍）价格走势的子序列动态时间规整中的积累代价矩阵
3. 积累代价矩阵中测试股票价格走势所有与上涨/下跌特征子序列距离小于4.3节所设定的阈值的点，都是相应的上涨/下跌预测点。如图4-2所示，其中标注的圆圈点既为预测点，本文提出的方案该点之后中长期会处于特征序列所代表的趋势，图示中所代表的是上涨。

图4-2 股票价格走势趋势的预测

## 3.5 本章小结

本章根据上文的研究结果，阐述了本文对于股票价格走势预测的见解和本文提出的基于模式发现的股票价格走势预测方法。通过对基于时间序列的shapelets算法和动态时间规整算法的整合形成了本文对中长期股票走势预测的一种方案。下面将通过A股市场的历史数据对本文提出的预测方案进行验证。

第4章 实验验证

在上文讨论的基础上，本章将本文提出的基于模式发现的股票价格走势预测方法应用于股票趋势预测。通过将该模型对沪深300股票的历史数据进行训练和评测，来评价该预测方案在股票预测领域的应用前景。

## 4.1 实验环境

本文实验环境是一台台式机，具体配置如下：

1. CPU：Intel Core i7-7700 3.60GHz
2. 内存：16GB
3. 存储：128GB SSD + 1T 硬盘
4. 操作系统：Win10 64位
5. 数据库：mongodb4.0
6. IDE：Eclipse

## 4.2 实验数据的选取

本文选择了上证50指数包含的50支股票的历史股价数据作为训练数据集和测试数据集。其中2014年09月01日至2014年12月31日和2015年06月01日至2015年07月31日的股票价格数据为训练数据集，2015年01月01日至2015年04月30日和2015年08月01日至2015年09月30日的股票价格数据为测试数据集，该段时间的上证指数走势如图5-1所示。这段时间是2015年股灾前后股票大幅上涨和急速下跌的阶段，在这种急速上涨和下跌阶段，如果能够较准确的预测股市走势，将能够获取相对丰厚的收益。选择这段具有标志性的数据正是希望检测本文提出的算法能否在国内股市这种急速上涨和下跌阶段能够较准确的预测股票走势。

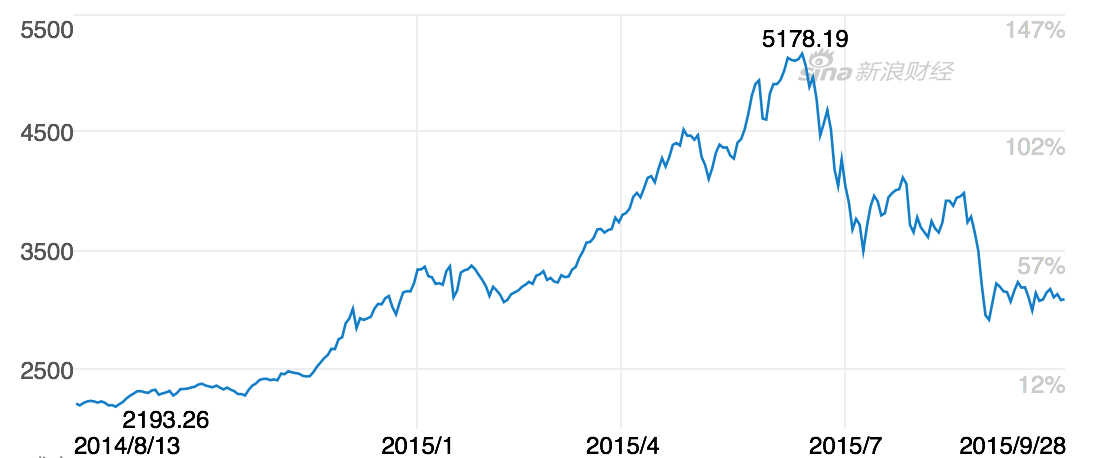


图5-1 训练数据和测试数据所属时间对应的上证指数

## 4.3 代码实现

本文实现的代码被上传到github上，地址为：<https://github.com/quantaolin/MyPaper>。代码使用python实现，下载股票数据使用了tushare开源包，数据库使用了mongodb。整体代码介绍如下：

1. downloadstockdata包是下载股票数据到本地的代码，其中有下载全部股票数据和各种成分股指数数据，本文中使用的是上证50指数。
2. datapreprocess包是数据预处理代码，其中计算了股票每日的涨跌幅（因为股票价格差别较大，所以本文使用涨跌幅作为股价值来进行归一化处理）、获取上涨/下跌序列
3. dataprocess包是特征提取的代码，进行了4.2节中的处理步骤（3）（4）（5）。
4. test包是测试数据集测试的代码，进行了4.2节中的处理步骤（6），并对测试结果的准确性进行了计算，另外对一些提取的数据进行绘图方便直观观察。

## 4.4 实验结果分析

### 4.4.1 预测性能

在不同信息熵增益阀值下，测试结果如表5-1所示。从表中可以看到上涨特征子序列的预测准确度明显高于下跌特征子序列的预测精度，这点与文献[38]中的结论相同。随着信息熵增益阀值的提高，预测的精度显著提高，但是特征子序列的数量在急剧减少，并且检测点也在减少，这就意味着漏报率也在上升，在急速上涨阶段会错失一些盈利的机会，但同时也降低了在大跌中判断错误造成损失的可能。从这里可以看出，提高信息熵增益阀值相当于投资决策趋于谨慎，这与其它投资策略中投资愈加谨慎获取超额收益的机会降低同时造成巨额损失的可能性也在下降的情况是相同的。从表中可以看出选择信息熵增益阀值为0.5具有比较好的预测精度和预测点数量。

