基于多种隐马尔可夫模型的量化择时研究

罗子恒

指导教师　齐官红　吴清强

厦门大学

厦門大學

本 科 毕 业 论 文 （设 计）

（主修专业）

基于多种隐马尔可夫模型的量化择时研究

Quantitative Timing Study Based On Multiple Hidden Markov Models

姓　　名：罗子恒

学　　号：24320142202473

学　　院：软件学院

专　　业：软件工程

年　　级：2014 级

校内指导教师：　齐官红　助教

　　吴清强　教授

校外指导教师：　(姓名)　(职务)

二〇一八年五月二日

厦门大学本科学位论文诚信承诺书

本人呈交的学位论文是在导师指导下独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合相关法律规范及《厦门大学本科毕业论文（设计）规范》。

该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明）。

本人承诺辅修专业毕业论文（设计）（如有）的内容与主修专业不存在相同与相近情况。

学生声明（签名）：

年　　月　　日

致 谢

关于完成这个论文的整个过程，有许多的人在这个过程中祝我一臂之力，帮助我顺利完成毕业设计。我对这些热情帮助我的人满怀感激之情。

毫无疑问，首先我需要向我论文的导师齐官红助教和吴清强教授致谢。从上学期的最初的选题和研究准备、研究过程中的指导和最后论文全文的修改审查，两位老师都给予了无微不至的指导。在大学本科四年，我也要感谢我们学院的所有的教职工，是他们默默的付出使我在大学本科中能有如此丰富的学习生活。

同时，我也要感谢吴清强教授项目组的傅中杰学长和梁巧梅学姐，是他们激发了我们对于此次研究的最初的灵感。当我遇到任何难题的时候，他们总是愿意毫无保留的帮助我。傅中杰学长原先通过因马尔可夫模型研究过量化问题，随后其通过深度学习继续进行量化策略研究。

其次，我也需要感觉和我一起完成整个量化投资团队的所有的项目组成员。这是一个团结、高效并且协作共进的团队。项目组的每个人都尽自己最大的努力保证这个项目的顺利推进。也恰恰是这个项目的顺利推进，本文才有机会顺利完成。

我也要感谢大学本科四年以来一直帮助我的朋友们。在本次论文完成过程中，有很多同学无私帮助我克服了很多难题。通过和许多朋友的交流后，笔者才了解到了耦合隐马尔可夫模型并且继续进行研究。

摘　要

量化择时是指通过一系列算法分析后在单个或多个标的发出做多、做空指令，以获得低风险套利的交易策略。股票市场预测是一个非常经典的问题，近年来有许多人使用机器学习和技术对此课题进行了深入研究。一些特征使得这种建模与众不同，其中包括时间依赖性，波动性以及其他类似的复杂依赖关系。

为了解决以上提出的问题，隐马尔可夫模型（HMM）开始被使用于预测股票市场。本文中提出了隐马尔可夫模型（HMM）的方法。并使用此方法预测相关市场的股票价格。本研究应用 HMM 来预测指数。HMM 已被广泛用于模式识别和分类问题。然而，使用 HMM 来预测不确定事件并不简单。本研究中首先使用一个 HMM，对选定的指数过去的数据集进行训练。然后，再划定一定数量的数据集进行测试，准备预测。最后发现，如果合理使用 HMM 可以获取到可观的收益。

随后，本研究通过耦合隐马尔可夫模型制定了上证 50 的交易策略，同时使用上证 50 和沪深 300 指数进行耦合研究。本研究的策略涉及使用耦合隐马尔可夫模型（CHMM）对上证 50 和沪深进行建模。观察指标是涨跌幅和成交量，这些指标将用作模型的交易信号的触发器。在每次迭代中解码建模时，模型可以得到下一个最可能的状态和下一个最可能的观察值。本研究希望利用市场分析和模型中隐含的马尔可夫属性，用这些最可能的价值进行交易将产生超额收益。同时，证明了 CHMM 相比于 HMM 表现会更加优秀。

关键词：隐马尔可夫模型；耦合隐马尔可夫模型；量化择时

Abstract

Quantitative timing is the analysis of a series of algorithms to issue long and short orders in single or multiple targets to obtain a low-risk arbitrage trading strategy. The stock market forecast is a very classic issue. In recent years, many people have conducted in-depth research on this subject using machine learning and technology. Some features make this model different from others, including time dependence, volatility, and other similar complex dependencies.

In order to integrate the issues raised above, Hidden Markov Models (HMM) have recently been used to forecast and forecast the stock market. This paper proposes a Hidden Markov Model (HMM) method to predict the stock price of the relevant market. We use HMM to predict the index. HMM has been widely used for pattern recognition and classification problems. However, using HMM to predict uncertain events is not simple. In this study, an HMM was first used to train past data sets of selected indices. Then, a number of data sets are demarcated for testing and preparation. Finally, it is found that if you use the HMM rationally, you can get considerable benefits.

Subsequently, the study formulated the trading strategy of Shanghai Stock Exchange 50 using a coupled hidden Markov model, and used the Shanghai 50 Index and the Shanghai-Shenzhen 300 Index to conduct coupled research. Our strategy involves the use of coupled hidden Markov models (CHMM) to model Shanghai 50 and Shanghai-Shenzhen. The observed indicators are ups and downs and volume, which will serve as a trigger for our trading signals. When decoding the modeling in each iteration, we can get the next most likely state and the next most likely observation. This study hopes to use the Markov attributes implied by market analysis and models. Trading with these most probable values will generate excess returns. At the same time, it is proved that the coupled hidden Markov model performs better than the HMM.

Key words: HMM; CHMM; Quantitative timing

目 录

Contents

第一章 绪论

1.1 引言

准确预判股票、债券等市场是一个非常困难的事情，但是预测预判市场在将来一小段时间内的趋势以及上升下跌状态是相对可行的。量化投资是使用多种计量方法，从历史行情数据中寻找能获得超额收益的方法。

下图为量化投资方法和传统投资方法的概括图。

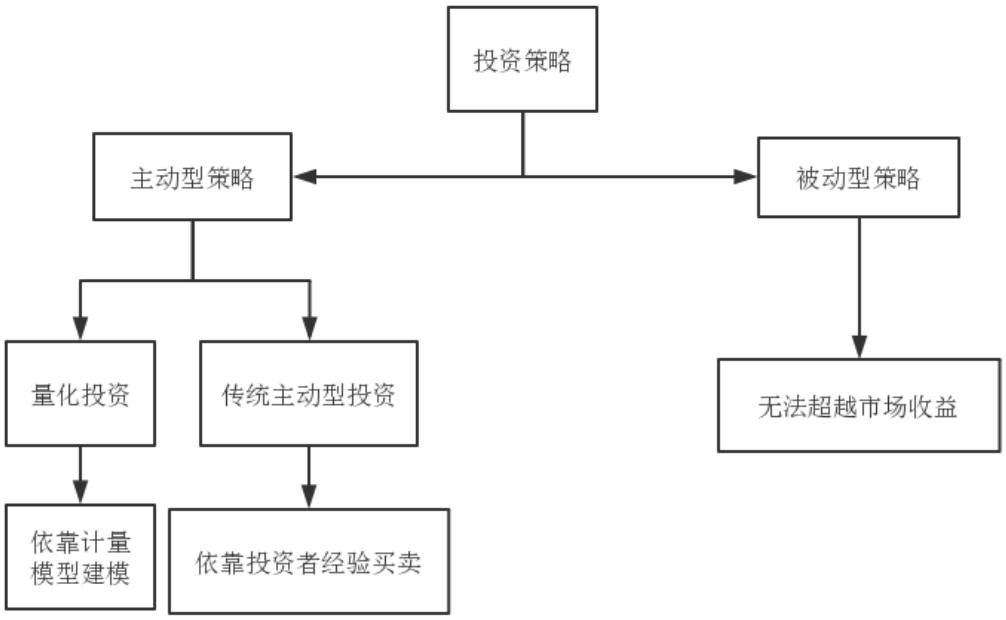

图 1-1 量化投资区别图

量化投资在国外有许多年的发展历史，但是在国内的大部分基金券商等买方中还很少被运用。在大部分涉及量化投资的公司中，量化投资仅仅是被当做投资研究中的辅助部分。通常由少数几位研究员在基本面研究的同时兼顾量化策略，并给出量化研究意见。由此可以知道，国内的量化投资研究道路道阻且长。

图 1-2 为笔者于 2016 年在香港招银国际资产管理有限公司实习时，公司内部的关于环保大健康行业的量化研究分析员给出的相关意见。其结合了量化分析

与基本面分析。在每天的投资委员会会议上汇报量化分析和基本面分析的汇总分析后进行做多、做空操作。其大环保大健康行业基金也获得了可观的收益。

股票名称	代码	ST RANKING 操作评级	每股收益一致预测 WIND	估值倍数短期目标PE	目标价 3个月内	现价 HKD	潜在上涨幅度	CATALYST 股价触发因素	LT RANKING 基本面评级
北控水务	0371.HK	3	0.33	13.3	5.5	4.95	11.1%	体外融资	5
中国光大国际	0257.HK	4	0.57	16.8	12	9.65	24.4%		5
东江环保	0895.HK	5	0.52	23.1	15	13.46	11.4%	国企保护	5
绿色动力环保	1330.HK	5	0.292	15.1	5.5	3.97	38.5%	A股上市进程	5
龙源电力	0916.HK	4	0.44	15.1	8.3	7	18.6%		5
华能新能源	0958.HK	4	0.24	11.7	3.5	3.01	16.3%		5
京能清洁能源	0579.HK	3	0.318	7.0	2.8	2.64	6.1%		4
中广核电力	1816.HK	4	0.16	13.5	2.7	2.43	11.1%	与神华合并	4
协和新能源	0182.HK	4	0.075	6.4	0.6	0.49	22.4%	低估值	4
金风科技	2208.HK	4	1.18	9.5	14	12.34	13.5%		4
昆仑能源	0135.HK	4	0.45	11.7	6.6	6.05	9.1%		4

图 1-2 量化策略图

一九七一年，美国一家基金公司发型了全球第一个量化基金，这开启了量化投资时代。现在，量化投资变为了世界上许多买方进行买卖研判的重要策略依据之一。

量化投资和一直以来的基本面投资分析也有着显著区别。量化投资是通过大量的历史数据作为训练集、测试集以及验证集，从而得出的卓越策略。而传统投资方法是通过对某一标的资产进行尽职调查从而得出的买卖结论。

尽管量化投资在我国还位于低级别阶段，有许多不足。但从国内的股票市场来看，恰恰是因为 A 股的发展时间短，相对于美国、欧洲等市场，属于非有效市场，沪深股市中存在着许多定价偏差的股票。与此同时，在国内非理性投资随处可见，比如复盘后的乐视网等股票。正是这些不理性的投资存在，才使得量化投资研究在国内股市有更多的发展机遇。

和原先的择时策略相对比，量化投资拥有三个的优点[1]：（1）拥有充足的精力：无论是基金公司中的专业研究员还是专业散户，他们都无法一个人覆盖整个股市中的所有股票。一个专业的买方研究员也无法一个人覆盖整个行业中的所有公司。大部分的行业研究员都是研究行业中的龙头股，但是许多投资机会往往也会出现在许多小市值股票中。但是量化投资可以同时监测、覆盖大量的股票，不

断的发掘获利机会。（2）充分利用胜率：在量化投资中，其不断的分析历史数据来找到规律，最终使用这个规律来获利。与此同时，量化投资也可以通过同时建仓多个股票来减小风险。（3）严格的纪律：在传统投资方式中，人们主观的人性与情绪波动对投资的收益影响非常大。在量化投资中，计算机按照策略执行并不会受到其他影响，因而摈弃了人性的弱点。

基于以上论述，量化的股票择时研究，寻找适合于国内资本市场的择时模型是非常必要的。

1.2 研究内容

1.2.1 研究目标

通过以上的论述可以发现，现在研究适用于我国市场的量化择时模型拥有着非常好的机遇。实际，这几年也逐渐有许多的证券公司、基金公司和资产管理公司在量化投资方面加大投入，致力于能够获得一个稳定收益的模型。

有许多论文的研究已经论证了金融市场是非线性的。有许多不确定的因素共同影响着股票价格的波动。与此同时，许多研究表明股票价格的波动存在着规律，合理的使用股票价格的历史行情以及其他基本面信息可以用来大致预测未来股票价格。机器学习在很多其他领域已经证明了其是针对非线性数据建模的有效策略。比如，搜索、自然语言处理、图像识别等。因此，选择机器学习中的算法来编写股票策略拥有了理论基础。许多的论文也在机器学习量化方面进行了不同方面的研究，得出了机器学习在量化投资方面的有效性的结论。因此，本文将主题定为基于机器学习的量化择时研究，以达到对于单个股票的买卖择时进行更有效的操作。

1.2.2 研究意义

（1）理论意义

本文把其他方面研究中非常成熟的机器学习的策略应用到量化择时。同时，结合机器学习与基本面分析手段，得到了更有效的量化择时策略。本文为量化择时领域提供了更完善的视角以及更全面的分析。本文中，本研究运用了许多机器

学习算法来预判股票的走势。同时，本研究也通过隐马尔可夫模型和耦合隐马尔可夫模型来预测股票价格，随后回测。

（2）实践意义

一个卓越的量化择时模型能帮助基金经理、行业研究员等解决许多数据分析问题，帮助他们克服自己在关键时间节点对于买卖的犹豫。同时，量化择时模型也拥有很好的现实意义。券商、基金等可以通过量化择时模型来操盘自营资金，为公司贡献丰厚收益。对于卖方来说，也可以通过发行量化基金产品进行募集，做大资产管理规模。在扩大规模的同时，通过管理费、申购费和超额利益分成等可以拥有无风险收益。如中欧基金发行了中欧量化驱动混合型基金，国泰君安发行了国泰量化收益灵活配置混合基金，并且都募集到了大量资金。

1.3 论文组织结构和研究方法

本文中一共有六个章节。第一章讲述了整体研究思路和研究内容。

第二章详述了在机器学习领域的各个研究的现状。从国外研究现状到国内研究现状，从历史研究中总结了经验。同时，结合了许多股票市场的其他研究方法，从而得出了本文的研究方向。

第三章综述了本研究中的基于隐马尔可夫模型的量化择时模型的研究过程。首先介绍了 HMM 的理论基础。随后，构建了 HMM 的择时模型和回测模型，得出了预测结果和回测结果。

第四章详细介绍耦合隐马尔可夫模型的择时策略研究。最初，本章论述了 CHMM 的理论基础和研究过程。然后，搭建了基于 CHMM 的择时模型和回测模型。同第三章一样，也得出了预测和回测结果。

第五章中，对多个量化策略进行了对比。将 HMM 和其他机器学习量化策略对比得出了优缺点。然后，对比了基于 CHMM 和 HMM 的策略，最后总结了各个策略的优缺点。

第六章总结了全文，并且对接下来的研究提出了展望。

本文主要使用方法如下：

（1） 文献与理论研究

撰写本文前阅读了大量的量化择时相关论文以及博客，了解了全球量化择时

研究概况。通过许多机器学习教程，笔者学习了机器学习的基础排序原理和相关算法。通过理解算法和原理对本研究中的模型进行初步构建。在浏览文献的同时，也试图去寻找现有方法的不足之处，并且作为研究的改良点。

（2） 实证研究

首先，本研究选取了时间跨度为十五年的沪深 300、上证 50 等数据。模型内将其分为训练集和测试集。本研究通过 Python 和 R 语言进行编程和数据处理以检测策略的是否有效。

后面几章介绍了利用 HMM 构建量化择时策略以及回测结果和分析。因为 HMM 存在一些缺点和不足，所以本文在第四章更进一步优化了 HMM，构建了耦合的 HMM 并且回测。

本文技术路线如下。

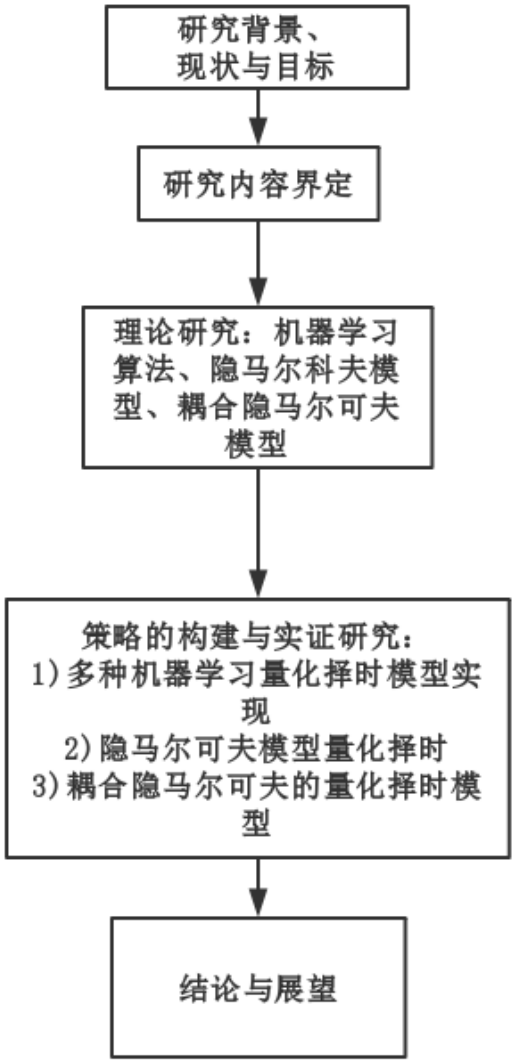
研究背景、
现状与目标
研究内容界定
理论研究：机器学习
算法、隐马尔科夫模
型、耦合隐马尔可夫
模型
策略的构建与实证研究：
1）多种机器学习量化择时模型实
现
2）隐马尔可夫模型量化择时
3）耦合隐马尔可夫的量化择时模
型
结论与展望

图 1-3 论文技术路线图

第二章 基于机器学习的量化择时研究现状

2.1 多种机器学习算法于股票市场的应用现状

2.1.1 国外研究现状

随着量化领域研究的深入，很多研究者已经将一些机器学习领域运用于金融市场，其中包括股票市场、外汇市场和黄金市场等。

自从查尔斯在一百多年前提出道氏理论后，技术分析在世界上迎来了黄金发展时期。当时的测试结果以及现实交易表明了量化方法的获利是不菲的。欧文[5]等在上世纪末为期货基金构建了自动化程序交易系统。最后结果表明该交易系统中的过半信号是有效的，胜率超过百分之五十。涅波茨[6]改进了这一系统，将此运用于道琼斯指数中。他发现了一百五十日均线有预测意义。随后的几十年来，人们在量化的道路上继续深入探索。

