计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于矩阵画像的金融时序数据预测方法
作者：高世乐，王滢，李海林，万校基
收稿日期：2020-05-31
网络首发日期：2020-08-03
引用格式：高世乐，王滢，李海林，万校基．基于矩阵画像的金融时序数据预测方法[J/OL]．计算机应用．
https://kns.cnki.net/kcms/detail/51.1307.TP.20200731.1052.010.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2020-08-03 12:02:31
网络首发地址：https://kns.cnki.net/kcms/detail/51.1307.TP.20200731.1052.010.html

Journal of Computer Applications　　ISSN 1001-9081
计算机应用　　CODEN JYIIDU　　http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2020060877

基于矩阵画像的金融时序数据预测方法

高世乐[1],王　滢[1],李海林[1,2*],万校基[1]
(1.华侨大学 工商管理学院,福建 泉州 362021;　2.华侨大学 应用统计与大数据研究中心,福建 厦门 361021)
(*通信作者电子邮箱hailin@hqu.edu.cn)

摘　要:针对金融市场中机构交易对股票市场中的散户投资行为具有较强的误导性的现象,提出了一种基于机构交易行为影响下的趋势预测方法。首先,利用时间序列的矩阵画像方法,以股票换手率数据为切入点,构建不同兴趣模式长度下的基于机构交易行为影响的换手率波动知识库。其次,确定待预测股票在兴趣模式长度取何值时,预测结果精确度高。最后,根据该兴趣模式长度下的知识库,预测在机构交易行为影响下单支股票的波动趋势。为验证趋势预测新方法的可行性和准确性,将其与自回归滑动平均模型(ARMA模型)和长短时记忆网络(LSTM网络)预测方法进行对比分析,运用均方根误差(RMSE)与平均绝对百分误差(MAPE)评价指标综合比较70支股票预测结果。实验结果分析表明,与ARMA模型和LSTM网络预测方法相比,在70支的股票价格趋势预测上,新方法有80%以上的股票预测结果更准确。

关键词:机构交易行为;股票趋势预测;兴趣模式发现;矩阵画像
中图分类号:TP273　　**文献标志码:**A

Forecasting method on financial time series data based on matrix profile

GAO Shile[1], WANG Ying[1], LI Hailin[1,2*], WAN Xiaoji[1]
(1. *College of Business Administration, Huaqiao University, Quanzhou Fujian* 362021, *China*;
2. *Research Center for Applied Statistics and Big Data, Huaqiao University, Xiamen Fujian* 362021, *China*)

Abstract: Aimed at the phenomenon that institutional transaction in the financial market is highly misleading to retail investors in the stock market, a trend prediction method based on the impact of institutional transaction was proposed. First, using the time series matrix profile algorithm and taking the stock turnover rate as the entry point, a knowledge base codeDB of turnover rate fluctuations based on the influence of institutional trading behavior under different lengths of motif was constructed. Second, the motif's length of the stock to be predicted was determined, and the accuracy of the prediction result was higher. Finally, the fluctuation trend of single stocks under the influence of institutional trading behavior was predicted through this knowledge base. In order to verify the feasibility and accuracy of the new method of trend prediction, compared it with Auto-Regressive Moving Average model (ARMA model) and Long Short Term Memory network (LSTM network), and the root-mean-square error (RMSE) and Mean Absolute Percentage Error (MAPE) evaluation indicators were used to compare the 70 stocks' prediction results of the three methods. The analysis of the experimental result was showed that, compared with the ARMA model and the LSTM network prediction method, in the prediction of 70 stock price trends, more than 80% of the stock prediction results of the new method were more accurate.

Key words: institutional trading behavior; stock trend forecast; motif discovery; matrix profile

0　引言

机构投资使股市环境产生了多元化的投资者结构,对股市具有一定的冲击影响,同时也可以帮助稳定金融市场。目前大部分股票交易都有机构投资者的参与,其行为对股价波动的影响较大。研究机构交易行为对股价波动的影响有利于帮助散户进行股票投资,本文将从股票总体市场出发找出机构交易行为,分析机构交易行为对个股价格波动的影响,预测个股股价波动趋势,识别机构投资者操纵股票的行为,进而降低散户投资风险和提高投资回报。

机构投资者的交易行为对股价波动是否具有一定的影响,部分学者对其进行了研究。何佳等[1]对机构投资者能否稳定股市进行了实证研究,得出了机构投资者对股价波动的影响并不是确定的。王咏梅和王亚平[2]从机构投资者与市场信息效率的关系出发,对深市A股的上市公司数据进行实证研究,得出机构投资者的过度交易行为会损害信息效率,加剧股市的不稳定性,造成股价波动。刘京军和徐浩萍[3]根据换手率特点将机构投资者分为长期投资者与短期机会主义者,经过实证分析得出较长期机构投资者而言,短期机构投资者的交易行为加剧了股市的不稳定性,加剧了市场波动,长期

收稿日期:2020-05-31;**修回日期:**2020-07-04;**录用日期:**2020-07-16。
基金项目:国家自然科学基金资助项目(71771094);福建省自然科学基金项目(2019J01067)。
作者简介:高世乐(1972—),男,黑龙江绥化人,讲师,博士,主要研究方向:数据挖掘、复杂网络;　王滢(1997—),女,福建宁德人,主要研究方向:时间序列数据挖掘;　李海林(1982—),男,福建龙岩人,教授,博士,CCF会员,主要研究方向:数据挖掘、商务分析;　万校基(1984—),男,江西南昌人,讲师,博士,主要研究方向:数据挖掘、金融分析。