表5-1 上证50成分股走势预测结果

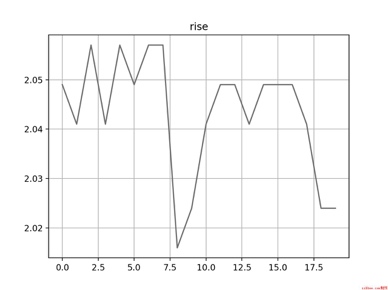
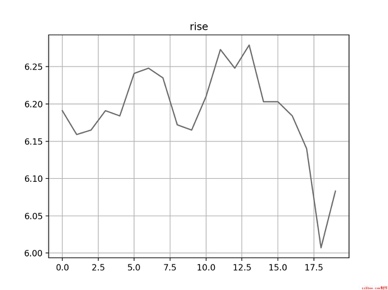
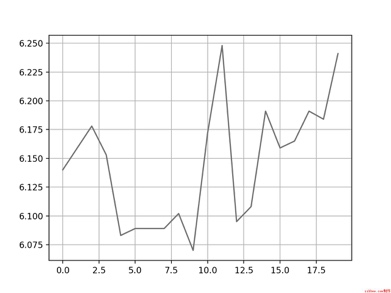
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 信息熵增益阀值 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
| 上涨特征序列数 | 13 | 11 | 10 | 7 | 1 |
| 上涨预测正确数 | 582 | 484 | 404 | 251 | 14 |
| 上涨预测错误数 | 72 | 56 | 51 | 31 | 0 |
| 上涨预测准确率 | 89% | 89.6% | 88.8% | 89% | 100% |
| 下跌特征序列数 | 21 | 20 | 17 | 15 | 8 |
| 下跌预测正确数 | 1645 | 1723 | 1375 | 1089 | 554 |
| 下跌预测错误数 | 900 | 821 | 570 | 443 | 199 |
| 下跌预测准确率 | 64.6% | 67.7% | 70.7% | 71.1% | 73.6% |
| 总预测准确率 | 69.6% | 71.6% | 74.1% | 73.9% | 74.1% |

比较表2-1和文献[38]的预测结果，本文的整体预测精度要好于这个模糊序列优化模型的预测精度。尤其是上涨的预测精度显著的高于其他模型，如表5.2所示。

|  |  |  |
| --- | --- | --- |
|  | 上证 | NYSE |
| Chen的模型 | 54% |  |
| 杨一文的模型 | 72% |  |
| 动态时间规整[38] |  | 58.08% |
| 本文的模型 | 74.1% |  |

### 4.4.2 模式挖掘结果

在信息增益阀值为0.5时，部分上涨特征序列如图5-2所示，从图中我们看到类似”W底”这样的走势模式，同时也有没有在传统股票走势预测中被总结出的走势图形，甚至有微跌的走势图形，说明在本文提出的算法中可以挖掘相对人工更多的走势特征子序列。同样的情况通过下跌特征序列图5-3也可以观察到。



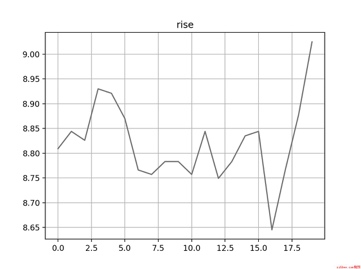
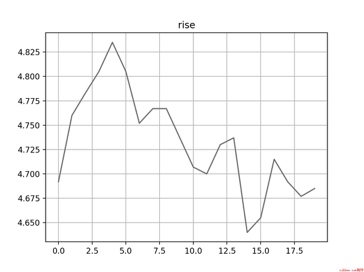
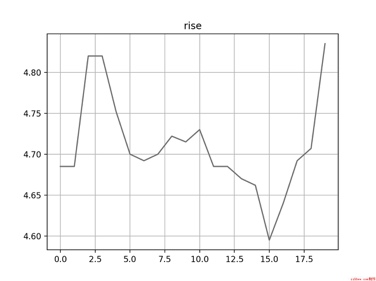