Khalid Alkhatib 在论文[2]中提出 KNN 算法可以有效地预测股市。其结果表明，KNN 算法鲁棒性好，误差小，结果合理。此外，根据实际股价数据，其预测结果接近实际未来走势。然而，由于信息系统技术的发展和充实，利用更加先进的预测模型来帮助市场预测，还有很长的路要走。此外，这可能削弱投资在约旦市场的吸引力，最终削弱市场回报。研究还表明，如今的数据挖掘技术为金融界提供了强有力的股票市场运动的预测分析能力。然而，KNN 算法仍然有许多的缺点。KNN 算法计算复杂度很高。其空间复杂度也很高。如果使用 KNN 算法，也有可能出现样本不均问题。同时，一个分类算法是无法太过精准的推算出未来股价涨跌走势。

同时，来自 CMU 的 Dung Tien Nguyen 教授[3]也提出了使用贝叶斯网络来进行股票价格的预测。他提出在一些基础算法诞生之后，贝叶斯推理在学界中一直非常流行。后验分布的有效性在很大程度上取决于似然函数和先验分布的组合。在某些情况下，可以获得后验分布的封闭形式。随后其使用贝叶斯网络来判断走势并且获得不错的收益。然而，朴素贝叶斯仍然存在着许许多多的问题。首先，NB 算法内的先验概率需要人们提出假设。由于人们是通过先验概率和输入来决

定后验概率因此决定分类，所以分类决策将会存在些许错误。因此，NB 模型的有效性相当一大部分取决于作者对于先验概率的定义。所以，该模型的可移植性也会被遭受质疑。

也有许多研究者使用支持向量机（SVM）来预测股票价格以获得收益。Huanhuan Yu 和 Rongda Chen[4]提出了 SVM 分类的股价预测模型。SVM 分类的准确性取决于训练集。其对数据非常敏感。文中为避免直接使用复杂高维财务比率，将主成分分析（PCA）引入 SVM 模型，提取低维高效的特征信息，提高训练的准确性和效率，同时保留特征初始数据。实证结果表明，基于支持向量机的规范化标准化后的 PCA 中，选股模型在训练集中达到 75.44%，在测试集中达到 61.79%的整体精度。此外，PCA-SVM 选股模型显着提高了股票投资组合的年收益率，超过了上海证券交易所 A 股指数。然而，当遇到大规模数据时，SVM 将消耗大量的系统内存与时间。但是，在股票预测领域，仅仅当训练过很多数据后才可以得出相对优秀的策略。另一方面，SVM 只能将状态二分类，而无法多分类。在金融市场中，人们不可以简单的将市场区分为牛市和熊市。

直到 2003 年，第一次有人将机器学习应用于量化交易上面。Kim K[7]使用 SVM 预测股价。其将结果与 BP 神经网络对比得出结论，SVM 可以发掘的有效获利模型。随后的十几年内，有大量关于机器学习的量化研究成果。

2.1.2 国内研究现状

国内许多研究人员也在量化投资上面进行了很多研究。在许多券商、基金内，有大量的研究员专门研究量化交易。现在市场上也涌现出许多专门做量化的私募基金。然而，因为我国一行三会对金融衍生品交易的严格管制。我国量化的总体市场规模并无法跟欧美等国对比。尽管受到了严格的管控，我国的量化市场总体发展并未因此受限制。

吴微[8]研究 BP 神经网络，对标的分类。他们结合了股市的特征，对指数的涨跌进行预测。他们通过大量的实验证明，机器学习对特殊标的的预判是有价值的。随后，周琳杰[9]运用重叠抽样方法，对上海和深圳股市的部分股票进行历史交易研究。此文研究了不同时间段的互相结合的情况，将其与其他特征相结合，最后发现了获得超额收益的方法。楼迎军[10]研究了遗传算法等在多个股票集中

的有效性。他们发现了可以有效提高收益的获利组合。吕琦[11]使用了基于时间序列的 SVM 股价预测方法。他们建立了股票回归模型，大量实验证明了 SVM 比许多其他方法有更好的收益，但是精度仍然存在改善。

尽管有大量的国内外研究证明了机器学习在量化投资领域的有效性。但是，大部分的研究局限于多个股票的量化选股模型，对于量化择时领域的研究仍然不足。同时，基于很多机器学习算法的量化择时模型，都有风险高、鲁棒性弱、交易过于高频、精度低等缺点。因此，本文将要探索一个更有效、低风险和相对低成本的交易策略。

2.2 股票技术分析方法现状

在股票市场内，应用较广泛和较容易入手学习的策略就是技术分析方法。大部分的方法，是通过每天的股票数据（包括成交价、交易量等）进行计量分析。然而，现行的 A 股策略中，如果仅仅考虑技术面指标而不考虑基本面指标，是不完全科学的。

石赛男[12]论述了如何使用 MACD 来检验股市的有效性问题。王俊华[13]使用股市的 K 线图来选择时机买点和卖点，并且得到了不错的结果。陈收[14]运用成交量来优化他们的投资组合，并且优化过的投资组合的收益率会有显著的提升。AR Gallant[15]在上个世纪首先提出了量价关系。其研究也被学界大量的引用和学习。

M WU[16]设计了模型并利用算法对股票市场中的三个指标进行了检验：移动平均收敛发散（MACD）、相对强度指数（RSI）和随机振子（KDJ）。测试数据为上海（SH）股票指数和深圳股票指数的 8 年数据，通过满足上述三个振荡指标，建立股票买卖模型，通过将不同的时间段分配为 3 天到一个月来计算涨跌幅度。结果表明，三个指标结合后可以预判短期变化。此外，还证明了不同的市场周期对指数效率的影响较小。

尽管本文将会分析如何找到合适的交易时机，选股仍然将影响一个择时策略的收益率。当在一个总体趋势向上的股票中进行量化择时，获得收益率将明显高于整体趋势向下的。

笔者还努力奉行一些简单的原则，例如：

第一，便宜才是硬道理。

第二，定价权是核心竞争力。

第三，独角兽（腾讯、茅台等公司早已经在各自行业中独占鳌头，但是这些股票仍然能给投资者带来巨大收益）。

因此，本文中的进行研究择时的个股标的都是经过时间验证的绩优股。同样的理论也可以应用在未来的量化选股模型中。

图 2-1 为 MACD 技术指标分析图。

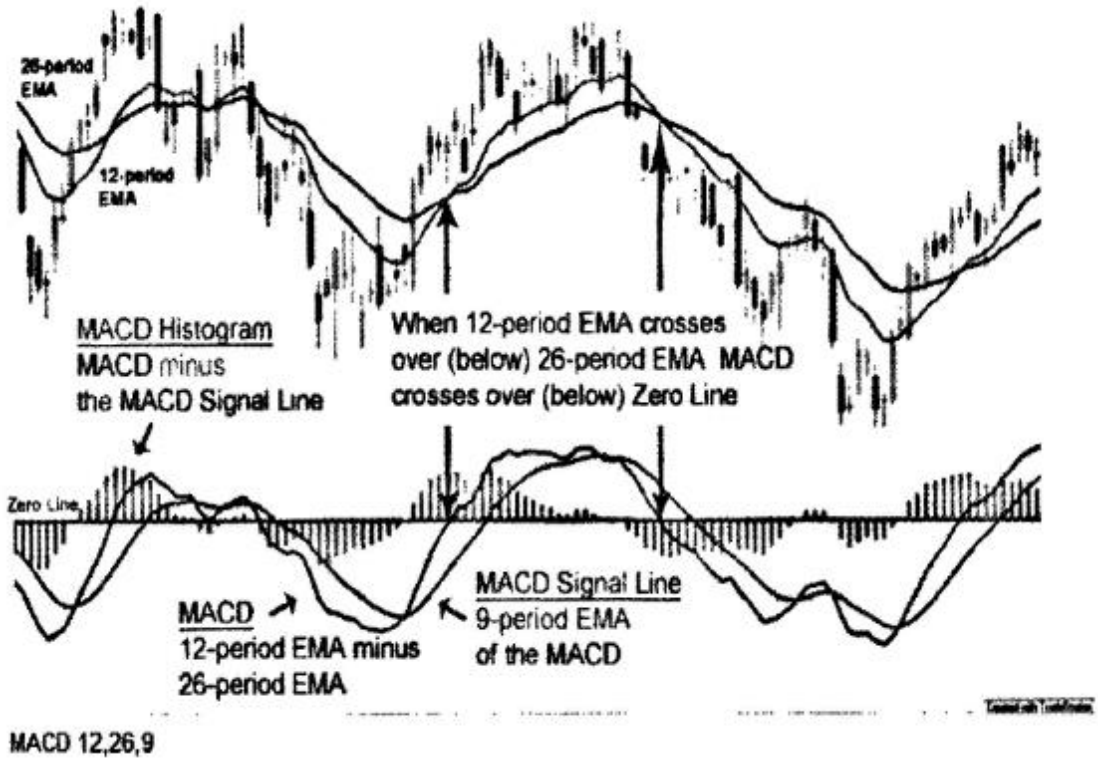

图 2-1 MACD 指标图

现有的很多买方或者散户也依靠某些指标进行买卖操作，并且能取得不错的收益。

2.3 创新点

由于以上的历史研究，笔者发现了 HMM 和 Coupled HMM 的股票择时研究的空缺。所以笔者选择了以下几点进行研究。

（1）将机器学习与其他股价分析理论结合，建立了 HMM 的量化择时模型。通过以上模型解决了其他机器学习算法的模式定义模糊、特征选择困难的难题。解决了某些动量反转的策略中，会有参数难以确定的问题。

（2）构建了基于 CHMM 的量化择时模型。此模型解决了 HMM 中的信息量不充分的缺点。此模型也加强了的风险控制能力、盈利能力，减少了交易次数。

（3）发现了更多的基于 HMM 和 CHMM 的可解决的问题。如果将来有更加完善的两种模型，会有更完美的收益。本文也提出了多耦合的 HMM，基于多个标的的耦合。将成为笔者继续探索的主题。

2.4 本章小结

由以上可以得出，国内外对于量化投资方面都进行了大量研究。但是基于 HMM 的择时策略和 CHMM 的择时策略尚未在公开文献中发现。因此，笔者选择了此命题来深入探究。

第三章 基于隐马尔可夫模型的股票择时模型研究

3.1 隐马尔可夫算法

在过去几十年中，隐马尔可夫模型被广泛运用在语音识别[17]、图像识别[18]等多个领域。近年来许多机构发现，HMM 在金融市场中仍然能帮助机构获取可观的收益。

首先，人们可以把金融市场分为牛市、熊市和震荡市三类。实际，按照股市状态，还可以更加细分股票市场的状态，但是这些状态难以观察，而且难以转换，所以需要一个过渡。图 3-1 为股票择时 HMM 模型量化金融示意图。HMM 假设当前交易日 t 的市场状态S_t只依赖前一天的状态S_{t-1}，市场状态间有着转移概率$P(S_t|S_{t-1})$。从而生成了转移概率矩阵 A。矩阵的每个元素代表着从一个市场形态变化到另一个市场形态的概率。

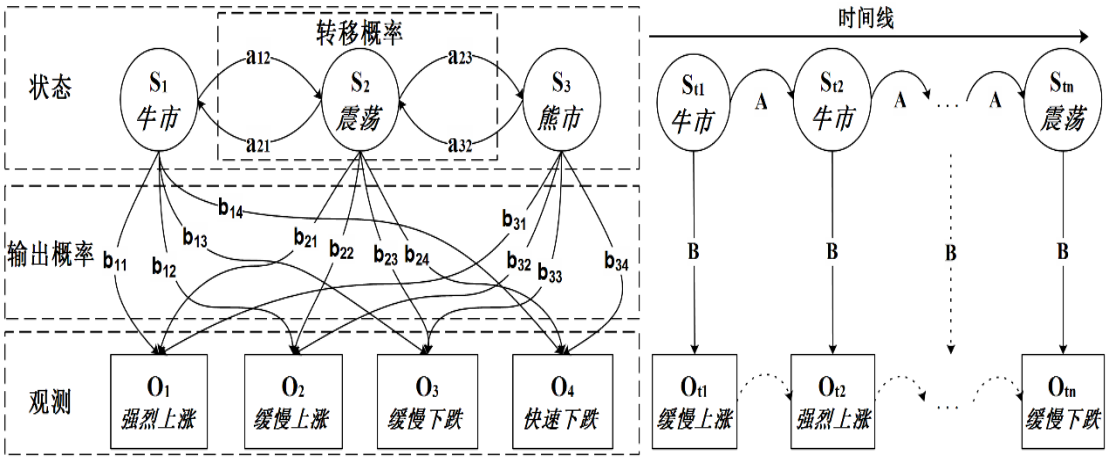

图 3-1 隐马尔可夫模型量化金融示例

在 Markov 模型中，Markov 链的状态与市场状态是一一对应的。如图 3-2 和 3-3 所示。

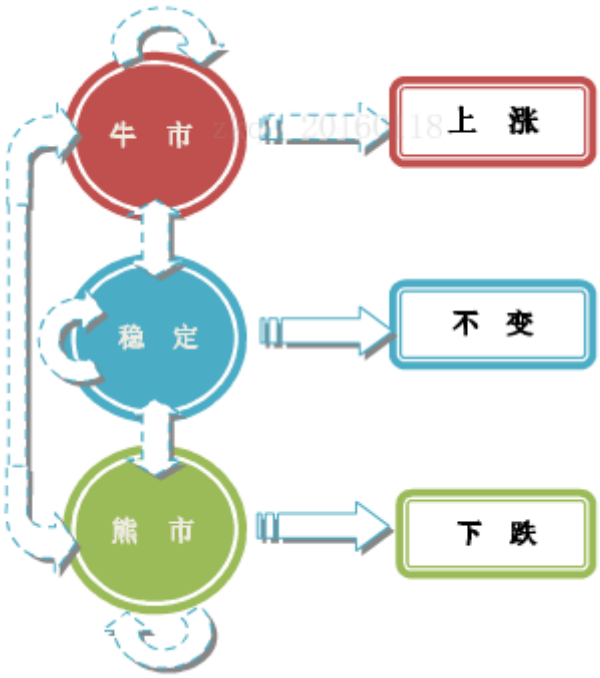

图 3-2 马尔科夫模型

在 HMM 中，就会有不同的对应情况。如下图所示。

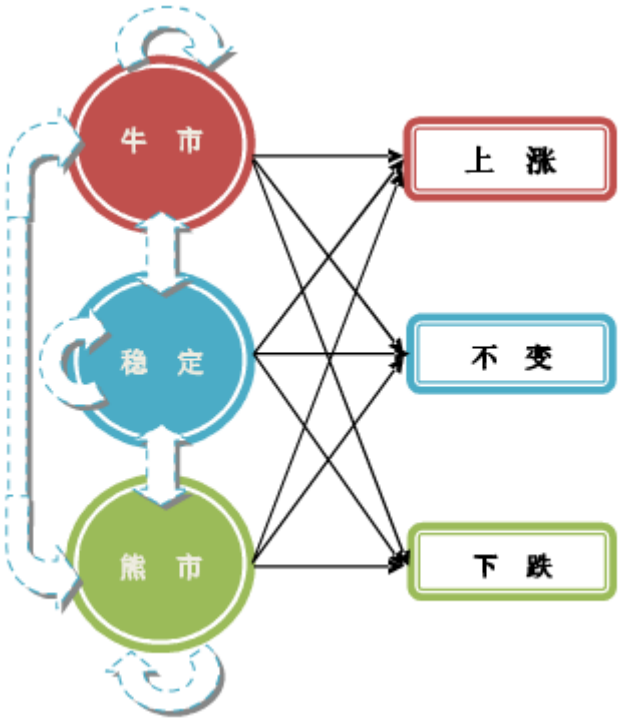

图 3-3 股市中的 HMM

同时，市场的每个状态都存在着相对应的观测状态的输出概率分布。如以下公式。

$$P(ot|st, st-1, \dots, s1, o1, \dots, ot-1) = P(ot|st) \quad \text{(公式 2-1)}$$

$$B = \{b_j(o_t)\} = P(O_t = o_t|S_t = j). \quad \text{(公式 2-2)}$$

当得到了转移概率和输出概率后，也就可以直接得到初始状态。

$$\pi_i = P(S_1 = i). \quad \text{(公式 2-3)}$$

所以，$\theta = (A, B, \pi)$可以用来表示一个完整 HMM 模型。在 HMM 训练过程中，当前并没有最优解的方法。一般来说，人们选择鲍姆-韦尔奇（Baum-Welch）算法，辅以 EM 原理来确定部分最优的θ。

在预测部分，本研究使用维特比(Viterbi)算法。首先，本文定义状态空间为 S，初始概率为π_i，两个状态之间的转移概率为a_{ij}，观测序列为$o_1, \dots, o_T$。然后，把$V_{t,k}$当做在已有的 t 个观测变量之下，概率最大的以状态 k 结尾的观测状态序列。所以当前状态S_T，可通过以下式子得到。

$$V_{1,k} = P(o_1|k) \cdot \pi_k, \quad \text{(公式 2-4)}$$

$$V_{t,k} = \max_{s \in S}(P(o_t|k) \cdot a_{s,k} \cdot V_{t-1,s}), \quad \text{(公式 2-5)}$$

$$S_T = \operatorname{argmax}_{s \in S}(V_{T,s}). \quad \text{(公式 2-6)}$$

HMM 会保存每个和当前状态具有强关联的前一交易日的数据。这样将会简化建模过程、减小复杂度。HMM 中假设每个观测量只和当前的隐藏状态相关，人们对于市场的观点恰恰与此相同。通过历史数也可以发现，在显著的牛市和显著地熊市期间，基准收益率等各项指标会有根本性差异。因此，人们可以推断观测变量将会拥有不同的分布，当变量处于不同的市场状态。相比深度学习等模型，HMM 有更强的可解释性。