图5-2 上涨特征序列

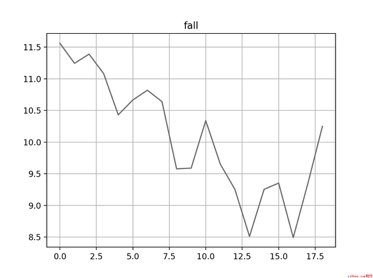
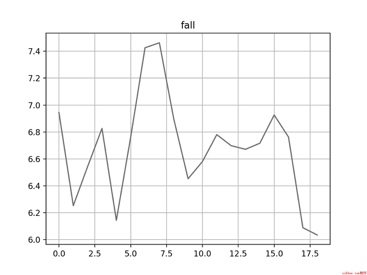
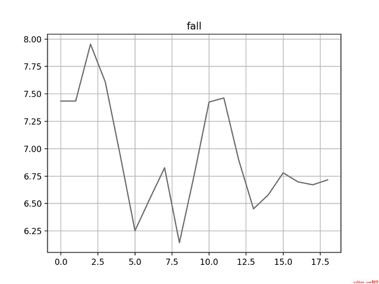
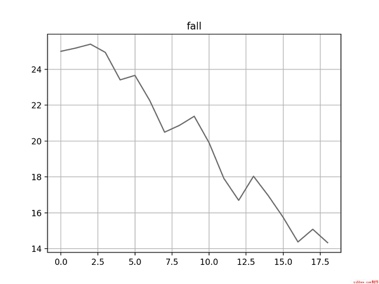
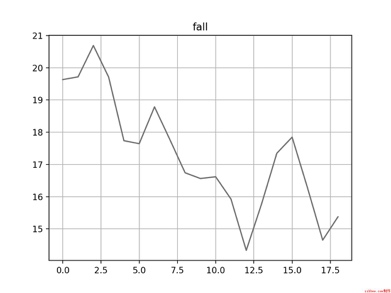
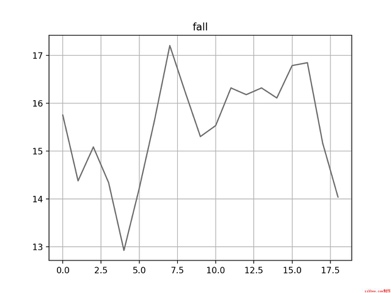
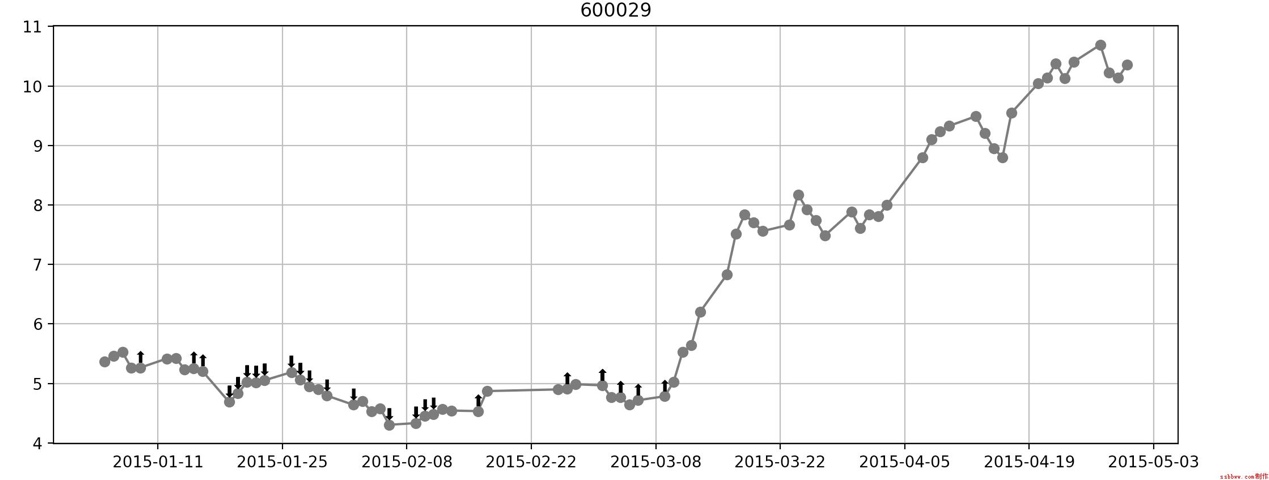
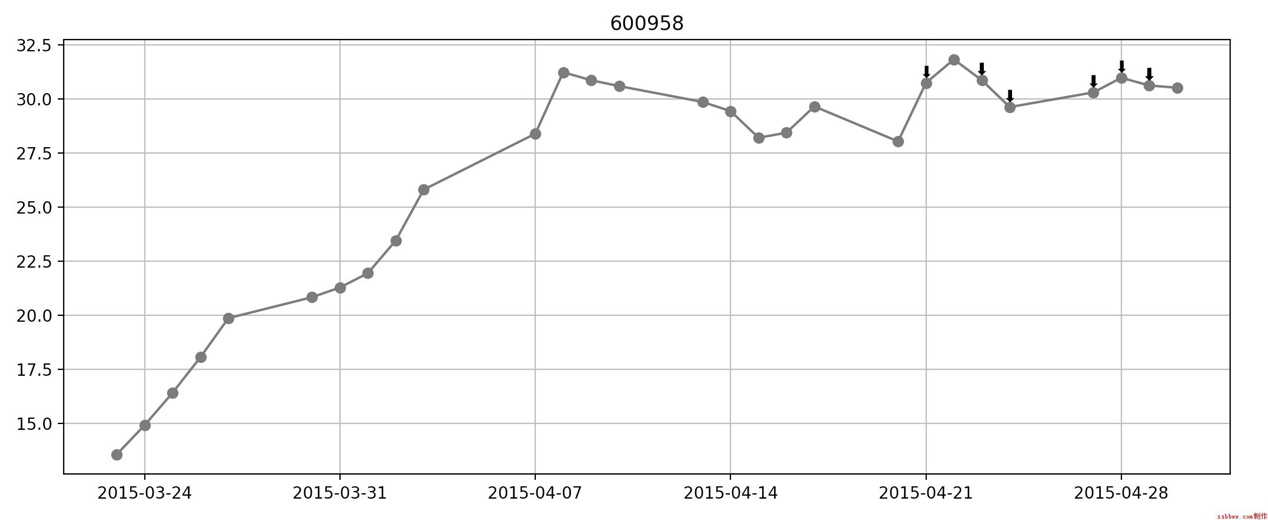
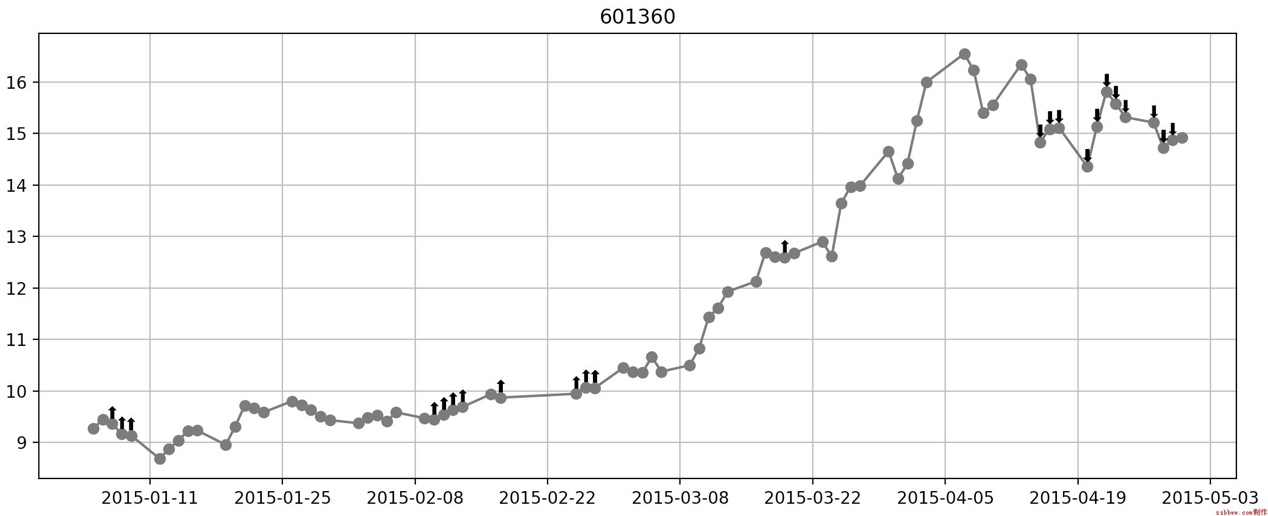
 

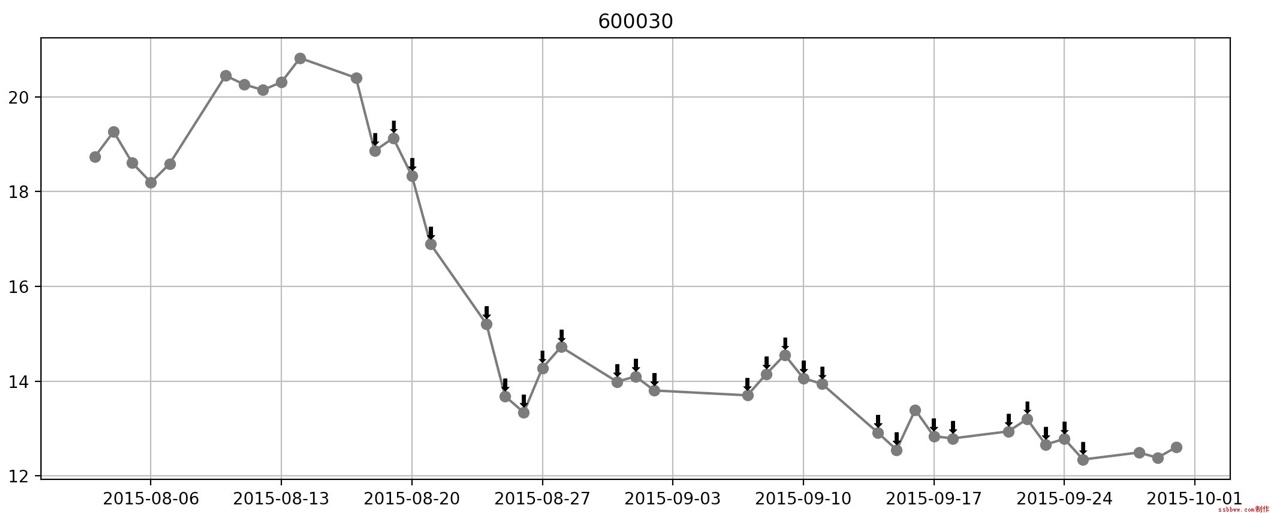
图5-3 下跌特征序列

对股票预测点如图5-4所示，其中向上箭头代表上涨预测点，向下箭头代表下跌预测点。可以看到部分在上涨时间片段中的下跌预测点在价格序列的尾部，而根据本文选择的测试数据集中上涨序列所处的历史价格位置可知，测试数据集中的上涨序列的尾部已经非常接近于2015年“股灾”的顶点，所以该处预测点之后的股价有可能已经开始下跌。但是因为受限于本文选择的测试数据集，所以该种预测被分为了错误的预测。而对于下跌股票时间序列的测试数据集，因为股价走势尾部之后股价进入了横盘震荡的走势，所以在尾部预测为上涨的点显著较少。这也是本文的检测结果中，下跌预测精度较低的一个原因。









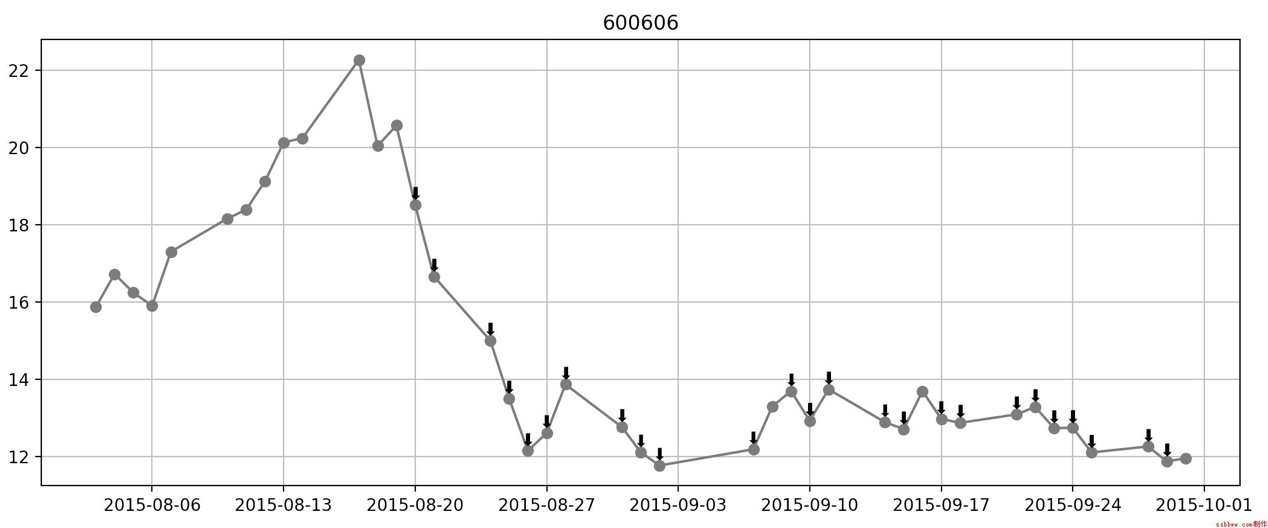


图5-4 股票预测结果图示

## 4.5 本章小结

本章通过实验验证了本文提出的基于模式发现的股票价格走势预测方法。实验过程使用python实现，并将该算法应用于上证50指数包含的50支股票的历史股价数据检测了该算法的准确性，并分析了模式挖掘的结果。