3.2 基于隐马尔可夫的量化模型

整个模型的流程图如图 3-4。本研究首先从开源的 Python 数据接口获得股票的各种数据，其中包括了每日的开盘价、收盘价、最高价、最低价以及每日成交量。随后，HMM 通过计量计算得到各个潜在特征。然后，本研究使用 HMM 对每个隐藏状态进行训练，得到各个状态，然后选取有效特征。最后，利用选取出的特征集合来对给定状态数目的 HMM 进行训练。回测时，HMM 会在某些时候

发出买卖信号。

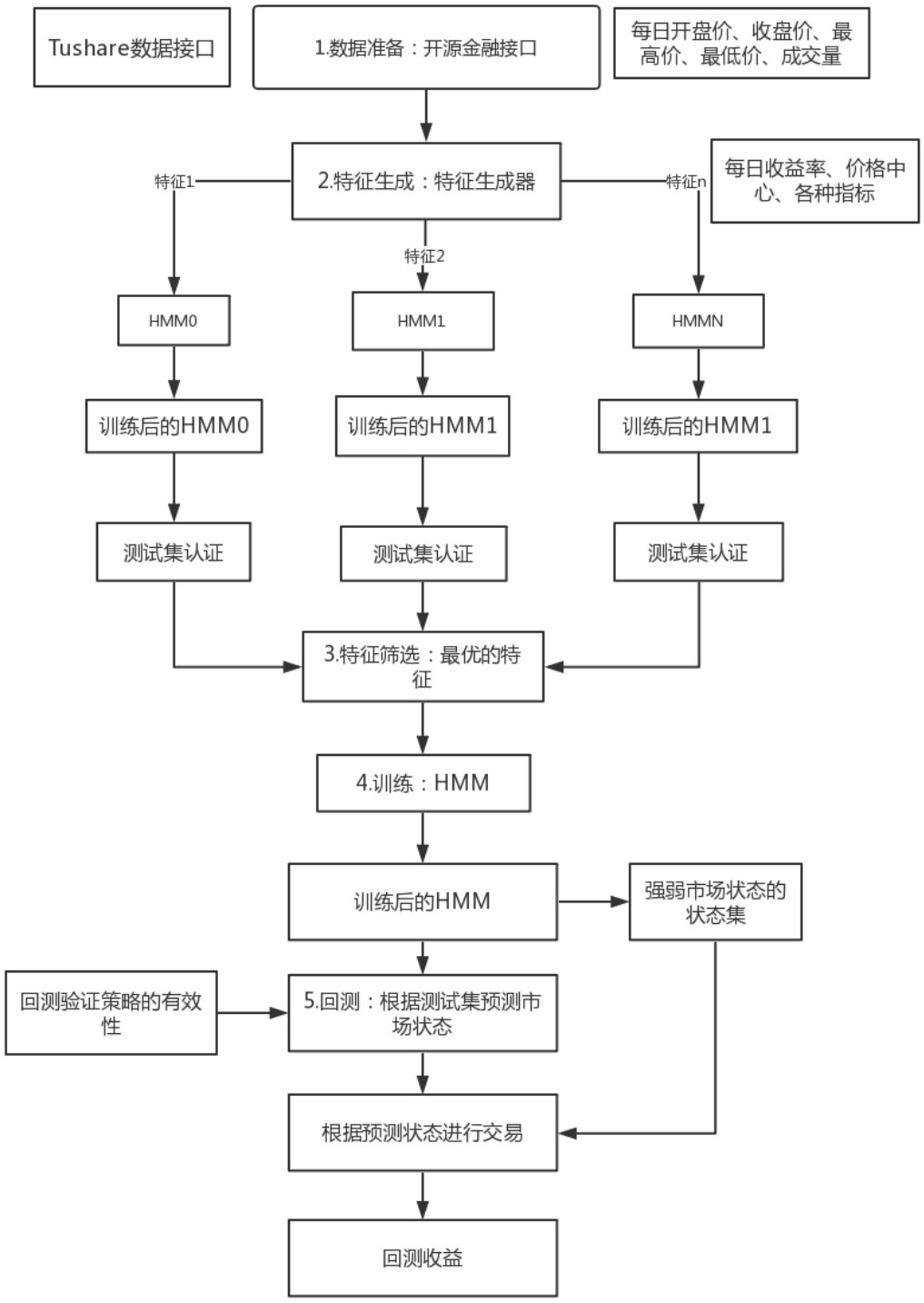

图 3-4 模型流程图

3.2.1 特征准备与选择

笔者使用了经过计量运算后的特征来代替从开源接口所获得的原始历史数据。在特征提取方面，本模型将特征分为技术面特征和基本面特征。技术面特征包括相对强弱指数（Relative Strength Index ，RSI）等。因为广泛的运用，这些指标也具有很强的鲁棒性。这些指标同时也被世界市场所公认。基本面特征包括了能反映国家整体经济发展的宏观指数。但是，不同国家可能拥有不同的计量标准，所以为了使得模型通用。本研究中主要使用技术面特征。

在特征筛选方面，本模型的首要任务是从所有的潜在特征中选择出高质量、更有代表性的特征。多个特征可以在多个方面来反映不同的市场状态。如果同时聚合这些不同的特征，它们可以得出相对卓越的结果，达到稳定收益。

根据以上图，在生成每个潜在特征后，为每个特征匹配独一的 HMM 然后进行检验。通过观察每个单独特征模型在测试集上面的收益率，模型可以对比这些特征。将这些特征相对比，模型也可以得到相对更有效的特征。在量化择时方面，一般使用、策略收益率、最大回撤和夏普比率等特征将成为人们的评判一个策略是否有效的依据。模型将指标占比权重设为w_t，指标的评判值设为c_i。根据以下公式，模型可以得到相关性强的特征，从而将它们共同训练得出最终的模型。

$$\text{Score} = \sum_{i=1}^{n} c_i w_i \qquad \text{(公式 2-7)}$$

3.2.2 状态选择

在大部分的 HMM 的金融市场运用中，人们都选择使用五状态模型。通过以上的特征选择，模型可以得到许多状态。因此，该模型需要识别哪些状态可以获得足够收益，哪些状态可能让模型亏损。模型需要为每个状态执行买入操作（只选择单个状态进行测试），然后得出它们的收益率。此时，这些状态可以被简单的分为获利（累计收益率大于零）和非获利（累计收益率不大于零）。然而，并非所有获利的状态都是有效状态。模型需要对获利的状态进行进一步的筛选，从而得出强力有效的状态。如果在测试期内有 N 个交易日，第 N 个交易日的累计收益为r_n，模型可以通过以下式子来判定状态是否强有效。

$$r_T > 15\% \ or (r_T > 5\% \ and \ r_T > -3\% \ for \ any \ t \leq T) \qquad \text{(公式 2-8)}$$

以上式子定义了两种获利状态，前面一类是具有明显的获利能力的状态，它的收益率明显为证。后面一类是微弱获利的情况，对于这种情况模型需要进一步分析。大部分时候短期的震荡行情、短期的波动行情都有可能带来微弱获利的情况，所以这种状态也可以被识别为盈利。

然而，一些获利的状态在某些时候可能会变得无法获利，一些状态也可能在特定时期从亏损变为获利。这是由于模型对于状态的定义不准确，所以不可避免会有一些误差存在于状态识别中。与此同时，状态的多少也有可能影响状态的识别。当设定更多的状态去识别股价的特征时，就越有可能出现不确定的状态。对于这个问题，需要建立一个能够筛选状态的机制。系统需要对每个状态每天的收益进行监控，套入上列判定式。如果某个状态能够达到上列判定式要求，其将会被添加为候选状态。同时，该状态也会被执行买入。相反，如果某个状态不能达到上列判定式要求，会被抛弃。

整个模型全部的最终状态需要通过全面性回测来进行筛选。为了获得更加有效的状态集合，需要将其中表现优异的状态筛选出来。将其与其他获利可观的状态共同组合获得一个状态数以成为模型的参数。

3.2.3 回测模型

随着量化金融的发展，出现了越来越多的在线量化策略网站。这些网站可以为用户提供在线策略编辑、在线回测、仿真交易和策略分享等功能。但是，对于跨模型的策略构建和对比来说，本地回测模型将会变得异常关键。在本系统中使用了本地回测模型对数据进行回测。

尽管中国内地的股市已经经历了三十年的时光，但是许多研究仍然可以证明[19]，中国的大部分 A 股股票是非有效的。所以，本文选择了沪深 300、上证 50 等指数作为研究标的，因为这两个指数中的样本股相对稳定性高，指数也调整设置了缓冲区。也有研究[20]得出上证 50 指数相对大部分个股的有效性。选用这两个指数可以有效的避免一些小盘股、次新股和业绩差的股票等黑五类股进入训练模型。

根据上述指数训练到的预测的市场状态，模型将对相对应基金进行策略交易。首先，需要定义将初始资金多少。随后，根据每日的状态的预测，对第二天的股

票进行做空与做多。由于我国各类市场中，做空、融券业务有各种各样的限制。所以在本回测系统中，只存在买入以及平仓（卖出）操作。因此，本回测系统无法对预测到的下跌状态进行价差套利。同时，由于 A 股市场是 T+1 制度，所以系统对于数据的选取为日频数据。如需要把次回测系统转换为境外股市(T+0)的回测,数据可以改为十分钟为单位等，也因此可以进行更高频的交易。

优秀的投资回报主要有三个方面组成。

（1） 策略收益率和基准收益率的对比。

由于许多指数随着时间的推移，国家整体经济向好，指数基准收益率已达到了相当可观的数值。因此策略的收益率不能仅仅与盈亏线进行对比，更需要与指数的基准收益率进行对比。当策略收益率超过基准收益率的差值，业内称之为超额收益。也就是单个策略战胜市场的收益。

（2） 对于风险的控制。

任何一个策略都需要有相对应的风险控制机制，如果回撤过大代表策略的风险控制机制不够完善。一个优秀的量化择时模型应该是相对低风险、高收益、低回撤的模型。

（3） 对于交易费用的控制。

在本研究中模型默认交易费用为 0。据调查，国内大部分券商的佣金最低底线为交易费用的万分之三。对于大额交易而言，这种小额花费基本能够被收益所覆盖。但是，如果运用到国外市场的高频交易，控制交易花费会成为重要的成本控制手段。

3.3 实验结果与分析

3.3.1 实验数据

本文中选择了上证 50 指数作为研究标的，因为单个股票会有更大的波动和会受到更多方面的影响。上证 50 指数是根据公认合理的方法，挑选上海证券市场相对更有价值的 50 只股票股本加权平均得到的市场指数。本文中的原始数据囊括了数据日期、开盘价、最高价、最低价、收盘价、成交量。

3.3.2 特征筛选

本文中选取了 MACD（指数平滑异动移动），调用伪代码和算法为以下。

```
get_macd
    hist ← talib.MACD(np.array(close), fastperiod, slowperiod, signalperiod)
    hist ← 2 *hist
    macd ← hist[max_period:]
```

也选取了 ATR（平均真实波动幅度），伪代码如下。

```
get_atr
    max_period← 10)
    atr ← talib.ATR(np.array(high), np.array(low),     np.array(close), timeperiod)
    maxlog ← boxcox(atr)
    atr ← atr[max_period:]
```

RSI 指数，伪代码如下。

```
get_rsi:
    rsi ← talib.RSI(np.array(close), timeperiod=period)
    maxlog ← boxcox(rsi)
    rsi ← rsi[max_period:]
```

N 日绝对收益率，代码如下。

```
get_return
    period ←1
    max_period ←10):
    绝对收益率 ← ((np.array(price_series[period:])-np.array(price_series[:-
period])) /np.array(price_series[:-period]))[(max_period-period):]
```

此为 N 日对数收益率的代码实现。其中，price_series 为价格序列、period 计算周期，n 日对数收益率、max_period 为最大计算周期，所有属性中最大的计算周期。

```
get_logreturn,
    period ←5
    max_period ←10
    对数收益率 ← (np.log(price_series[period:])-np.log(np.array(price_series[:-
period])))[max_period:]
```

以下为 N 日绝对收益率。

```
get_return
    period ← 1
    max_period ← 10
    绝对收益率 ← ((np.array(price_series[period:])-np.array(price_series[:-
period])) /np.array(price_series[:-period]))[(max_period-period):]
```

以下为 N 日对数成交量变化率。

```
get_logvol
    period ← 5
    max_period ← 10
    对数成家量变化率 ← (np.log(np.array(vol_series[period:]))-
np.log(np.array(vol_series[:-period])))[(max_period-period):]
```

以下为 N 日成交量变化率。

```
get_volume_change_rate
    period ← 1
    max_period ← 10
    成交量变化率 ← ((np.array(vol_series[period:])-np.array(vol_series[:-
period]))/np.array(vol_series[:-period]))[(max_period-period):]
```

以下为计算 N 日振幅。

```
get_amplitude
    period ← 5
    max_period ← 10
    high_period ← np.array(high)      # 每日最高价
    low_period ← np.array(low)        # 每日最低价
    for i ← 1 to 10
        high_period ← np.maximum(high_period[1:], np.array(high)[:-i])  # 计
算 N 日的最高价
    for i ← 1 to 10
        low_period ← np.minimum(low_period[1:], np.array(low)[:-i])      # 计
算 N 日的最低价
    amplitude ← (high_period[1:]-low_period[1:])/np.array(close[:-period])
    振幅 ← amplitude[(max_period-period):]
```

以下为计算 N 日价格轨迹效率。

```
get_price_efficiency
    period ← 5
    max_period ← 10
    # 绝对位移长度
    abs_dis ← np.abs(np.array(price_series[period:])-np.array(price_series[:-
period]))
    # 价格路程长度
    path_dis ← 0
    for i ← 1 to period+1
        IF i == period THEN
            Do path_dis ← path_dis + np.abs(np.array(price_series[i:]) -
np.array(price_series[(i-1):-1]))
        else:
            Do path_dis ← path_dis + np.abs(np.array(price_series[i:])-
np.array(price_series[(i-1):-1]))[:(i-period)]
    轨迹效率 ← (abs_dis / path_dis)[(max_period-period):]
```

以下为计算 AR 人气指标。

```
# AR 人气指标
get_ar
    period ← 14
    max_period ← 10
    ar_result ← []  # AR 值数组
    df_len ← len(open_p)          # 获取数据长度
    if df_len < period:         # 数据长度小于计算周期
        return None
    high_open ← np.array(high) - np.array(open_p)          # 当天最高价-当天
开盘价
    open_low ← np.array(open_p) - np.array(low)             # 当天开盘价-当天
最低价
    for i ← 1, df_len + 1
        if i <= period:          # 当 i 小于数据长度(无法计算 AR)
            ar_result.append(0)        # 将 0 赋值给 AR 数组的头 n 天
            continue
        h_o ← high_open[(i-period):i]        # 获取从第(i-period)天至第 i 天的最
高价-开盘价
        o_l ← open_low[(i-period):i]          # 获取从第(i-period)天至第 i 天的
开盘价-最低价
        h_o_sum ← sum(h_o)                    # 计算 N 天内的最高价-开盘价的
和
        o_l_sum ← sum(o_l)                    # 计算 N 天内的开盘价-最低价的
和
        if o_l_sum == 0:                            # 保证分母不为 0
            o_l_sum ← 0.01
        # 人气指数（AR）
        ar ← 100 * h_o_sum / o_l_sum
        ar_result.append(ar)
    # 返回 AR 结果
    AR ← np.array(ar_result)[max_period:]
```

以上这些特征都经过系统单独测试，并且有不错的表现。某些特征会在特定状态之下获得很可观的收益。

3.3.3 状态选择

状态选择的目的就是确定某种意义上的“最佳”状态。在学界中，有两种最常

用的选择方法。

（1） AIC=$-2\log L+2p$

（2） BIC=$-2\log L+p*\log N$

其中，L 是拟合模型的可能性。p 是参数的个数。n 是数据点的个数。$-2\log L$ 项随模型复杂度的增加而减小。同时，惩罚因子 2p 或 p*logN 随着复杂性的增加而增加，当 n>e^2=7.4 时，BIC 会用较大的惩罚因子。

首先，得出隐含状态的数量，相当于找出最大边缘似然值的模型。本研究中使用了 BIC 来得出隐状态数量。该模型就是在不完全的数据下，对隐藏的状态采用概率估计。随后，使用贝叶斯公式运算发生概率。最后，使用期望值和调整后的概率来选择最好的可能。

于是，本文中的模型选择使用了五状态模型。根据图 3-5 的状态转换图可以发现，状态 2（绿色）和状态 3（红色）代表着该指数处于上升通道。状态 0（蓝色）和状态 4（紫色）代表着该指数处于下跌通道。状态 1（橙色）代表着该指数处于震荡、波动趋势。可以观测到，状态 0、2、3 都可以带来可观的收益。其中 2014 年底的阶段牛市的启动，状态 2 和 3 都能及时的识别。图 3-6 为各个状态的收益图。

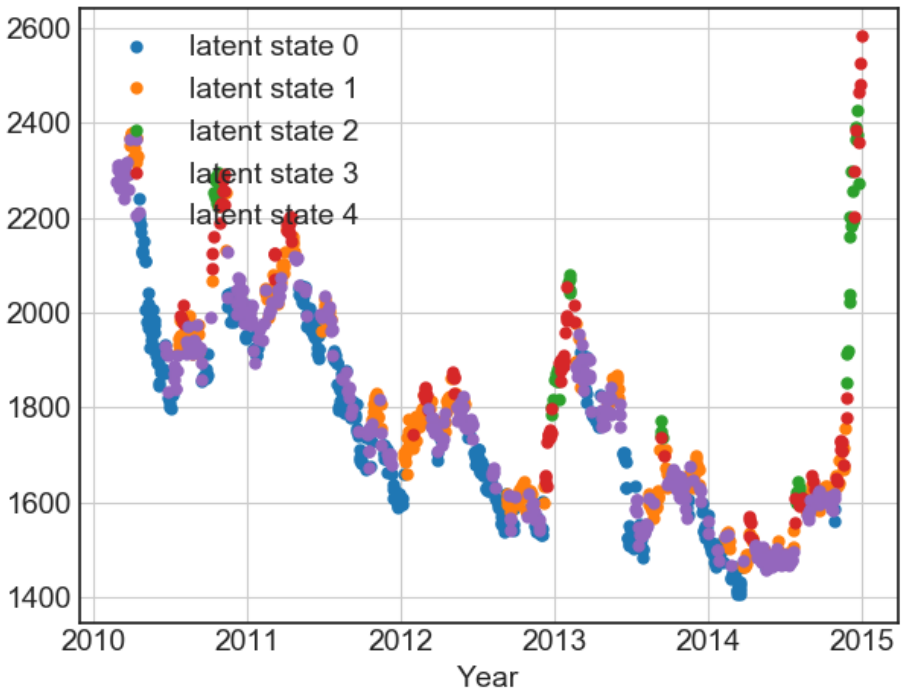

图 3-5 状态转换图

图 3-6 状态收益图

最后，本文选择 2015 年 1 月 1 日开始至 2016 年 12 月 31 日作为该模型的回测区间。在这个区间内 A 股度过了牛熊市的变化，因此模型可以被充分的检测其有效性。同时，在回测区间内包含更多的市场状态也可以更充分的发挥每个状态的效用。