第5章 总结与展望

## 5.1 全文总结

本文将基于时间序列的shapelets算法和动态时间规整算法进行整合调整后应用到股票走势预测中，并使用上证50的历史数据对该算法的预测准确度进行了测试，本文的主要研究内容如下：

1. 对量化交易的现状进行了研究，并深入研究了现有的股票走势预测算法，比较了不同算法的特点。重点对模糊时间序列算法、小波变换算法、隐式马尔可夫模型在股票趋势预测的应用进行了研究。通过总结不同的预测算法，发现现有的预测算法主要围绕股票走势的时间序列进行分析这样的特点。
2. 对时间序列相似性度量算法进行了研究。相似性度量算法是测量两个时间序列相似度的算法，是时间序列特征提取算法的基础，本文重点研究了动态时间规整算法及其优化算法。尤其是子序列动态时间规整算法，因为可预知的股票走势特征序列会远小于股票走势序列，所以该优化算法会被应用于本文。
3. 对基于时间序列的shapelets算法进行了研究。该算法最初是用于图片识别的一种图片边缘序列识别分类算法，通过计算子序列对于训练数据集分类的信息熵增益来获取特征子序列。并通过对股票走势的分析论述了该算法应用于股票走势预测的可行性。
4. 通过对国内股票市场特点的分析和现有股票走势预测算法的研究，分析了现有股票走势预测算法应用于国内股票市场股票走势预测的一些问题，并通过将基于时间序列的shapelets算法中的相似性度量算法替换为动态时间规整算法和调整了判定特征子序列的标准提出了本文的股票趋势预测算法。最后将该算法应用于上证50成分股的历史走势检测了该算法的预测准确度。

## 5.2 研究展望

因为本文研究环境的局限性，在算法选择和应用范围上都受到了一定的约束，我认为在以下几个方面有进一步研究改进的价值：

1. 子序列动态时间规整算法的时间复杂度是O(mn)。基于时间序列的shaplet算法的时间复杂度依据子序列选择方式的不同有所不同，但是时间复杂度都在O(n^2)以上。所以本文提出的股票趋势预测算法时间复杂度高，计算耗时长，本文因为是算法验证性质，所以采用python实现单机处理。如果要应用于实际股票预测场景，对预测的时效性有要求，需要使用并行处理框架实现算法，通过分机并行处理的方式提高算法计算速度。
2. 因为受到计算能力的限制，本文只选择了长度为20的子序列作为可能的备选子序列，实际可能的特征子序列可能在其他长度的子序列存在，所以研究不同长度子序列的预测精度是一个可能的研究方向
3. 受到研究时间的限制和计算能力的限制，本文只选取了相对较短的股市急速上涨和急速下跌的阶段进行了分析，算法也围绕区分急速上涨和急速下跌的股票走势为目标。实际股票走势还分为震荡上涨、震荡下跌、横盘震荡这些走势。增加3个股价分类意味着基于时间序列的shapelets算法的决策树需要增加层数，并且分类情况更加复杂，需要对算法进行进一步的优化。
4. 本文选择的特征监测点方法是有某一个特征子序列匹配到该点即认为该点是买入/卖出点。但是我们通过观察可以发现有些特征子序列之间是相似的，有些是不相似的，这里选择k-nn算法作为分类算法是不合适的，因为不同特征子序列之间可能是完全不同的原因导致的，不具备相关性。但是采用本文的算法完全忽略了特征子序列间可能存在的相关性，因此在分类算法方向有进行进一步的研究的价值。