以下是该策略的部分交易详细表。

表 3-1 HMM 策略部分交易图

交易指示	交易时间	持仓时间	
OpenLong:20150120			
买入:	20150120		
CloseLong:20150210			
卖出:	20150210	持仓:	15 天
OpenLong:20150212			
买入:	20150212		
CloseLong:20150213			
卖出:	20150213	持仓:	1 天
OpenLong:20150216			
买入:	20150216		
CloseLong:20150217			
卖出:	20150217	持仓:	1 天

表 3-1（续）

OpenLong:20150302			
买入:	20150302		
CloseLong:20150306			
卖出:	20150306	持仓:	4 天
OpenLong:20150320			
买入:	20150320		
CloseLong:20150323			
卖出:	20150323	持仓:	1 天
OpenLong:20150330			
买入:	20150330		
CloseLong:20150402			
卖出:	20150402	持仓:	3 天
OpenLong:20150408			
买入:	20150408		
CloseLong:20150417			
卖出:	20150417	持仓:	7 天
OpenLong:20150515			
买入:	20150515		
CloseLong:20150519			
卖出:	20150519	持仓:	2 天
OpenLong:20150520			
买入:	20150520		
CloseLong:20150526			
卖出:	20150526	持仓:	4 天
OpenLong:20150528			
买入:	20150528		
CloseLong:20150603			
卖出:	20150603	持仓:	4 天
OpenLong:20150615			
买入:	20150615		
CloseLong:20150624			
卖出:	20150624	持仓:	6 天
OpenLong:20150706			
买入:	20150706		
CloseLong:20150713			
卖出:	20150713	持仓:	5 天
OpenLong:20150805			
买入:	20150805		
CloseLong:20150812			
卖出:	20150812	持仓:	5 天
OpenLong:20150824			
买入:	20150824		

表 3-1（续）

CloseLong:20150827			
卖出:	20150827	持仓:	3 天
OpenLong:20150911			
买入:	20150911		
CloseLong:20150921			
卖出:	20150921	持仓:	6 天
OpenLong:20151013			
买入:	20151013		
CloseLong:20151016			
卖出:	20151016	持仓:	3 天
OpenLong:20151102			
买入:	20151102		
CloseLong:20151104			
卖出:	20151104	持仓:	2 天
OpenLong:20151116			
买入:	20151116		
CloseLong:20151119			
卖出:	20151119	持仓:	3 天
OpenLong:20151123			
买入:	20151123		
CloseLong:20151127			
卖出:	20151127	持仓:	4 天
OpenLong:20151202			
买入:	20151202		
CloseLong:20151207			
卖出:	20151207	持仓:	3 天
OpenLong:20151209			
买入:	20151209		
CloseLong:20151211			
卖出:	20151211	持仓:	2 天
OpenLong:20151215			
买入:	20151215		
CloseLong:20151216			
卖出:	20151216	持仓:	1 天
OpenLong:20151225			
买入:	20151225		
CloseLong:20151231			
卖出:	20151231	持仓:	4 天

从以上部分的交易详细图不难看出，该模型的交易次数偏多。本研究总的回测时间为 2015-2017 年，但是仅 2015 年就已经有了近六十次操作。但是仍可

以有较可观的利润。

表 3-2 为该模型的指数结果。

表 3-2 上证 50 指数结果（2015-2017）

状态个数	策略收益率	基准收益率	策略年化收益率	基准年化收益率	最大回撤	夏普比率	胜率	买入次数
3	11.35%	7.95%	3.87%	2.71%	26.82%	7.99%	54.0%	50

其中，夏普比率为策略评价的标准化指标。其计算公式为以下。

$$\mathrm{SR} = \frac{E(R_p) - R_f}{\sigma_p} \qquad \text{(公式 2-9)}$$

其中，$E(R_p)$为投资组合预期报酬率。R_f为无风险利率。σ_p为策略标准差。

胜率，代表着一个策略运行固定的周期内，盈利的操作占所有交易操作的百分比。

图 3-9 为回测时的收益率曲线。

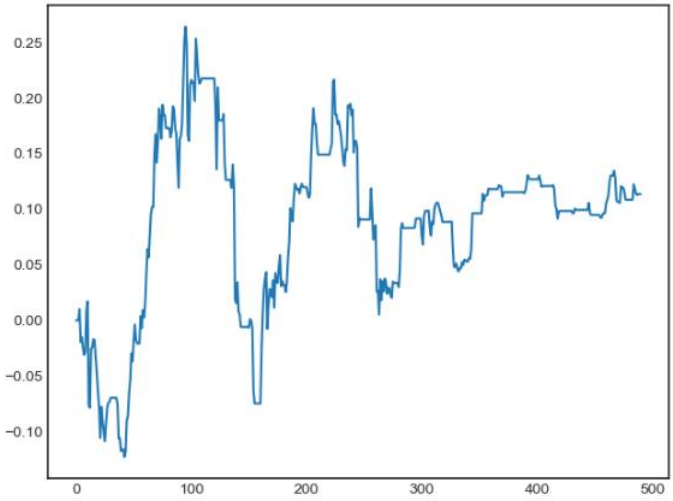

图 3-9 策略收益率曲线

同时，为了测试该模型在个股上面的表现。本文选取了深康佳（000016）作为该模型测试标的。策略选择 2005 年到 2010 年底作为训练集，2011 年到

2017 年为测试集。图 3-10 和图 3-11 为各个状态的分布和收益。

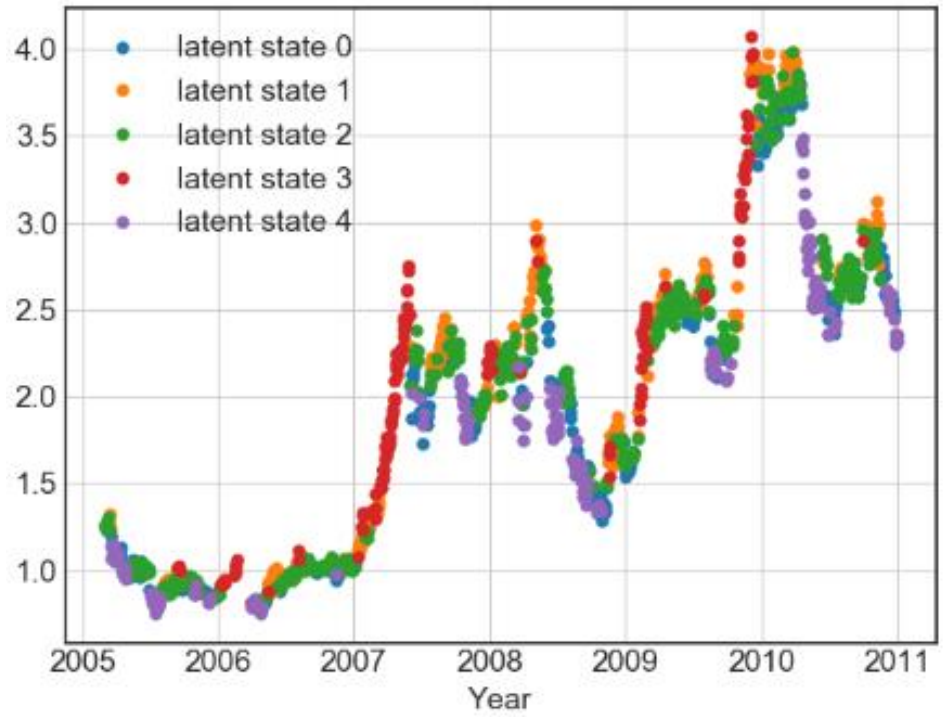

图 3-10 状态转换图

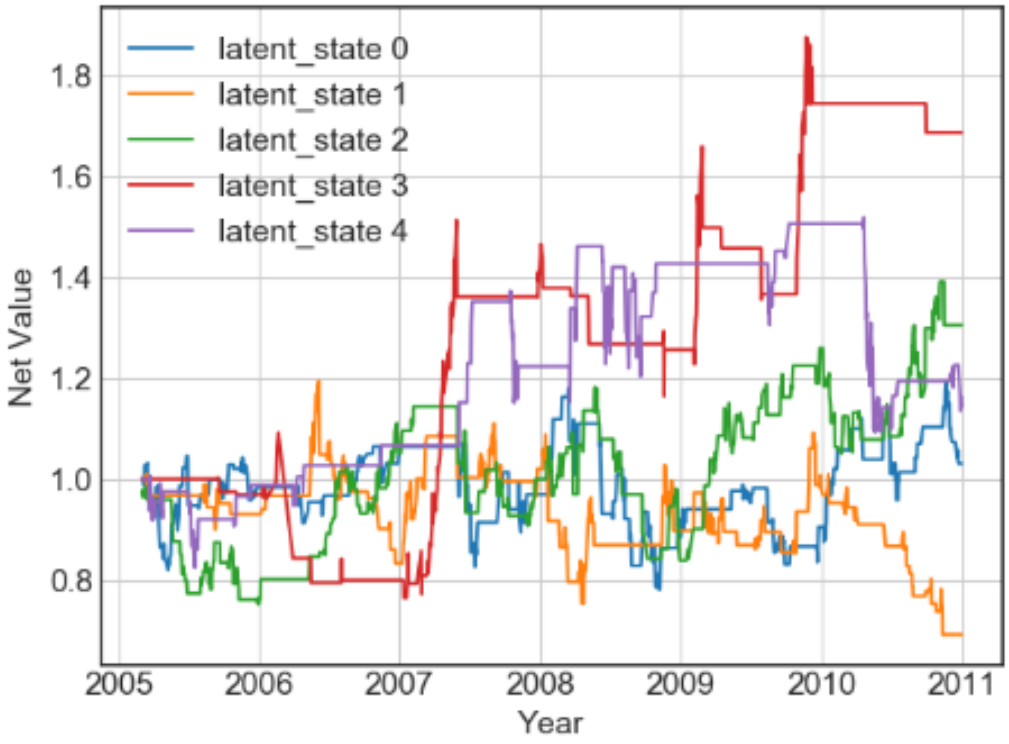

图 3-11 状态收益图

在回测区间得到了两倍于基准收益的收益。不可否认的是，这种高额的收益是有极大的随机性的。在个股的某些时候，可能会拥有极大的收益，但是这种收益同时是高风险的。在获得这个收益的策略中，其最大回撤达到了百分之五十。当模型选择其他测试区间时，其也可能会远远低于基准收益率。所以这种具有偶然性的策略是不值得推荐的。如表 3-3。

表 3-3 策略收益对比表

策略收益率	240.45%
基准收益率	98.23%

系统也可以按照每日的收盘价减去开盘价，得到每日的价差。结合价差和交易量，构建新的 HMM。下图为依照每日涨跌幅和交易量构建的分解的隐藏状态的转移概率图。

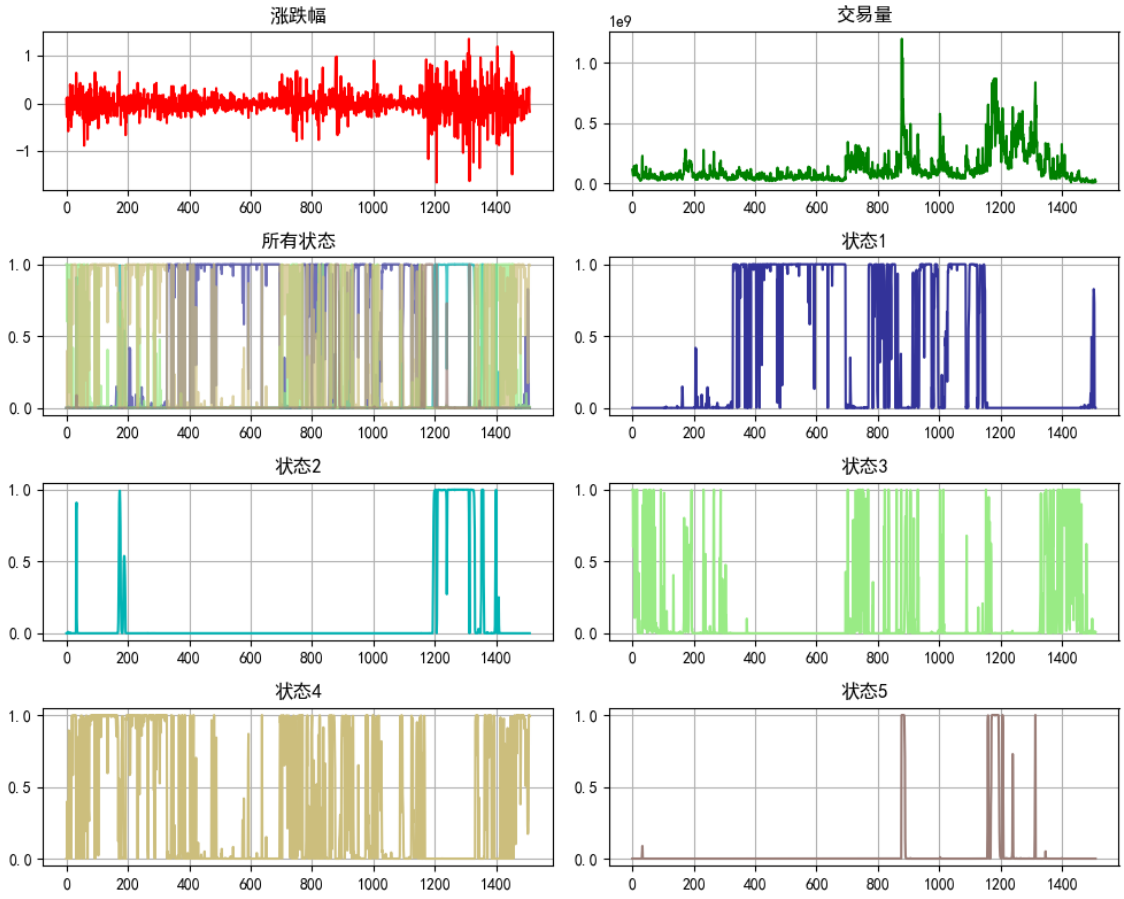

图 3-12 状态转移概率图

3.4 本章小结

本章构建了基于 HMM 的策略模型，相比于前一章的机器学习算法，HMM 能给出更有效、高收益的择时选择。本章也实现了 HMM 的现实回测，能有一个完整的系统来判断一个本模型是否有效。

与前一章的各个模型相比，HMM 可以选择选择更加合适的交易时间节点。同时，在市场暴跌的时候，HMM 也可以更加有效的避免过大幅度的回撤。在效果上面来说，HMM 也相比前面的算法更加稳定和敏感。

第四章 基于耦合隐马尔可夫模型的股票择时研究

4.1 耦合隐马尔可夫模型概述

4.1.1 耦合隐马尔可夫模型引言

根据 HMM 的理论，继续研究两种不同 HMM 在其状态演进中具有相关性的情况。这种模型被称为耦合的隐马尔可夫模型（Coupled Hidden Markov model, CHMM）。一个 HMM 状态的切换将取决于 HMM 自己的状态和另一个 HMM 的状态，其自身的状态转移也将取决于其本身和另一个 HMM。

在其他领域[21]，HMM 容易把许多有用的特征当做噪声而进行处理。这样将降低模型的有效性。因为，提出了因子隐马尔可夫模型（Factorial Hidden Markov Model, FHMM）、分层隐马尔可夫模型（Layered Hidden Markov Model, LHMM）和 CHMM。其中，对于事件检测，LHMM[22]提供了一种解决推理问题的方法。通过分层结构将运动分析解耦合为不同的时间粒度，使得该算法能够检测突然的变化，并且对低阶误差具有鲁棒性。正如稍后将详细说明的那样，需要严格的推导来描述这种耦合关系。CHMM 的主要优点是其可以模拟多个相互作用的序列，这是现实世界中非常常见的一种情况。

从各个研究院、高校中所使用的 CHMM 的公式都不完全一样，然而这些有一些相同特征。他们都选择两个 HMM，并拟合这两个的转移概率。

其中也有许多不同领域的研究者[23]将 CHMM 运用于不同的领域，并且得到了不错的结果。

4.1.2 耦合隐马尔可夫模型结构

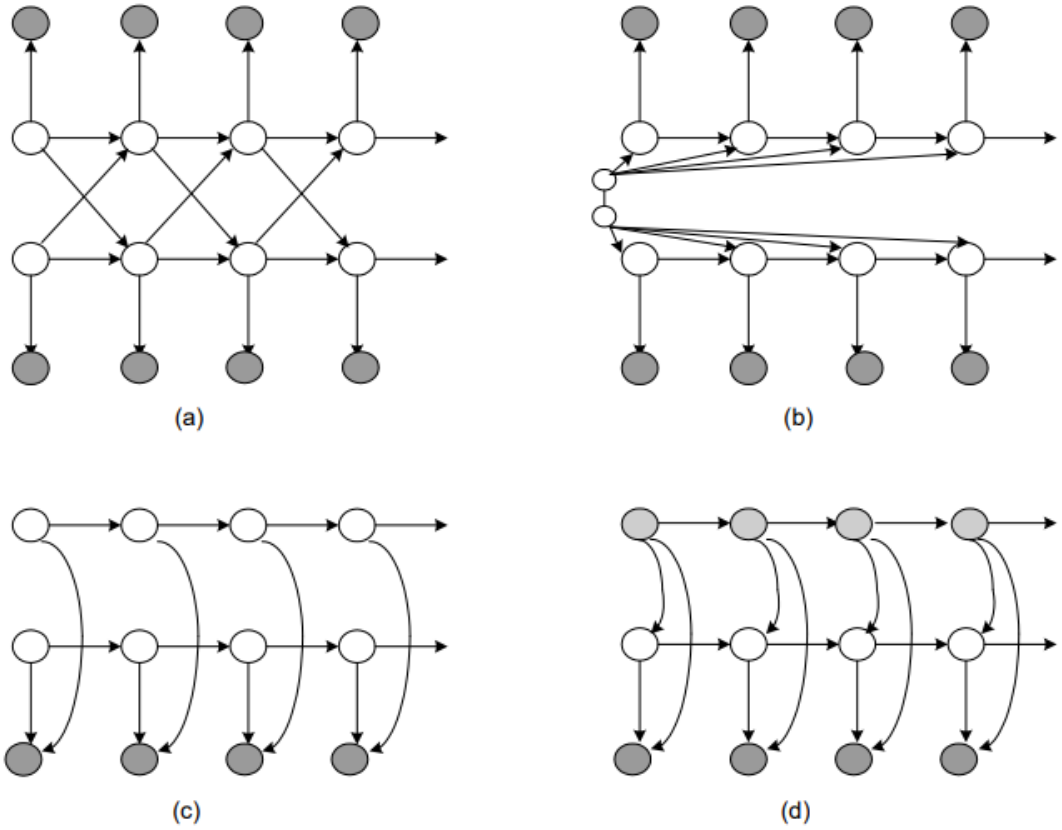

图 4-1 多种耦合隐马尔可夫模型

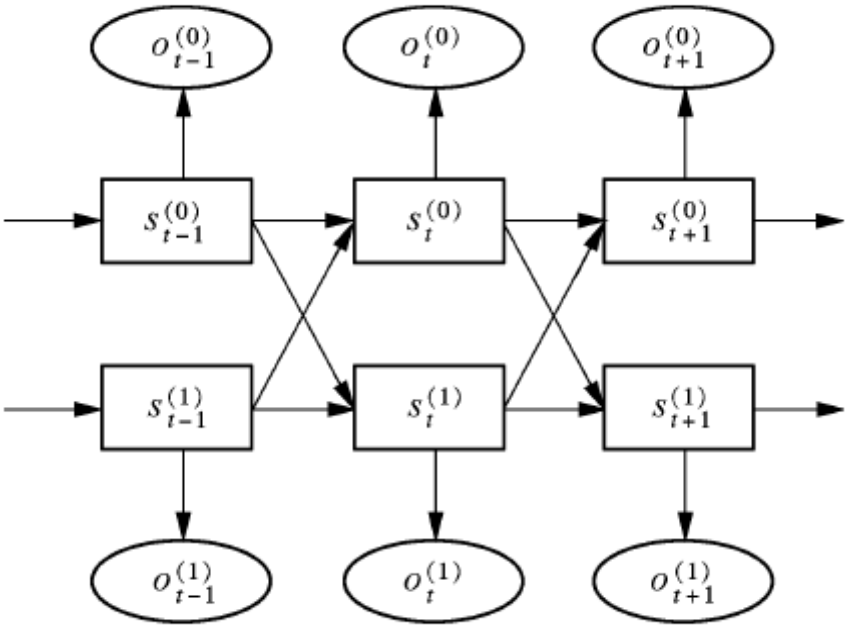

图 4-2 本文使用耦合隐马尔可夫模型

如图4-2，本研究将要使用的一个称为标准的耦合HMM。它指的是一组HMM模型，其中一个模型在时间 t 的状态取决于在时间 t-1 时所有模型（包括它本身）的状态。虽然这架构可以扩展到任意数量的 HMM 耦合在一起，但是本文只限在两个 HMM 中进行耦合。两个节点的节点数 N 相同。CHMM 中任何一个模型的状态转换概率取决于其 HMM 的当前状态和另一 HMM 前一状态。

两个 HMM 的状态的转移概率都和单个 HMM 不同。所以，转移概率不再是$P(S_t|S_{t-1})$，而是P$(S_t^c|S_{t-1}^1 S_{t-1}^2)$。在 CHMM 中，也将会有四个转移概率矩阵，因此模型将变得比单个 HMM 复杂。

一种流行的方法是使用索尔和约旦(1999)的混合模型公式。这个公式替换了$N^C * N^C$的转移矩阵，使用转移矩阵代表如下：

$$\mathrm{P}\left(X_t^{(i)}\middle|x_{t-1}^{(1:C)}\right)=\sum_{k=1}^{C}\omega_{ki}\,\mathrm{P}\left(X_t^{(i)}\middle|x_{t-1}^{(k)}\right) \qquad \text{(公式 3-1)}$$

其中，ω_{ki}可被认为是耦合权重或者是链 k 对链 i 的影响强度。以上式子就只涉及了一个参数。

随后，在 2013 年 Sherlock 用每个链的结构化转换矩阵替换完整转换矩阵，其中转换概率包括了另一个链作为协变量的逻辑回归建模。

本文中使用的是以上式子中的相对简便的 CHMM。

如果使用多个 HMM 或者对图像耦合 HMM，那么有可能出现离散的情况。

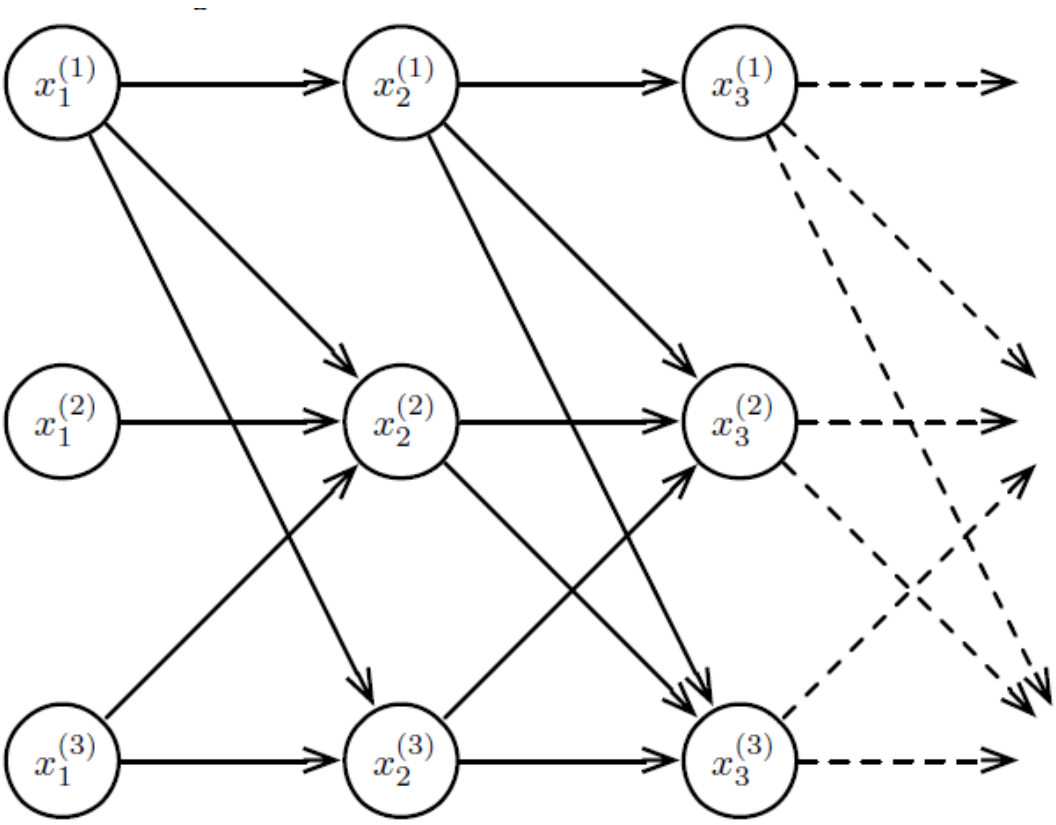

图 4-3 多耦合 HMM 图

在图像识别的应用的背景下，主要的问题是推断图形结构。这样预期将是离散的。

4.1.3 耦合隐马尔可夫模型的应用准备

本研究的目标是首先制定一个 CHMM 策略，并利用它来模拟沪深 300 和上证 50 的走势。然后设计一种交易策略，利用它们的耦合作用，准确地预测这两个指数的未来走势。根据 CHMM 的模型，观察结果可以是价格或技术指标，如 SMA，RSI 或资产的随机指标。

本研究希望通过使用 CHMM 分析两个指数获得更好的结果，这使研究能够利用 CHMM 的两个特征之间相关特征的联系。在股票交易中，普遍认为上证 50 和沪深 300 是强相关的指数。假设可以用 CHMM 来表现这种相关性，希望在交易指数时，来自市场的信息将会被更有效的分析利用。

如果选择耦合的标的关系是强联系或毫无联系，CHMM 就可以获得很大利

润。CHMM 和 HMM 的基础原理可以优秀的分析两个标的的相关性。同时，它们也可以预测一个标的的转换概率对另一个标的走势的作用。

由于其优点，人们希望开发一个以 CHMM 为中心的整个交易引擎，如图 4-4 所示。

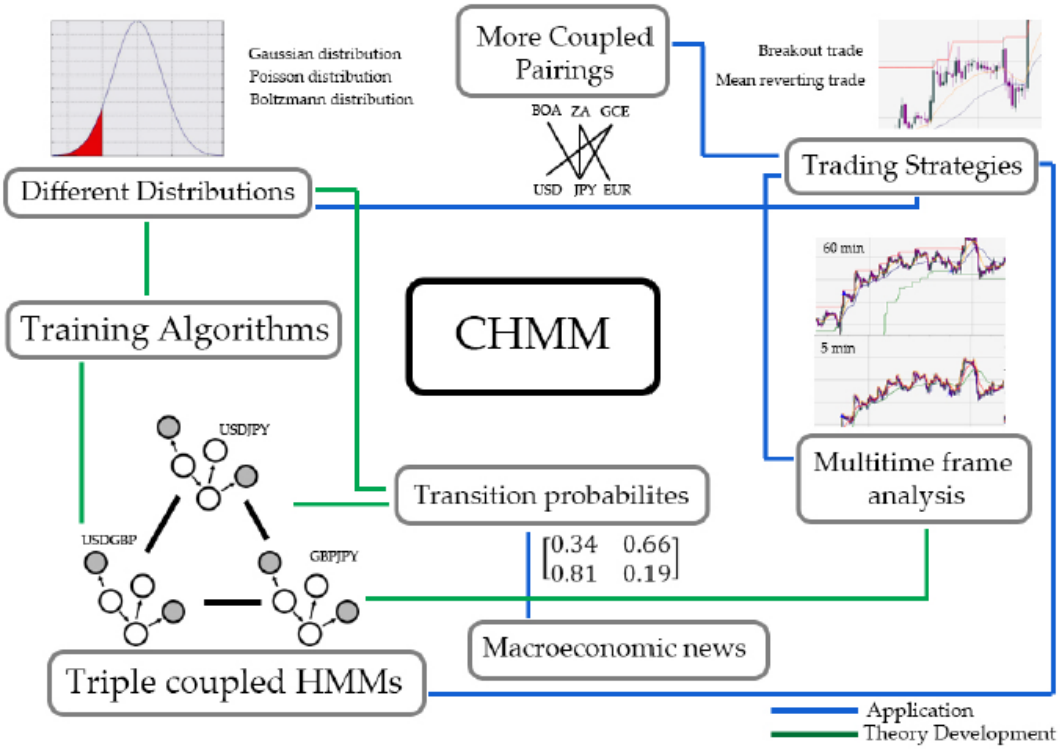

图 4-4 耦合隐马尔可夫模型量化系统流程图

上图代表着利用 CHMM 的全流程的交易引擎。从不同的特征数据，获得转移概率矩阵，随后训练耦合模型，辅以多维时间数据分析，进而获得交易策略。

4.2 基于耦合隐马尔可夫的量化模型

4.2.1 特征准备与选择

与 HMM 模型一样，CHMM 需要设计一种方法来从观察中了解 CHMM 的参数。虽然可以将用于 HMM 的 EM 等算法移植到 CHMM 上，但这样做会使在计

算上难以处理。在目前已开发的大部分算法中，都无法顺利完成这种计算。因此，这需要将回到经典的优化技术。

为了能够检索最优路径，本文需要跟踪每个 t,(i,j)配对和 C 的最大参数。文中创建数组$\alpha_t^{(c)}k$来实现这一点。为 CHMM 找到最优路径的完整过程如下：

（1） 初始化

$$\delta_1^c(i) = \pi_i^c b_i^c(o_i^c), 1 < i < N \quad \text{(公式 3-2)}$$

$$\varphi_1^c = 0 \quad \text{(公式 3-3)}$$

（2） 递归

$$\delta_1^c(k) = \delta_t(i)^c a_{ik}^{(1,c)}, 2 < t < T, 1 < k < N \quad \text{(公式 3-4)}$$

（3） 得出结果

$$P_i^c = \max_{1<i<N}[\delta_T^c(i)] \quad \text{(公式 3-5)}$$

（4） 最优路径回溯

$$q_t^c = \varphi_{t+1}^c(\mathrm{q}_{t+1}^c), \mathrm{t} = \mathrm{T}-1, \mathrm{T}-2, \ldots\ldots, 1 \quad \text{(公式 3-6)}$$

用于训练 HMM 的 Baum-Welch 方法在其中包含定理，它可以导出 CHMM 中的参数的重估公式。这给出了关于 HMM 的模型参数的重新确认过的值。就像学习 HMM 一样，对于每次迭代，需要使用当前 HMM 的参数来计算所有新参数，将当前参数与新的参数交换。

同时，本研究选择了沪深 300 和上证 50 两种指数作为训练标的。首先，由于中国境内上海交易所和深圳交易所两种指数有相当大的相关性。图 4-5 为上证指数和深证成指从 2010 年 1 月 1 日到 2017 年 12 月 31 日的走势图。

图 4-5 上证指数和深证成指走势图

不难发现，两个指数是呈高度相关的状态。因此，上证 50 和沪深 300 将会有很强的相关性。

图 4-6 上证 50 和沪深 300 指数走势图

从以上两个图对比可以发现，在某些大盘指数急跌或者微涨的时候，可能对应着上证 50 和沪深 300 的微跌和急涨。这是由于沪深 300 和上证 50 大部分由蓝筹股、大盘股组成，当许多其他股票杀跌的时候，蓝筹股等会相对抗跌，甚至会成为回避风险、资金多元配置的选择。因此，选择上证 50 和沪深 300 相比较于大盘指数建模也是更加有效的选择。

同时，本研究也必须谨慎地选择适当的时间框架，达到良好的可继承性。在选择一个时间间隔时，本研究假设市场状态服从马尔可夫属性，即转换一次时间，

状态就会发生一次变化。本研究假设预测下一个状态所需的所有信息都是在该区间内捕获的。这代表着太长的间隔，模型可能会失去一些信息，而且这些信息对于确定下一个状态是有意义的。

选择合适的时间框架本身就是一项研究课题。许多交易爱好者甚至认为，一个优秀的策略应该包括多个时间框架。在本研究中，将使用从 Tushare 开源接口获取的日频数据。如果需要将 CHMM 应用到外汇等市场，为了能够达到高频交易，普遍是以十分钟为单位。

本研究选择了每日股价涨跌和交易量作为观测量。在 CHMM 中，如果单独选择每日收盘价为观测量，两个不同量级的指数是无法有效的对比的。在外汇黄金等领域，也有人选用 RSI 指标等作为观测特征。因为如外汇和黄金的价格相差非常之大，许多时候外币与美元相比的量级为 1 以内。然而，黄金价格却为上千元。因此，若直接对两个量级的观测标的直接进行耦合，效果会变得非常差。但是在经济学原理内，量价关系是十分重要的内在关系。正所谓，军马未动粮草先行。股票价格是由每天的大量的交易而决定的。对于在中国最火热的房地产市场，一个房地产的完整景气周期包括了四个阶段。包括衰退、萧条、复苏和繁荣。衰退时为量缩价平。萧条时为量缩价跌。复苏时为量平价稳。繁荣时为量价齐升。在股市中也有同样的量价关系。下图是多种量价关系所对应的情况。

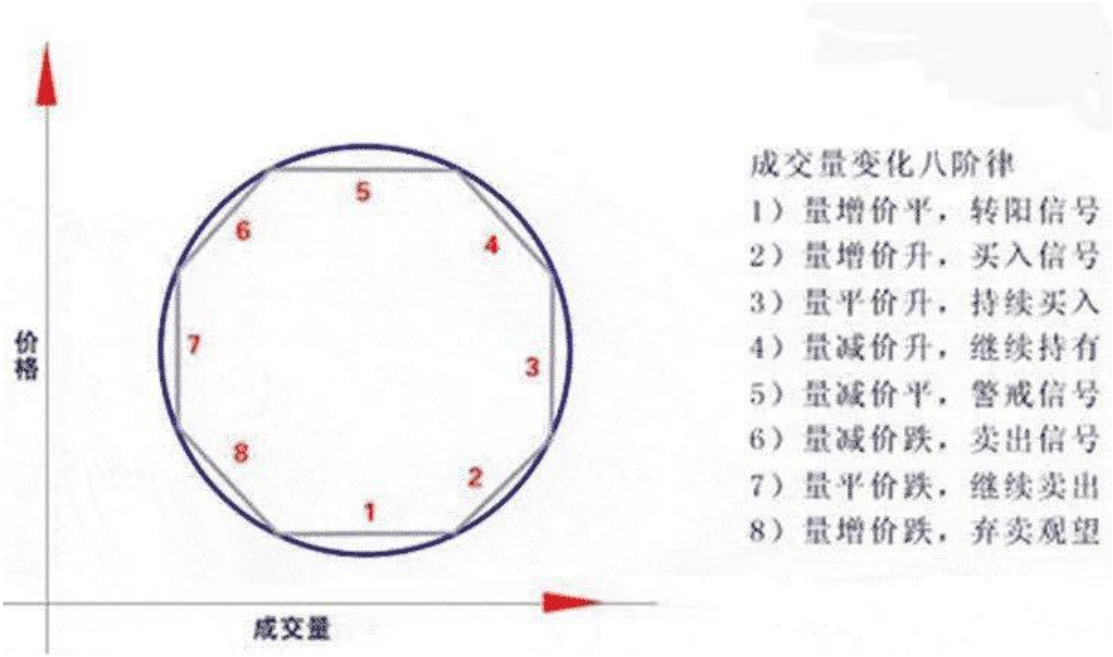

图 4-7 量价关系对应图

从上图不难得出，成交量与价格之间有相当紧密的联系。随后，本模型得出了这两个指数各自的转移概率矩阵。这五个隐藏状态可能代表指数市场的急跌、微跌、震荡、微涨和暴涨。

给定一个观测序列，模型希望可以选出潜在的隐藏路径或最优路径，从而输出该路径。模型内将最优路径 Q 定义为使$\mathrm{P}(\mathrm{Q}|\mathrm{O},\varphi)$最大化的路径，它等价于最大化$\mathrm{P}(\mathrm{Q}|\mathrm{O},\varphi)$。对于 HMM，有一个维特比迭代算法来寻找这样的 Q。本研究中扩展了它在 HMM 的实现来解决 CHMM。在此需要考虑两个观测序列的两个最优路径，而不是仅考虑一条最佳路径。因此，本研究希望找到一个最优路径$Q^{(1)} = \{q_1^{(1)}, q_2^{(1)}, q_3^{(1)}, q_4^{(1)}, q_5^{(1)}, \ldots . q_n^{(1)}\}$和$Q^{(2)} = \{q_1^{(2)}, q_2^{(2)}, q_3^{(2)}, q_4^{(2)}, q_5^{(2)}, \ldots . q_n^{(2)}\}$。