参考文献

1. 李子睿. 量化投资交易策略研究[硕士学位论文]. 天津: 天津大学, 2013.
2. FT中文网. 量化投资——从西蒙斯谈起. [2015-12-31]. <http://www.ftchinese.com/story/001065528?page=1>.
3. [郑文](http://eproxy2.lib.tsinghua.edu.cn/kcms/detail/search.aspx?dbcode=CMFD&sfield=au&skey=%e9%83%91%e6%96%87&code=28421846;). 数据挖掘在股票分析中的应用[硕士学位论文]. 上海: 复旦大学, 2012.
4. [陈金佑](http://eproxy2.lib.tsinghua.edu.cn/kcms/detail/search.aspx?dbcode=CMFD&sfield=au&skey=%e9%99%88%e9%87%91%e4%bd%91&code=31975776;). 数据挖掘在股票分析中的应用研究[硕士学位论文]. 广州: 华南理工大学, 2014.
5. [KROLAK PD](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5EsGkgXpHWPFCfEJF5K&field=AU&value=KROLAK,%20PD&ut=4174273&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [LINDAHL CE](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5EsGkgXpHWPFCfEJF5K&field=AU&value=LINDAHL,%20CE&ut=24707664&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [HERNDON JM](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5EsGkgXpHWPFCfEJF5K&field=AU&value=HERNDON,%20JM&ut=934554&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). CONSTRUCTION OF A ROBOT INVESTOR. SIMULATION, 1973, 21(4):97-109.
6. [Malagrino LS](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6Be2w1UqX83MLTyM8K9&field=AU&value=Malagrino,%20LS&ut=24317108&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage&cacheurlFromRightClick=no), [Roman NT](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6Be2w1UqX83MLTyM8K9&field=AU&value=Roman,%20NT&ut=4957773&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Monteiro AM](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6Be2w1UqX83MLTyM8K9&field=AU&value=Monteiro,%20AM&ut=1487899&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Forecasting stock market index daily direction: A Bayesian Network approach. EXPERT SYSTEMS WITH APPLICATIONS, 2018, 105:11-22.
7. [Tsionas MG](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8EMitW1eSJMcdnVnEis&field=AU&value=Tsionas,%20MG&ut=1120651&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Michaelides PG](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8EMitW1eSJMcdnVnEis&field=AU&value=Michaelides,%20PG&ut=667001&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage). Neglected chaos in international stock markets: Bayesian analysis of the joint return-volatility dynamical system. PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS, 2017, 482:95-107.
8. [Billio M](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8EMitW1eSJMcdnVnEis&field=AU&value=Billio,%20M&ut=1337016&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Casarin R](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8EMitW1eSJMcdnVnEis&field=AU&value=Casarin,%20R&ut=1664326&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Osuntuyi A](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8EMitW1eSJMcdnVnEis&field=AU&value=Osuntuyi,%20A&ut=12258807&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Markov switching GARCH models for Bayesian hedging on energy futures markets. ENERGY ECONOMICS, 2018, 70:545-562.
9. Zadeh L A. Fuzzy set[J]. Information and Control, 1965, 8:338-353.
10. Song Q, Chissom B S. Forecasting enroollments with fuzzy time series Part I[J]. Fuzzy Set System, 1993, 54(1):1-9.
11. Song Q. Chissom B S. Fuzzy Time Series and Its Models [J]. Fuzzy Sets and Systems, 1993.
12. [Cheng CH](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Cheng,%20CH&ut=224058&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Yang JH](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Yang,%20JH&ut=6059430&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage). Fuzzy time-series model based on rough set rule induction for forecasting stock price. NEUROCOMPUTING, 2018, 302:33-45.
13. [Fang Y](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Fang,%20Y&ut=1162998&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Bo L](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Bo,%20L&ut=432707&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Zhao DP](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Zhao,%20DP&ut=8115310&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage), [Wang SY](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Wang,%20SY&ut=27200&pos=4&excludeEventConfig=ExcludeIfFromFullRecPage). Fuzzy Views on Black-Litterman Portfolio Selection Model. JOURNAL OF SYSTEMS SCIENCE & COMPLEXITY, 2018, 31(4):975-987.
14. [Efendi R](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Efendi,%20R&ut=3219023&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Arbaiy N](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Arbaiy,%20N&ut=2030014&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Deris MM](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Deris,%20MM&ut=246512&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. INFORMATION SCIENCES, 2018, 441:113-132.
15. [Soto J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Soto,%20J&ut=1430255&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Melin P](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Melin,%20P&ut=8462&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Castillo O](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Castillo,%20O&ut=3992&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). A New Approach for Time Series Prediction Using Ensembles of IT2FNN Models with Optimization of Fuzzy Integrators. INTERNATIONAL JOURNAL OF FUZZY SYSTEMS, 2018, 20(3):701-728.
16. [Wang WN](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FbZeFl8VMTKAhQCpCK&field=AU&value=Wang,%20WN&ut=292136&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage). A big data framework for stock price forecasting using fuzzy time series. MULTIMEDIA TOOLS AND APPLICATIONS, 2018, 77(8):10123-10134.
17. 杨一文, 蔺玉佩. 模糊时间序列建模及股票市场多步预测. 计算机工程与应用, 2014, 05:252-256.
18. 李小琳, 孙玥, 刘洋. 基于SVM修正的模糊时间序列模型在沪指预测中的应用. 中国科学技术大学学报, 2016, 03:238-246.
19. [Naranjo R](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8Bm8N8A56wZTr4fHGzw&field=AU&value=Naranjo,%20R&ut=6371455&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Arroyo J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8Bm8N8A56wZTr4fHGzw&field=AU&value=Arroyo,%20J&ut=2199171&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Santos M](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=8Bm8N8A56wZTr4fHGzw&field=AU&value=Santos,%20M&ut=567934&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Fuzzy modeling of stock trading with fuzzy can-dlesticks. EXPERT SYSTEMS WITH APPLICATIONS, 2017, 93:15-27.