因此，本研究定义下式。

$$\delta_t^{(c)} = \max P[q_1^{(c)}, q_2^{(c)}, q_3^{(c)}, q_4^{(c)}, q_5^{(c)}, \ldots . q_n^{(c)} = s_i^{(c)}, o_1^{(c)}, o_2^{(c)}, , \ldots . o_t^{(c)}|\varphi] \text{(公式 3-7)}$$

这意味着在时间 t 上，HMM（1）输出第一个 t 观测值，并以状态 s(C)i 结束的最大概率。本研究中，创建了数组$\tau_t^{(c)}(k)$来实现这一点。

在学习 CHMM 的参数时，模型会重复更新参数，每次迭代三次。在训练过程中，一个 HMM 如何影响其他 HMM 参数的回归，可能尚未彻底不清楚。因为所有的模型中调用的公式都是重新设计的。

在 CHMM 中，为了可以把回测结果与上一章中的 HMM 进行比较，本策略也选择了五状态模型。

因此，本研究针对两个指数进行 CHMM 的训练。首先可以得到两个模型单独的转移概率。随后根据转移概率来进行耦合。图 4-8 和图 4-9 为一段时间段内的两个指数的转移概率图。

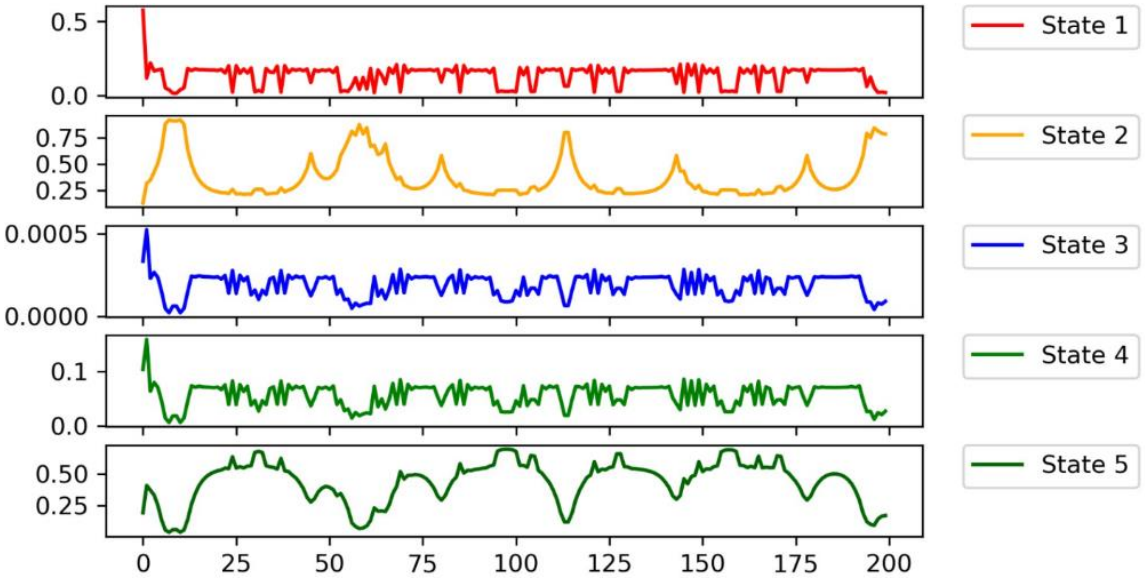

图 4-8 上证 50 的转移概率图

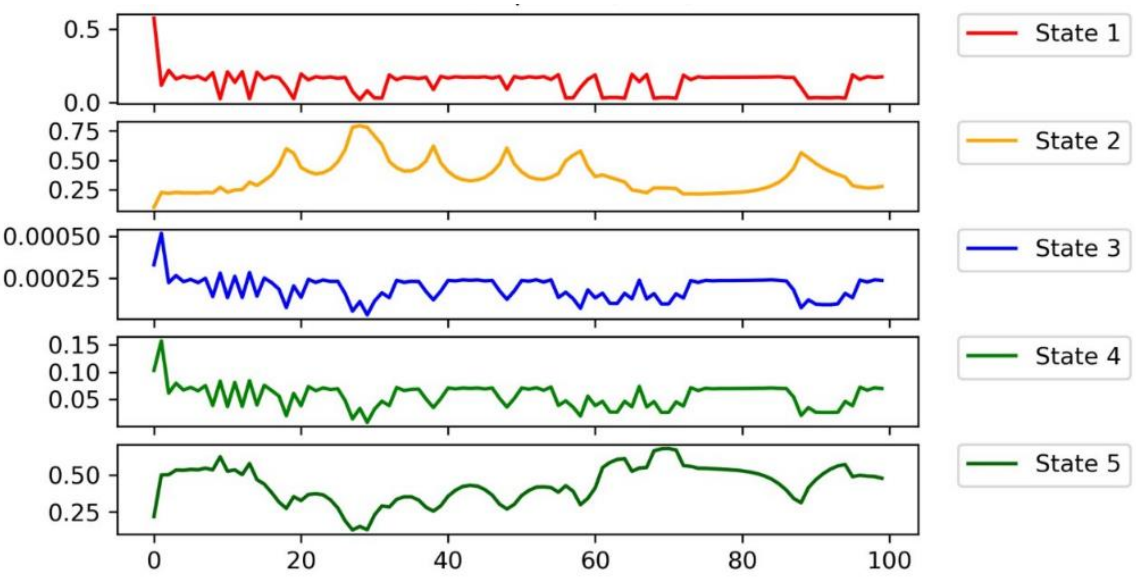

图 4-9 沪深 300 的转移概率图

从以上转移概率图不难发现，两个指数的由于整体涨跌趋势相同，转移概率也有所相近。

4.2.2 状态选择

首先，策略可以得到其中隐藏状态的相关参数。

当拥有了足够多的特征状态的训练参数后，本模型首先将计算两个训练模型的状态相似度。这个相似度也将代表耦合程度。

在上证 50 和沪深 300 指数的耦合模型中，耦合度为 65.66%。

```
['similarity: ', 0.655266757865937]
```

图 4-10 耦合数值图

在本模型中，若状态耦合程度越高，其回测结果将会更好。代表着两个模型的高度耦合。下图为状态耦合矩阵。

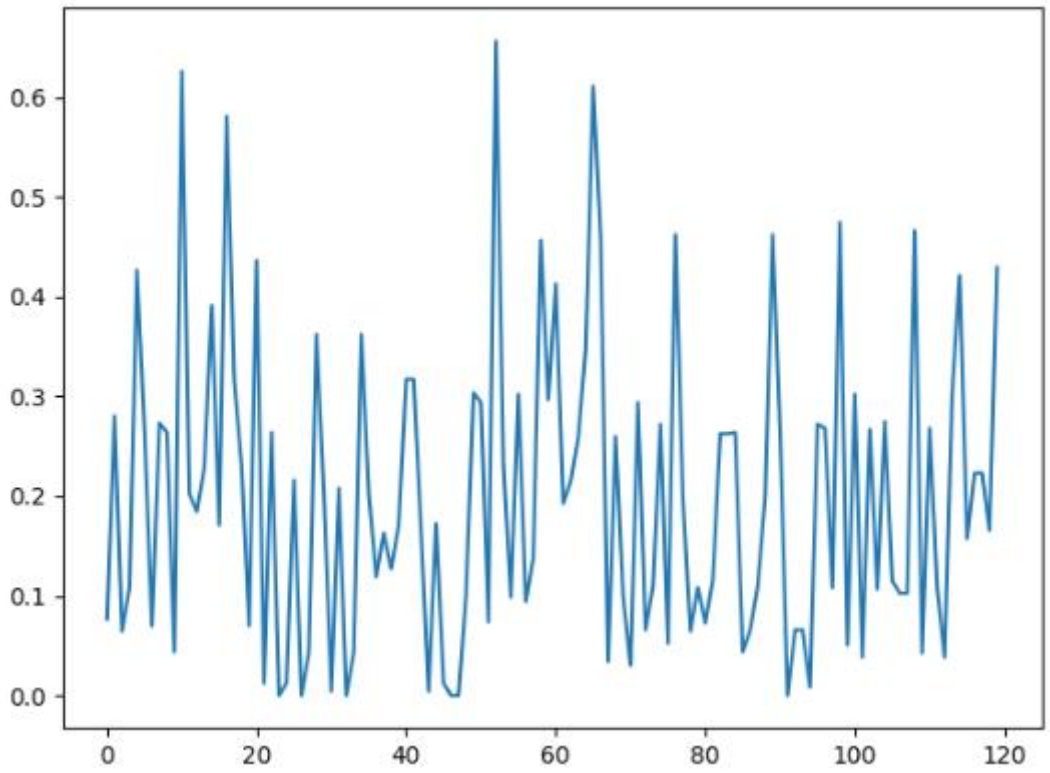

图 4-11 状态耦合图

上图代表，当其中耦合系数为 55 时，状态耦合程度最高。其中 120 代表着五个状态的全分布。若当耦合系数为 1 时，耦合程度最高。代表着链一的{1,2,3,4,5}状态对应着链二的{1,2,3,5,4}时相似度最高,每个状态一一对应。

随后，本研究也快速分析了使用维特比和非维特比分析时下一个可能状态的差异程度。由于两者的表现有所不同，但仅略有差异，本研究预计两种方法预测的可能状态也会略有不同。能够了解维特比和非维特比方法是否可以实际运用可

以帮助人们放弃计算复杂度高的计算方法。

4.2.3 回测模型

在 CHMM 中，回测模型基本按照上一章的回测模型框架构建。本系统也选择使用本地回测。由于耦合了两个观测标的，本模型可以同时回测两个标的的结果。笔者最初希望通过 MATLAB 来进行回测，但是发现许多欧美的投行和对冲基金普遍在用 Python 的技术栈。这代表着使用 Python 回测即不需要跟任何一家私有的公司或者平台捆绑。而且，许多量化平台也都搭建了开源的平台帮助其他人的策略进行回测。

4.3 实验结果与分析

首先，回测系统会把在测试期内的买卖操作单独列出，供用户检查。

表 4-1 CHMM 策略详细交易记录表（2015-2017）

操作指示	操作时期	持仓时间	
OpenLong:20150106			
买入:	20150106		
CloseLong:20150107			
卖出:	20150107	持仓:	1 天
OpenLong:20150206			
买入:	20150206		
CloseLong:20150225			
卖出:	20150225	持仓:	8 天
OpenLong:20150909			
买入:	20150909		
CloseLong:20150925			
卖出:	20150925	持仓:	12 天
OpenLong:20150930			
买入:	20150930		
CloseLong:20151009			
卖出:	20151009	持仓:	2 天
OpenLong:20151027			
买入:	20151027		
CloseLong:20151103			
卖出:	20151103	持仓:	5 天

表 4-1（续）

OpenLong:20151118			
买入:	20151118		
CloseLong:20151126			
卖出:	20151126	持仓:	6 天
OpenLong:20151203			
买入:	20151203		
CloseLong:20151216			
卖出:	20151216	持仓:	9 天
OpenLong:20151223			
买入:	20151223		
CloseLong:20160106			
卖出:	20160106	持仓:	9 天
OpenLong:20160111			
买入:	20160111		
CloseLong:20160129			
卖出:	20160129	持仓:	14 天
OpenLong:20160215			
买入:	20160215		
CloseLong:20160301			
卖出:	20160301	持仓:	11 天
OpenLong:20160304			
买入:	20160304		
CloseLong:20160309			
卖出:	20160309	持仓:	3 天
OpenLong:20160311			
买入:	20160311		
CloseLong:20160317			
卖出:	20160317	持仓:	4 天
OpenLong:20160321			
买入:	20160321		
CloseLong:20160421			
卖出:	20160421	持仓:	22 天
OpenLong:20160422			
买入:	20160422		
CloseLong:20160429			
卖出:	20160429	持仓:	5 天
OpenLong:20160518			
买入:	20160518		
CloseLong:20160527			
卖出:	20160527	持仓:	7 天
OpenLong:20160530			
买入:	20160530		

表 4-1（续）

CloseLong:20160601			
卖出:	20160601	持仓:	2 天
OpenLong:20160701			
买入:	20160701		
CloseLong:20160718			
卖出:	20160718	持仓:	11 天
OpenLong:20160726			
买入:	20160726		
CloseLong:20160728			
卖出:	20160728	持仓:	2 天
OpenLong:20160811			
买入:	20160811		
CloseLong:20160812			
卖出:	20160812	持仓:	1 天
OpenLong:20160816			
买入:	20160816		
CloseLong:20160822			
卖出:	20160822	持仓:	4 天
OpenLong:20160914			
买入:	20160914		
CloseLong:20160930			
卖出:	20160930	持仓:	10 天
OpenLong:20161020			
买入:	20161020		
CloseLong:20161025			
卖出:	20161025	持仓:	3 天
OpenLong:20161102			
买入:	20161102		
CloseLong:20161110			
卖出:	20161110	持仓:	6 天
OpenLong:20161114			
买入:	20161114		
CloseLong:20161125			
卖出:	20161125	持仓:	9 天
OpenLong:20161205			
买入:	20161205		
CloseLong:20161215			
卖出:	20161215	持仓:	8 天
OpenLong:20170113			
买入:	20170113		
CloseLong:20170116			
卖出:	20170116	持仓:	1 天

表 4-1（续）

OpenLong:20170123			
买入:	20170123		
CloseLong:20170207			
卖出:	20170207	持仓:	6 天
OpenLong:20170209			
买入:	20170209		
CloseLong:20170222			
卖出:	20170222	持仓:	9 天
OpenLong:20170323			
买入:	20170323		
CloseLong:20170508			
卖出:	20170508	持仓:	29 天
OpenLong:20170524			
买入:	20170524		
CloseLong:20170601			
卖出:	20170601	持仓:	4 天
OpenLong:20170621			
买入:	20170621		
CloseLong:20170626			
卖出:	20170626	持仓:	3 天
OpenLong:20170706			
买入:	20170706		
CloseLong:20170921			
卖出:	20170921	持仓:	55 天
OpenLong:20170929			
买入:	20170929		
CloseLong:20171016			
卖出:	20171016	持仓:	6 天
OpenLong:20171025			
买入:	20171025		
CloseLong:20171206			
卖出:	20171206	持仓:	30 天

由于买入以及卖出的次数达到六十六次，次数虽多但是仍然相比 HMM 有所减少。在本实验模型中，仍然模型交易费用为零。当卖出股票时，会显示持仓天数。从回测结果看，此模型的交易次数并不会太过高频，因此就算存在交易费用，此模型的收益将不会受到太大影响。

此模型为了与上一章的 HMM 进行对比，仍然选择 2015 年 1 月 1 日开始至 2017 年 12 月 31 日的上证 50 指数作为该模型的回测区间。同样，这个区间的

波动相较于其他的区间较大。该区间也包含了多种股市状态。以下为 CHMM 的回测结果。

图 4-12 为 CHMM 的收益率走势图。表 4-2 为回测结果。

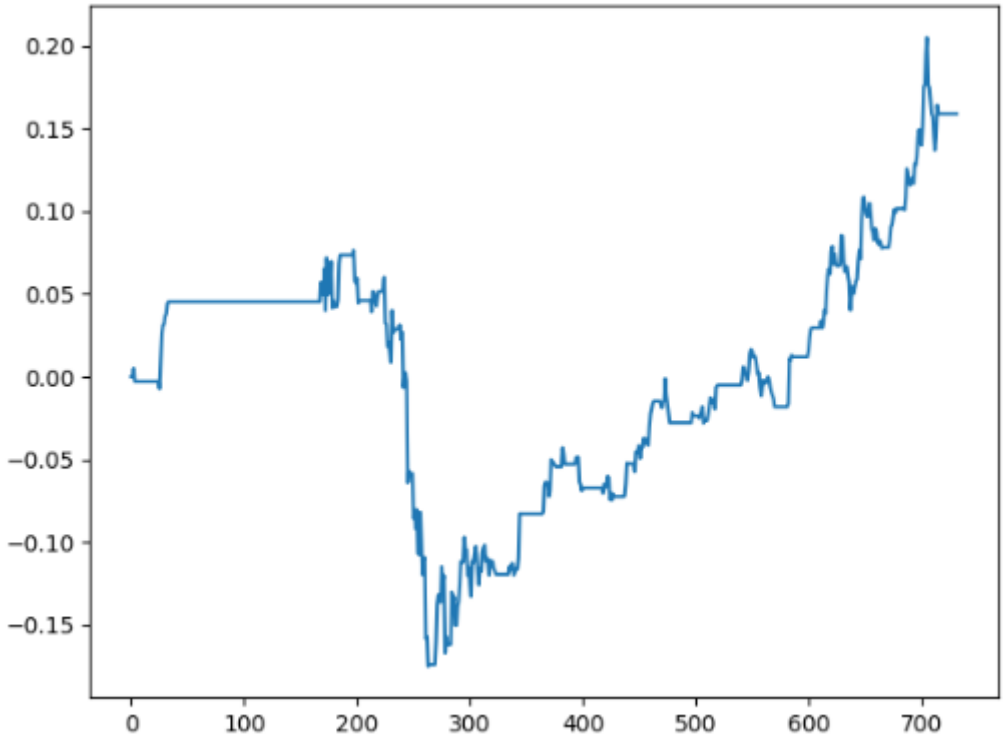

图 4-12 收益率走势图

表 4-2 上证 50 指数回测结果（2015-2017）

模型	策略收益率	基准收益率	策略年化收益率	基准年化收益率	最大回撤	夏普比率	胜率	买入次数
CHMM	16.22%	7.95%	5.54%	2.71%	23.36%	11.70%	66.66%	33

从以上走势图和表格不难看出，CHMM 能获得可观的收益。同时，CHMM 可以有效的减小模型的风险。在选取其他一些时间段进行回测时，CHMM 可能会达到更大的收益。诚然，其中有随机的因素，但是 CHMM 获得的巨额收益也是和策略的有效性相关的。通过模型训练，CHMM 可以更综合、全面的捕捉买

入和卖出时机。当两个观测标的都提示为买入信号时，CHMM 将会更加坚定的做多。

4.4 本章小结

根据以上的理论以及实验，CHMM 相比 HMM 模型的优越性已经得到证实。

目前可以得出结论为，CHMM 提供更好的指标来产生合适的交易信号。如果需要进一步证实这一理论。需要研究各种交易系统中常见的交易问题，即随机性、MACD 散度、MA 交叉。采用 CHMM 和适当的跨市场资产，令其观测值是系统关联的指标。然后比较性能数据。通过这样做，基于 CHMM 的有益特征将被改善。

本章的结果使人们相信在制定低风险套利模型的交易策略时，CHMM 的价值所在。CHMM 的马尔可夫特征以及将不同市场链接起来的能力，为人们提供了一个更加完美的交易信号。

第五章 多策略对比与总结

5.1 基于隐马尔可夫模型量化择时策略与其他对比

在本研究中对比的其他策略，也都基于机器学习。机器学习的主要逻辑是从训练集中训练出能尽量逼近训练器的函数。图 5-1 为机器学习总体模型。

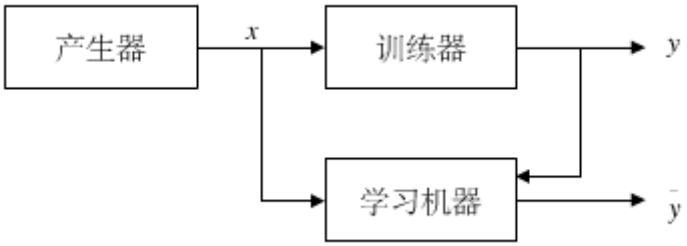

图 5-1 机器学习模型

本文中，笔者为了对比 HMM 与其他机器学习分类算法的结果，笔者复现了 KNN、支持向量机和线性回归等算法的股票预测模型并且得出结果。

首先，笔者复现了经典的朴素贝叶斯算法对股票涨跌的预测[24]。可能由于本文中没有对先验概率等进行详细研究，所以最后得出的结果并不理想。从表 5-1 可以得知，朴素贝叶斯的预测准确率只有百分之四十。

表 5-1 朴素贝叶斯准确率图

	Actual down	**Actual up**
Predicted down	5	10
Predicted up	32	23

然后，本研究使用 KNN 算法对指数进行预测[25]。

图 5-2 为 KNN 的分类逻辑图。

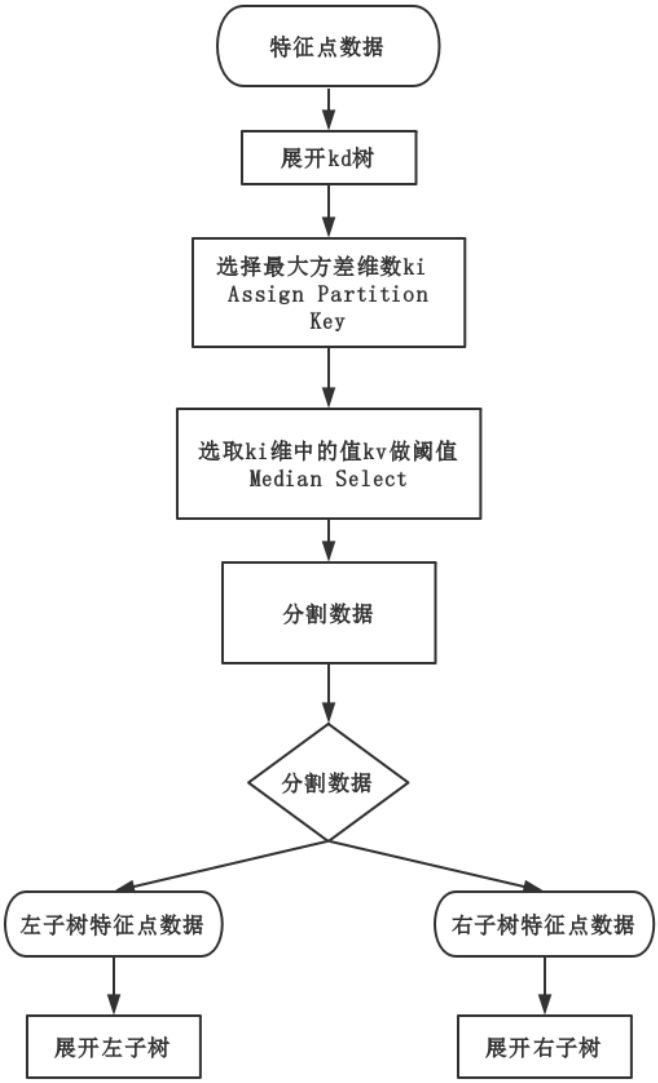

图 5-2 KNN 分类逻辑图

随后，笔者对该方法进行预测。以下的输入的数据为以时间序列的每日相对前一日的涨跌情况。表格中为预测准确度和。

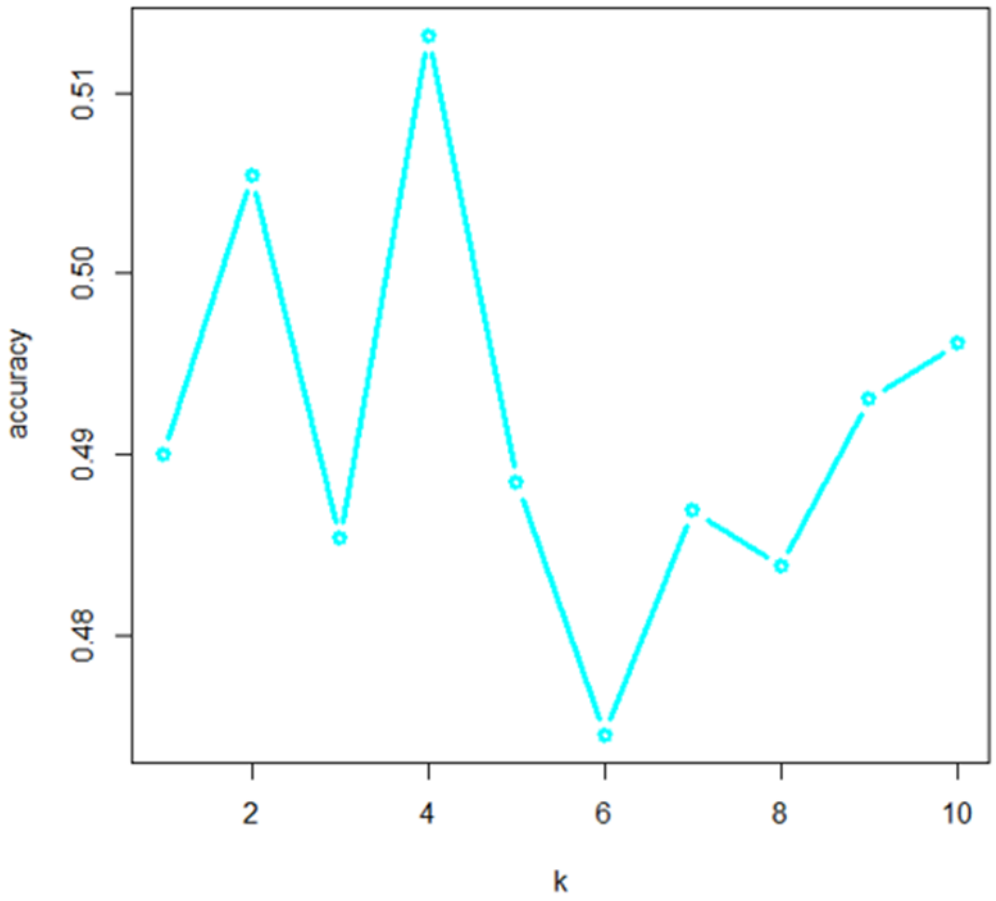

图 5-3 KNN 算法准确率图

通过上图可以得知，当 k=5 时，准确率最高。此时准确率达到了 52.5%。然而，对于一个预测算法而言，准确率仅仅超过一半是基本没有任何作用的。如将此模型运用到真实市场中，大概率将会是亏损。

随后，本研究针对 SVM 算法进行建模。在 SVM 方面，有研究者[26]将其运用于股票量化策略方面，得到了不错的结果。也有研究者[27]将其与 HMM 结合，得到了更为优秀的模型。支持向量机（SVM）作为一种强有力的时间序列预测工具，在许多方面具有很好的性能。有科学家[28]提出了一种新的混合模型，改善了 SVM 中存在的某些问题。

以下图 5-4 为 SVM 的算法流程图。

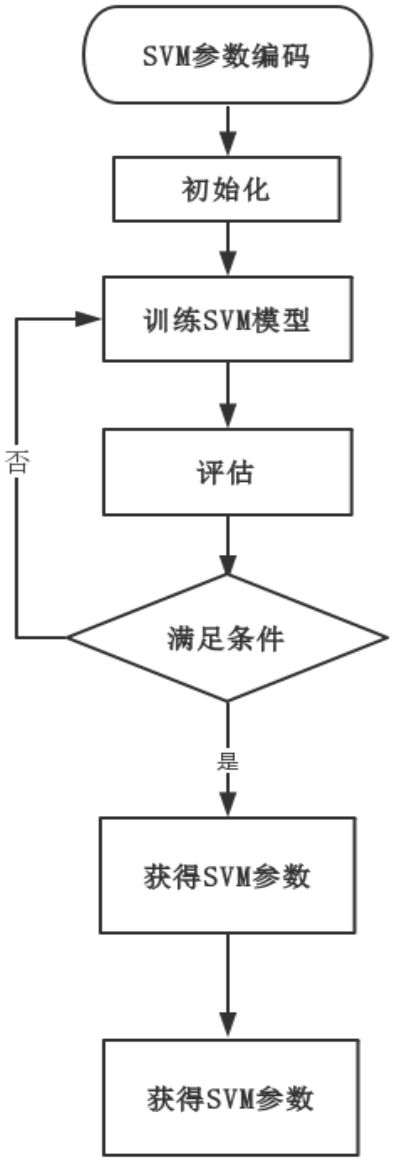

图 5-4 SVM 算法流程图

SVM 相比 KNN 算法，能得到更高的准确度。同时，为了提升本模型的准确度，此处的数据输入为粒度更细的从雅虎财经接口获取的美股亚马逊股票。该模型的数据为每日开盘价、每日收盘价、三十分钟数据等。如图 5-5。

	Open	High	Low	Close	Volume
Date					
2018-03-21	1586.45	1590.00	1563.17	1581.86	4667291.0
2018-03-22	1565.47	1573.85	1542.40	1544.10	6177737.0
2018-03-23	1539.01	1549.02	1495.36	1495.56	7843966.0
2018-03-26	1530.00	1556.99	1499.25	1555.86	5547618.0
2018-03-27	1572.40	1575.96	1482.32	1497.05	6793279.0

图 5-5 SVM 数据图

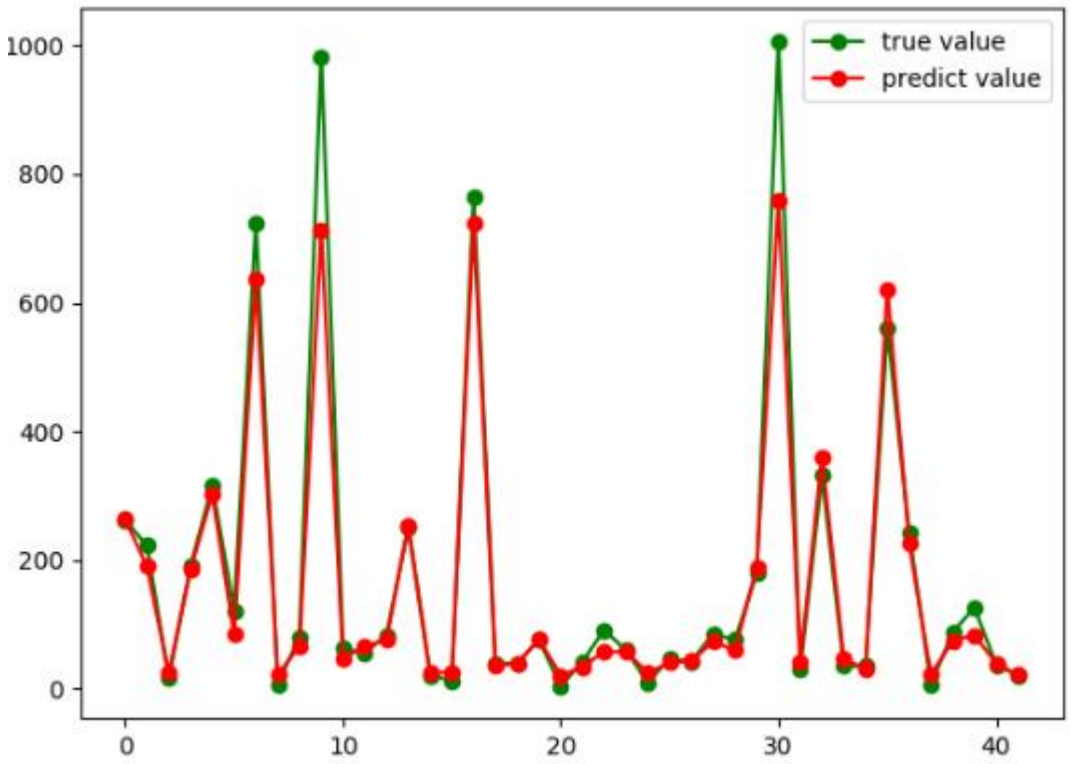

图 5-6 SVM 准确率图

以上预测数据是随机从测试集中抽取并且预测的。绿线代表真实的价格，红色代表预测价格。从上图可以得出，当股票价格较低时，预测准确度很高。然而，当股价在高位的时候，预测价格与现实股价会有比较大幅度的偏差。

随后，本文也完成了线性回归的预测[29]。此算法对于股票价格的预测准确率十分高。在该模型中，数据输入与 SVM 相同。

图 5-7 为线性回归策略的流程图。

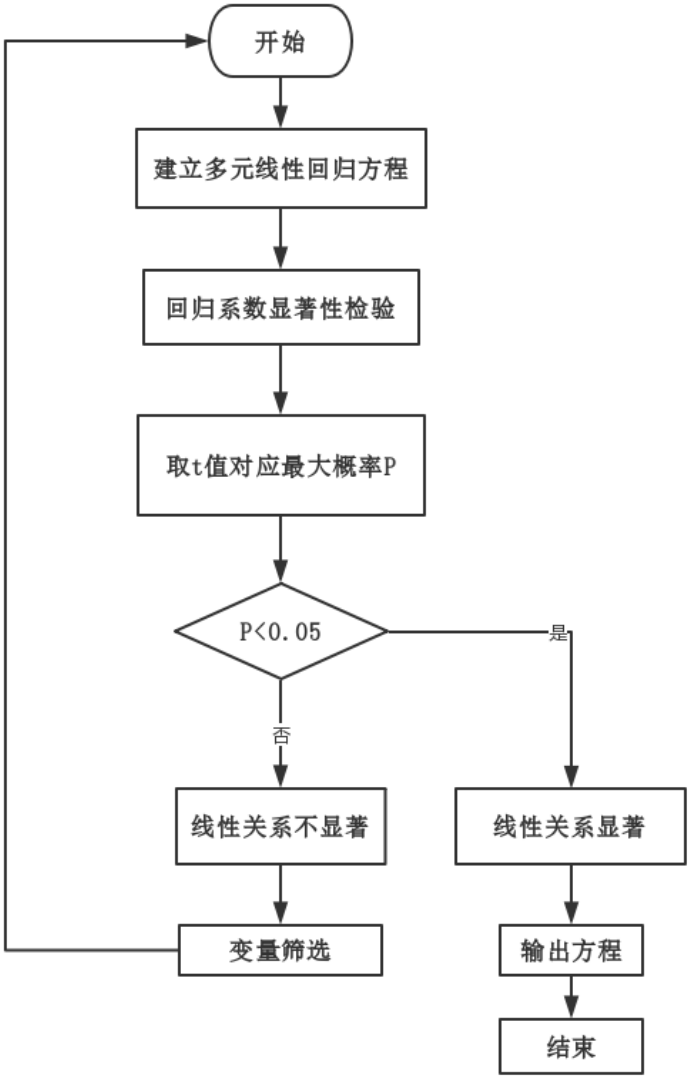

图 5-7 线性回归策略图

图 5-8 为线性回归的预测股价和回测区间真实股价的对比图。

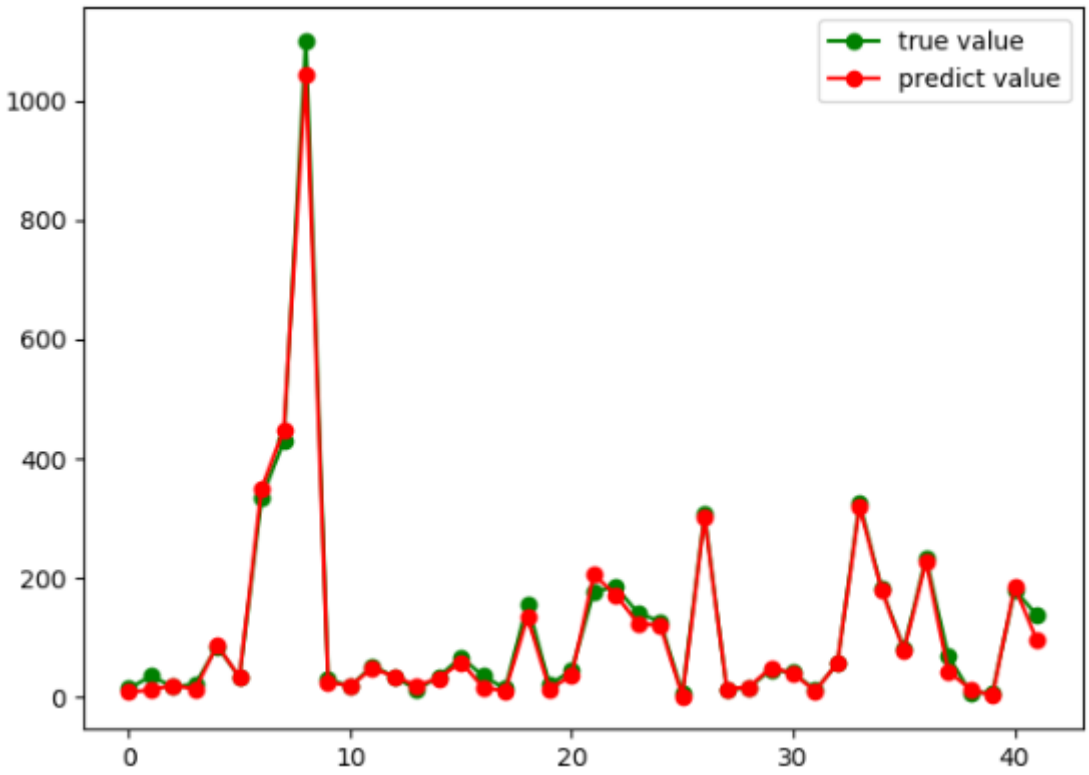

图 5-8 三十天预测图

与此同时，本模型可以给出将来三十天的股价预测供参考。如图 5-9。

	Open	High	Low	Close	Volume
Date					
2018-03-21	1586.45	1590.00	1563.17	1581.86	4667291.0
2018-03-22	1565.47	1573.85	1542.40	1544.10	6177737.0
2018-03-23	1539.01	1549.02	1495.36	1495.56	7843966.0
2018-03-26	1530.00	1556.99	1499.25	1555.86	5547618.0
2018-03-27	1572.40	1575.96	1482.32	1497.05	6793279.0