20. [Luo SH](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Luo,%20SH&ut=683464&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Huo JY](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Huo,%20JY&ut=26927690&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Dai ZA](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Dai,%20ZA&ut=26996681&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Frequency-Division Combination Forecasting of Stock Market Based on Wavelet Multiresolution Analysis[J]. DISCRETE DYNAMICS IN NATURE AND SOCIETY, 2018.
21. [Huang LL](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Huang,%20LL&ut=24608&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Wang J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Wang,%20J&ut=322076&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage). Forecasting energy fluctuation model by wavelet decomposition and stochastic recurrent waveletneural network. NEUROCOMPUTING, 2018, 309:70-82.
22. [Arevalo A](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Arevalo,+A.), [Nino J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Nino,+J.), [Leon D](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Leon,+D.), [Hernandez G](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Hernandez,+G.), [Sandoval J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Sandoval,+J.). Deep Learning and Wavelets for High-Frequency Price Forecasting. Computational Science, 2018:385-399.
23. 王刚,许晓兵. 基于小波分析与神经网络时间序列的股票预测方法. 金融经济, 2013, 12:161-163.
24. Sri Endah Moelya Artha, Hasbi Yasin, Budi Warsito, Rukun Santoso1 and Suparti. Application of Wavelet Neuro-Fuzzy System (WNFS) method for stock forecasting. Journal of Physics: Conference Series, 2018.
25. 李坤, 谭梦羽. 基于小波支持向量机回归的股票预测.统计与决策, 2014, 06:32-36.
26. [Raimundo MS](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Raimundo,%20MS&ut=26924804&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Okamoto J](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=6EMpmhVHNdc8nM4aCsB&field=AU&value=Okamoto,%20J&ut=37970&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage). SVR-Wavelet Adaptive Model for Forecasting Financial Time Series .CONFERENCE PROCEEDINGS OF 2018 INTERNATIONAL CONFERENCE ON INFORMATION AND COMPUTER TECHNOLOGIES (ICICT), 2018:111-114.
27. 席雯雯, 姬强, 范英. 国际油气价格动态关系对中国天然气定价机制改革的启示. 中国科学院院刊, 2016, 31(7):812-819.
28. Toni Giorgino. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. Journal of Statistical Software, 2009, 31(7).
29. Michail Vlachos, Marios Hadjieleftheriou, Dimitrios Gunopulos, Eamonn Keogh. Indexing Multi-Dimensional Time-Series with Support for Multiple Distance Measures. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2003, 15(1):216-225.
30. Juan P, Caraça-Valente, Ignacio López-Chavarrías. Discovering similar patterns in time series. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2000:497-505.
31. ak-chung Fu. A review on time series data mining. Engineering Applications of Artificial Intelligence, 1976, 24(1):164-181.
32. James D. HAMILTON. ANALYSIS OF TIME SERIES SUBJECT TO CHANGES IN REGIME. Journal of Econometrics , 1990, 45:39-70.
33. 侯雅文. 基于隐马尔可夫模型的股票价格指数预测[硕士学位论文]. 广州: 暨南大学, 2007.
34. 朱嘉瑜, 叶海燕, 高鹰. 基于隐马尔可夫模型的股票价格预测组合模型. 计算机工程与设计, 2009, 30(21):4945-4948.
35. 李嵩松. 基于隐马尔可夫模型和计算智能的股票价格时间序列预测[博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2011.
36. [陈子冰](http://xueshu.baidu.com/s?wd=authoruri:(7afbfbec29c8ee12)%20author:(%E9%99%88%E5%AD%90%E5%86%B0)%20%E8%8B%B1%E5%9B%BD%E8%AF%BA%E4%B8%B9%E6%AF%94%E4%BA%9A%E5%A4%A7%E5%AD%A6%E9%A9%AC%E6%9D%A5%E8%A5%BF%E4%BA%9A%E5%8F%B2%E4%B8%B9%E7%A6%8F%E5%AD%A6%E9%99%A2%E5%90%89%E9%9A%86%E5%9D%A1&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person&sort=sc_cited). 评索罗斯的反身性理论. 前沿, 2006(8):28-31.
37. [邓佳](http://xueshu.baidu.com/s?wd=authoruri:(ff1e5dc2be7a61f0)%20author:(%E9%82%93%E4%BD%B3)%20%E5%9B%9B%E5%B7%9D%E5%A4%A7%E5%AD%A6%E7%BB%8F%E6%B5%8E%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person&sort=sc_cited). 股票价格决定及其波动分析——基于索罗斯的反身性理论. 时代金融旬刊, 2011, 8:144-144.
38. [Tsinaslanidis PE](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Tsinaslanidis,%20PE&ut=7434314&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage). Subsequence dynamic time warping for charting: Bullish and bear-ish class predictions for NYSEstocks. EXPERT SYSTEMS WITH APPLICATIONS, 2018, 94:193-204.
39. [Jeon S](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Jeon,%20S&ut=1805630&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Hong B](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Hong,%20B&ut=1245300&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Chang V](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Chang,%20V&ut=135809&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Pattern graph tracking-based stock price prediction using big data. FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE, 2018, 80:171-187.
40. [Xinxin Yao](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Xinxin+Yao), [Hua-Liang Wei](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Hua-Liang+Wei). Short-term stock price forecasting based on similar historical patterns extraction. 2017 23rd International Conference on Automation and Computing (ICAC), 2017:6.
41. [Bagheri A](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Bagheri,%20A&ut=557304&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Peyhani HM](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Peyhani,%20HM&ut=14641074&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Akbari M](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=5FE6yxFTyvX8zSaG8xZ&field=AU&value=Akbari,%20M&ut=402566&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage). Financial forecasting using ANFIS networks with Quantum-behaved Particle Swarm Optimization. EXPERT SYSTEMS WITH APPLICATIONS, 2014, 41(14):6235-6250.
42. Huarng K , Yu H K . Ratio-based lengths of intervals to improve fuzzy time series forecasting[J]. IEEE Trans Syst Man Cybern B Cybern, 2006, 36(2):328-340.
43. 鲜思东. 基于SVM-FCE 的网络学习评价系统[J]. 数学的实践与认, 2007, 37( 20) :53-60.
44. Li S T, Cheng Y C, Lin S Y. A FCM-based deterministic forecasting model for fuzzy time series[J]. Computers ＆ Mathematics with Applications, 2008, 56(12):3052-3063.
45. Chen S M. Forecasting enrollments based on fuzzy time series[J]. Fuzzy Sets and Systems, 1996, 81:311-319.
46. R Agrawal, C Faloutsos, A Swami. Efficient similarity serarch in sequence databases, In Int. Conference on Foundations of Data Organization(FODO), 1993:69-94.