图 5-9 三十天预测股价图

这个模型的准确度为 99%。可见线性回归对于股票的预测是相对有效的。然而，上述中的准确度 accuracy 并不是意味着 99%的预测完全准确。它表示的是线性模型能够描述统计数据的信息的一个统计概念。

尽管这些机器学习算法都能作出大致有效的预测。但是，与 HMM 对比，这些机器学习算法都无法达到如此好的预测效果。

5.2 基于隐马尔可夫模型量化择时策略的优缺点

图 5-10 SVM 一个月内收益图

图 5-11 线性回归策略一个月内收益图

从对比结果综合来看，HMM 拥有参数少、鲁棒性强和可解释性强的优势。实验结果也表明了相较于其他常见策略，HMM 在规避风险的敏感性和在控制回撤的有效性上表现更佳。从期望收益率的角度来说，基于 HMM 的择时策略的表

现优于线性回归和 SVM。这主要源于 HMM 合理的马尔可夫性质假设。线性回归策略由于本身的延迟性，在控制回撤风险上表现糟糕。另外，HMM 策略带来了长期高于 50%的交易胜率，以及更低的交易频率。从平均意义上来说，它产生更少的交易费用。

尽管 HMM 在量化择时上表现出较为优越的表现，但是仍然存在一些不足，在后续的研究中将进一步改进。其当前模型的状态选择依然属于静态规则，不能有效监控各个状态的动态变化。同时，整个金融市场是个整体。HMM 仅仅能对单个指数或者个股进行建模，却没有考虑到其他指数或者个股对该标的的影响程度。通过多个指数的同时训练，必然将得到更好的收益。

5.3 基于耦合隐马尔可夫模型量化择时策略的对比

在许多其他领域，CHMM 都可以很大幅度的改善 HMM 的结果。在动作识别方面，有人[30]提出了一种新的多传感器融合框架的手语识别（SLR）使用耦合 Hidden Markov 模型（CHMM）。CHMM 在状态空间中提供交互，而不是在经典 HMM 中使用的观察状态。HMM 不能模拟模式间依赖关系之间的相关程度。该框架已被用来识别动态的孤立的手势以帮助听力受损的人。使用现有的数据耦合技术对数据集进行了测试。用 CHMM 获得的识别准确率高达 90.80%。在那个领域的基于 CHMM 的方法改进后的识别性能，远远优于 HMM 的数据融合技术。

在本文中，CHMM 为了与前一章的 HMM 进行对比，仍然选择 2015 年 1 月 1 日开始至 2017 年 12 月 31 日的上证 50 指数作为该模型的回测区间。同样，这个区间的波动相较于其他的区间较大。该区间也包含了多种股市状态。表 5-1 为 HMM 和 CHMM 的回测结果对比。

表 5-1 CHMM 和 HMM 对比表

模型	策略收益率	基准收益率	策略年化收益率	基准年化收益率	最大回撤	夏普比率	胜率	买入次数
HMM	11.35%	7.95%	3.87%	2.71%	26.82%	7.99%	54.0%	50
CHMM	16.22%	7.95%	5.54%	2.71%	23.36%	11.70%	66.66%	33

从以上表格的对比不难看出，CHMM 相比 HMM 能获得更多的收益。同时，其回撤幅度也相比 HMM 更小，所以 CHMM 可以有效的减小模型的风险。CHMM 的夏普比率和胜率也是高于 HMM 模型的，所以 CHMM 的模型获利能力是明显强于 HMM 的。本研究也在其他时间段之内同时进行了 HMM 和 CHMM 的回测，在某些时间段 CHMM 最高收益率会比 HMM 高将近一倍，特别是当回测部分为整体牛市的时候，因为 CHMM 中的交易次数会相对少一些，能守住收益。同时，在波动、震荡情况下，CHMM 也会比 HMM 更加合理交易。当两个高度相关的标的被同时提取特征状态时，耦合后的分析毫无疑问会比单个标的更加合理。

相比 HMM 的交易次数过多，CHMM 相对少的交易次数也可以帮助策略获得更高收益。因为以上的回测中并未考虑交易成本，所以 CHMM 实际将相对比 HMM 获得更多收益。

所有使用 CHMM 的回测结果都比单独使用 HMM 的标准版本更胜一筹。回想一下，模型的策略设置采取了一种测试策略，并使用 CHMM 来预测系统中使用的参数的未来值。如果使用 HMM，尽管能跑赢基准收益率，但是无法达到 CHMM 一样的收益水平。

5.4 基于耦合隐马尔可夫模型量化择时策略的优缺点

在音频识别中[31，32]，CHMM 的统计特性允许人们模拟音频和视觉观测序列的状态异步。随着时间的推移，仍然保持它们的自相关性。实验结果表明，在音频以及视觉识别中，CHMM 优于 HMM。

本文中，CHMM 可以同时训练两个标的，从而得到综合的训练模型。对比

HMM，CHMM 可以更全局的预测。同时，CHMM 也有更强的风险控制能力。在任何量化策略中，风险控制都是至关重要的。对于机构或个人，低风险获利毫无疑问是最理想的获利模式。

本研究中的 CHMM 仍然也存在着一些缺点。首先，该模型的特征选取方面依旧可以改进。一个股票或者指数有许多个特征，如何合理提取特征、筛选特征也将影响到模型的训练。如果能找出更贴合模型的特征、能更合理的处理数据，该模型效果将会更好。另一方面，本文也可以更进一步优化 CHMM 的耦合算法。学界中提出了多种 CHMM，本文只实现了其中一种。若实现更为复杂的、耦合程度更高的，有可能将得到更好的收益。

同时，相比 HMM，CHMM 对耦合标的的要求也相当高。如果两个耦合的标的毫无关系或者关联不大，耦合效果将会非常差。当两个标的关联性越大，耦合效果将会更好。而 HMM 对单个标的进行训练，可移植性也更强。

5.5 本章小结

从本章最开始介绍的多种机器学习算法的量化模型到 HMM，再继续优化至 CHMM。每个模型都各自有它们的优点，但是对于金融市场而言，利润是唯一准则。一个量化策略的建立目标便是为了获取更多的利润。

因此，除去各种外界条件限制外，现有的 CHMM 毫无疑问是非常优秀的。

第六章 总结与展望

6.1 总结

显而易见，机器学习的普及令量化择时领域的研究更加充实。从最初的 K 近邻、支持向量机到隐马尔可夫模型，都可以相对于简单的量化择时模型更加有规律的预测股票的短期、中期走势。尽管股票的长期走势取决于基本面情况，但是按照时间序列的多种特征仍然可以帮助人们预测股票在一段时间内的走势。

本研究主要聚焦机器学习算法在股市量化策略上面的运用。通过 HMM、CHMM 完成了量化择时策略的完整流程。相对比前面的其他算法，证明了 HMM 能够识别市场一定期间内的变动能力。HMM 的关键点在与马尔可夫性质的设定，按照时间序列下的数据可否被有效的提取特征。当隐藏状态的转移概率矩阵得出后，策略可以选择合适的交易时间节点，并在市场启动之前进入市场、市场开始急跌时及时止损。

对于 CHMM，本文中完成了基础的 CHMM 的量化择时策略建模。随后，将 CHMM 和 HMM 进行对比，最后得出了客观结论。

显而易见，CHMM 可以很优秀的避免回撤幅度过大，获得更高的收益。同时，CHMM 也可以通过与别的标的耦合，进而拥有更稳定的策略。不可否认，单个标的的 HMM 和两个标的的 CHMM 相对比，多标的将明显从两个指数中获得更多的信息，从而可以分析出更优秀的策略。

6.2 展望

在 HMM 方面，本研究构建的 HMM 模型为静态的。即本模型不能动态、实时的监督每个状态的变化。在某些动态 HMM[33]中，可以有效地增强模型能力。若有一些状态在某些时候不能发出明显信号，该模型有可能错过时机很好的买入时间点。当某些状态滞后的变化后，若买入也可能会经受大幅度回撤。因此，对状态的动态监控是非常必要的。

在模型中，如果能同时监控单个标的的多个指标变化，也会使模型更加有效。荣腾中[34]提出使用多元马尔科夫模型来监控单个股票的走势。

同时，本模型只聚焦于单只股票，但是一支股票或者单个指数是不可能经常存在利润空间的。当该股票或者指数经历了暴涨后，将会有一定时间段的回调期。如果希望该模型能够持续获利，那么就需要加入选股模型。JB Guerard [35]构建了一个选股模型，并且将其运用在美国和日本市场。如果本研究中的模型可以新设立一个股票池，系统实时筛选选定范围内的股票。当模型同时运用在多个股票时，策略收益率将会大大提升。

对于 CHMM 而言，如果坚持耦合资产之间是高度相关或高度不相关，那么 CHMM 能保持其盈利性。标的筛选和划定的观测特征都将影响 CHMM 的有效性。Lee D[36]设计了一个使用黄金和 USDCHF（美元瑞士法郎指数）交易策略。策略涉及使用 CHMM 建模两个指数。因为国际市场中美元与黄金相绑定，因此美元瑞士法郎指数将和黄金价格成反比。最后仍热发现了巨大的获利空间。

如果是多个资产进行耦合而不只是两个，尽管会大大提升计算复杂度，但是效果应该会是非常有效的。学界有人提出非常好的目标，但是一个非常长远的工作目标。这个目标为探索三个耦合的 HMM 对三种货币，即日元美元指数，英镑日元指数和美元英镑指数。在货币交易中，这样的三元交易如果和各国的利率结合，会存在套利机会。若能够发现套利时机，那将会是三元的无风险套利。

如果拥有足够多的数据，也可以进行多时间帧的分析。吕巧云[37]利用了高频量化交易对上海和深圳股市中的某些股票进行操作分析。优秀的高频交易模型将会大大提高模型策略的收益率，充分挖掘一个股票的收益潜力。对于本文，首先，可以使用高粒度的 HMM 的耦合来研究两个标的之间的关系，例如每 30 分钟的深证成指和上证指数。在制定交易策略时，30 分钟内出现的突破可能与价格在临近阶段时间内发生的变化有关。

参考文献

[1] 胡谦. 基于机器学习的量化选股研究[D]. 山东大学, 2016.

[2] Alkhatib K, Najadat H, Hmeidi I, et al. Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm[J]. Ijbhtnet Com, 2013.

[3] Nguyen D T, Nguyen S P, Pham U H, et al. A Calibration-Based Method in Computing Bayesian Posterior Distributions with Applications in Stock Market[M]// Predictive Econometrics and Big Data. 2018.

[4] Yu H, Chen R, Zhang G. A SVM Stock Selection Model within PCA [J]. Procedia Computer Science, 2014, 31(31):406-412.

[5] Louis P. Lukac, B. Wade Brorsen, Scott H. Irwin. A test of futures market disequilibrium using twelve different technical trading systems[J]. Applied Economics, 2015, 20(5):623-639.

[6] Neftci S N. Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Prediction Theory: A Study of "Technical Analysis"[J]. Insurance Mathematics & Economics, 1991, 12(4):549-571.

[7] Kim K J. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1):307-319.

[8] 张玉林, 吴微. 用 BP 神经网络捕捉股市黑马初探[J]. 运筹与管理, 2004, 13(2):123-126.

[9] 周琳杰. 中国股票市场动量策略赢利性研究[J]. 世界经济, 2002(8):60-64.

[10] 楼迎军. 基于 EGARCH 模型的我国股市杠杆效应研究[J]. 中国软科学, 2003(10):49-51.

[11] 吕琦. 基于 SVM 的股票时间序列的预测研究[J]. 吉林工程技术师范学院学报, 2011, 27(7):48-49.

[12] 石赛男. 股票技术分析中 MACD 指标的有效性检验[D]. 西南财经大学, 2011.

[13] 王俊华, 张锡琴, 冯敏敏. 解股票 k 线组合准确性问题的一种时态关联规则算法[J]. 经济论坛, 2008(18):102-103.

[14] 陈收, 杨宽, 廖懿,等. 证券市场中股票成交量对投资组合优化的影响[J]. 管理科学学报, 2002, 5(5):6-10.

[15] Gallant A R, Rossi P E, Tauchen G. Stock Prices and Volume[J]. Review of Financial Studies, 1992, 5(2):199-242.

[16] Wu M, Diao X. Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH & SZ stock markets[C]// International Conference on Computer Science and Network Technology. IEEE, 2016:320-323.

[17] CJ Wellekens. Explicit time correlation in Hidden Markov Models for speech recognition [J]. Proc. ICASSP-87, 1987, 12(3):384-386.

[18] 金辉, 高文. 基于 HMM 的面部表情图像序列的分析与识别[J]. 自动化学报, 2002, 28(4):646-650.

[19] 张亦春, 周颖刚. 论基本分析流派与中国股市有效性--分析以混沌等非线性特征的新范式[J]. 华北金融, 2002(8):8-10..

[20］雷书达，吴文锋. 关于上证 50ETF 期权价格有效性研究——基于期权平价理论分析[J]，价格理论与实践, 2017(4)：116-119

[21] 黄湘松. 基于 HMM 噪声背景下的语音识别方法的研究[D]. 哈尔滨工程大学, 2005.

[22] N Thome ， S Miguet ， S Ambellouis. A Real-Time, Multiview Fall Detection System: A LHMM-Based Approach [D]. IEEE Transactions on Circuits & Systems for Video Technology, 2008, 18(11) :1522-1532.

[23］刘江华, 陈佳品, 程君实. 基于光流及耦合隐马尔可夫模型的动态手势识别[J]. 上海交通大学学报, 2003, 37(5):720-723.

[24] Eisuke Kita, Yi Zuo, Masaaki Harada, Takao Mizuno，Application of Bayesian Network to stock price prediction[M]，Artificial Intelligence Research，2012

[25] K Alkhatib，H Najadat，I Hmeidi，MKA Shatnawi. Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm[M]，Ijbhtnet Com , 2013

[26] Lin Y, Guo H, Hu J. An SVM-based approach for stock market trend prediction[J], International Joint Conference on Neural Networks. 2013:1-7.

[27] Kumawat A K, Khandelwal S. Analysis of timing constraint on combined SVM-HMM classifier and SVM classifier[J], Innovation and Technology in Education. IEEE, 2014:214-218.

[28] SC Huang，HW Wang. Combining Time-Scale Feature Extractions with SVMs for Stock Index Forecasting, Neural Information Processing, International Conference，2010 , 4234 (4) :390-399

[29] E Altay， MH Satman.Stock. Market Forecasting: Artificial Neural Network and Linear Regression Comparison in An Emerging Market, Journal of Financial Management & Analysis , 2005, 18(2) :18-33

[30] P Kumar，H Gauba，PP Roy，DP Dogra. Coupled HMM-based multi-sensor data fusion for sign language recognition[J], Pattern Recognition Letters, 2017 , 86(C) :1-8

[31] Nefian A V, Liang L, Pi X, et al. A coupled HMM for audio-visual speech recognition[J]. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2002:II-2013-II-2016.

[32] Brand M, Oliver N, Pentland A. Coupled hmm for complex action recognition[M], Conference on Computer Vision & Pattern Recognition. 1997.

[33] 张毅,姚圆圆,罗元. 基于 B 参数的改进 HMM 动态手势识别算法[J]. 华中科技大学学报(自然科学版), 2015(s1):416-419.

[34] 荣腾中,肖智,刘朝林. 股票分类指数的多元马尔可夫链模型[J]. 统计与决策, 2012(10):55-57.

[35] Guerard J B. Quantitative Stock Selection in Japan and the United States: Some Past and Current Issues[J]. Journal of Investing, 2006, 15(1):43-49.

[36] Lee D. Trading USDCHF filtered by Gold dynamics via HMM coupling[J]. Computer Science, 2013.

[37] 吕巧云. 面向高频量化交易的沪深 300 股指期货跨期套利研究[D]. 哈尔滨工业大学, 2012.