47. C Faloutsos, M Ranganathan, Y Manolopoulos. Fast Subsequence matching in time-series databases. ACM-SIGMOD Conf, 1994:419-429.
48. 冯琦森. 基于出租车轨迹的居民出行热点路径和区域挖掘[硕士学位论文]. 重庆：重庆大学, 2016.
49. Muller M. Dynamic Time Warping (DTW). In Information Retrieval for Music and Motion, 2007:69-84.
50. C S Myers, L R Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal, 1981, 60(7):1389-1409.
51. Law[rence Rabiner](https://book.douban.com/search/Lawrence%20Rabiner), [Biing-Hwang Juang](https://book.douban.com/search/Biing-Hwang%20Juang). Fundamentals of speech recognition. Prentice Hall, 1993.
52. Tian Y, Wang ZL, Lu C. Self-adaptive bearing fault diagnosis based on permutation entropy and manifold-based dynamictime warping. MECHANICAL SYSTEMS AND SIGNAL PROCESSING, 2019, 114:658-673.
53. [Liu S](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Liu,%20S), [Liu CL](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Liu,%20CL). Scale-varying dynamic time warping based on hesitant fuzzy sets for multivariate time series classification. MEASUREMENT, 2018, 130:290-297.
54. [Berman EE](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Berman,%20EE), [Bolton DK](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Bolton,%20DK), Coops NC, [Mityok ZK](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Mityok,%20ZK), [Stenhouse GB](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Stenhouse,%20GB), [Moore RD](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&excludeEventConfig=ExcludeIfFromFullRecPage&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Moore,%20RD). Daily estimates of Landsat fractional snow cover driven by MODIS and dynamic time-warping. REMOTE SENSING OF ENVIRONMENT, 2018, 216:635-646.
55. [Sharabiani A](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Sharabiani,%20A&ut=12163919&pos=1&excludeEventConfig=ExcludeIfFromFullRecPage), [Darabi H](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Darabi,%20H&ut=1969813&pos=2&excludeEventConfig=ExcludeIfFromFullRecPage), [Harford S](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Harford,%20S&ut=7122865&pos=3&excludeEventConfig=ExcludeIfFromFullRecPage), [Douzali E](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Douzali,%20E&ut=16147666&pos=4&excludeEventConfig=ExcludeIfFromFullRecPage), Karim F, [Johnson H](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Johnson,%20H&ut=25944630&pos=6&excludeEventConfig=ExcludeIfFromFullRecPage), [Chen S](http://eproxy2.lib.tsinghua.edu.cn/OneClickSearch.do?product=UA&search_mode=OneClickSearch&SID=7BQkxYsqo9bSVciitCM&field=AU&value=Chen,%20S&ut=16571177&pos=7&excludeEventConfig=ExcludeIfFromFullRecPage). Asymptotic Dynamic Time Warping calculation with utilizing value repetition. KNOWLEDGE AND INFORMATION SYSTEMS, 2018, 57:359-388.
56. 蒋志浩, 于群, 董骊. 基于改进动态时间规整算法的断路器故障诊断[J]. 工矿自动化, 2016, 42(8):52-55.
57. [S Salvador,](http://xueshu.baidu.com/s?wd=author:(Salvador,%20Stan)%20Correspd.%203639%20Robin%20Ln.,%20Charlotte,%20NC%2028269,%20USA.%20Tel.:%20+1%20704%20599%202993;%20Fax:%20+1%20980%20235%202396;%20E-mail:%20ssalvador@gdatp.com)%20General%20Dynamics,%20Armament%20and%20Technical%20Products,%204205%20Westinghouse&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person) [P Chan](http://xueshu.baidu.com/s?wd=author:(Chan,%20Philip)%20Department%20of%20Computer%20Sciences,%20Florida%20Institute%20of%20Technology,%20150%20W%20University%20Blvd.,%20Melbourne,%20FL%2032901,%20USA.%20E-mail:%20pkc@cs.fit.edu&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person). FastDTW:Toward Accurate Dynamic Time Warping in Linear Time and Space. [Intelligent Data Analysis](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri:(541571e6a45be846)%20%E3%80%8AIntelligent%20Data%20Analysis%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited), 2007, 11(5):561-580.
58. [M Vlachos](http://xueshu.baidu.com/s?wd=author:(Michail%20Vlachos)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person), [M Hadjieleftheriou](http://xueshu.baidu.com/s?wd=author:(Marios%20Hadjieleftheriou)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person), [D Gunopulos](http://xueshu.baidu.com/s?wd=author:(Dimitrios%20Gunopulos)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person), [E Keogh](http://xueshu.baidu.com/s?wd=author:(Eamonn%20Keogh)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person). [Indexing multi-dimensional time-series with support for multiple distance measures](http://xueshu.baidu.com/s?wd=paperuri:(61aed034e2b1bbc47b239a65468fc9f3)&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http://dl.acm.org/citation.cfm?id=956777&ie=utf-8&sc_us=17984197990573670505). Proc Acm Sigkdd International Conference on Knowl, 2003 , 15(1):216-225.
59. Lexiang Ye, Eamonn Keogh. Time Series Shapelets:A New Primitive for Data Mining. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2009:947-956.
60. Lexiang Ye,Eamonn Keogh. Time Series Shapelets:a novel technique that allows accurate,interpretable and fast classification. Data Mining & Knowledge Discovery, 2011, 22(1-2):149-182.
61. A Mueen, E Keogh, N Young. Logical-shapelets:an expressive primitive for time series classification. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2011:1154-1162.
62. T Rakthanmanon, E Keogh. Fast Shapelets: A Scalable Algorithm for Disvoering Time Series Shapelets. Society for Industrial and Applied Mathematics, 2013:668-676.
63. 杜建卫, 王超峰. 小波分析方法在金融股票数据预测中的应用. 数学的实践与认识, 2008, 38(7):68-75.
64. [段文秀](http://xueshu.baidu.com/s?wd=authoruri:(27799bb05cd15e02)%20author:(%E6%AE%B5%E6%96%87%E7%A7%80)%20%E5%B1%B1%E4%B8%9C%E7%A7%91%E6%8A%80%E5%A4%A7%E5%AD%A6&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person&sort=sc_cited). 基于小波网络的数据挖掘技术及在股市预测中的应用[硕士学位论文]. 山东:山东科技大学, 2006.
65. 常松, 何建敏. 基于小波包和神经网络的股票价格预测模型. 中国管理科学, 2001, 9(5):9-15.
66. 大数据部落. 隐马尔科夫模型(HMM)股指预测实战. [2017-08-29]. <https://blog.csdn.net/qq_19600291/article/details/77659501>
67. Robi Polikar. The Wavelet Tutorial. 1996. https://engineering.rowan.edu/~polikar/WAVELETS/WTtutorial.html

致 谢

衷心感谢导师邓仰东教授在论文开题和写作阶段给予的精心指导。邓老师的言传身教将使我受益终生。

作为在职研究生，在论文写作阶段需要平衡工作、学习、家庭的方方面面，其中面临了诸多的困难和迷惑。邓老师在此期间对于论文的研究方向和研究中遇到的困难给予了耐心的指导，对于我完成论文的写作提供了巨大的帮助。

另外我还要感谢我的太太赵忠伟，在论文的写作阶段，我们的儿子出生了。在她本职工作已经很忙的情况下，我太太付出了大量的时间在孩子的照顾上，为我能够安心完成论文的写作提供了良好的环境和保障。在此衷心感谢她的巨大付出。

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

1987年07月24日出生于天津市。

2006年9月考入天津大学电子工程学院电子工程专业，2010年7月本科毕业并获得工学学士学位。