机构投资者在稳定市场方面具有一定的作用。史永东和王瑾乐[4]通过得分匹配模型验证了机构投资者的频繁交易会加剧市场的不稳定性，产生股价波动。因此，机构投资者的短期交易行为会加剧市场的不稳定性，进而导致股价波动，以机构投资者的短期交易行为为出发点，研究个股价格波动趋势是可行的。

在股票价格和趋势等预测方面，学者们也进行了大量研究[5-12]，提出了各种不同且有效的预测方法，例如自回归滑动平均模型（Auto-Regressive Moving Average Model，ARMA 模型）、支持向量机和神经网络等。许多学者又在传统方法的基础上进行了改进，以取得更好的预测效果。张贵生和张信东[6]在 ARMA-GARCH（Auto-Regressive Moving Average—Generalized Auto-Regressive Conditional Heteroskedasticity）模型的基础上引入因变量滞后项的微分信息，提出了 ARMAD-GARCH 模型，较之原模型取得了更准确的预测结果。吴少聪[7]通过对具有代表性的 13 支 A 股股票建立混合模型进行股票趋势预测，并据此建立了股票信息服务平台，且验证了其比长短时记忆网络（Long Short Term Memory，LSTM）模型和差分整合移动平均自回归模型（Auto-regressive Integrated Moving Average Model，ARIMA）的预测准确率高。宋刚等[8]提出了基于自适应粒子群优化的 LSTM 股票价格预测模型，对 LSTM 模型进行了改进，提高了准确率且具有普适性。石浩[9]通过建立基于递归神经网络的股票预测模型，并与传统的神经网络模型进行比较，突出其所建模型的价值。谢琪等[10]建立了一种基于长短记忆神经网络集成学习的金融时间序列预测模型，并使用准确率、精确率、召回率、F1 值与曲线下面积（Area Under Curve，AUC）这 5 个评价分类算法的指标对传统神经网络模型与该模型的预测结果进行评价，从而验证该模型优于其他传统神经网络模型。Kei Nakagawa 等人[11]对股票价格波动模式进行了 k-medoids 聚类，并利用索引动态时间规整法[12]提取了代表性波动模式作为预测的特征值，并据此对股价进行预测。

目前在机构交易行为对于个股趋势影响以及通过机构交易行为来预测股价波动等方面的研究甚少，学者们更多的是从股票市场的总体范围来研究机构投资行为对股市稳定性以及股价波动的影响，在预测股价波动方面更多的是基于收盘价序列数据进行预测。相对而言，机构对股票的操纵行为通常是间断性的且时间持续性不长，使得股票时间序列数据的局部性信息显得更为重要。然而，传统的时间序列预测方法是基于数据的整体信息考虑，缺乏对局部性数据的重视。鉴于传统模型和方法对数据具有研究假设前提的要求以及局部性时间片段的重要性，使用时间序列数据挖掘的相关技术和方法对其进行研究显得尤为重要。且矩阵画像算法在时间序列的局部性研究上具有一定的优越性。因此，本文借助时间序列矩阵画像算法对深市 A 股主要股票历史换手率数据建立基于机构交易行为的序列片段知识库 codeDB，利用知识库 codeDB 可从单支股票出发对个股价格波动趋势进行预测。与传统 ARMA 回归模型和 LSTM 网络等预测方法相比，新方法不仅从新的视角对股票时间序列数据进行预测，还对个股价格波动分析具有更好的预测效果。

1 矩阵画像相关理论

矩阵画像（Matrix Profile，MP）[13-24]是一种用于时间序列数据挖掘的数据结构，可用于主题发现、密度 估计、异常检测、规则发现、分割和聚类等。

定义 1 时间序列数据是按时间顺序排列的实数值数据，用序列 T 表示，且 $T = t_1, t_2, t_3, ..., t_n$，其中 n 是 T 的长度。

定义 2 子序列表示在原始序列 T 中截取长度为 m 的一段序列，用 $T_{i,m}$ 表示，即是从 T 中第 i 个位置开始的长度为 m 的连续子集。形式上表示为 $T_{i,m} = t_i, t_{i+1}, ..., t_{i+m-1}$，其中 $1 \leqslant i \leqslant n - m + 1$。

定义 3 距离画像 $\boldsymbol{D}$ 是时间序列 T 中不同的子序列间的距离矩阵。给定一个时间序列 T，子序列长度为 m，从 $T_{i,m}$（$i = 1, 2, 3, \ldots$）开始计算其与其他子序列片段的距离，得到一个距离矩阵 $\boldsymbol{D}$，即 $\boldsymbol{D}_i$ 是给定查询子序列与时间序列中的每个子序列之间的距离向量。

形式上表现为 $\boldsymbol{D}_i = [d_{i,1}, d_{i,2}, ..., d_{i,n-m+1}]$，其中 $d_{i,j}$ 是 $T_{i,m}$ 和 $T_{j,m}$ 之间的距离，计算子序列片段间的距离公式为：

$$d_{i,j} = \sqrt{2m\left(1 - \frac{QT_{i,j} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)} \tag{1}$$

m 表示子序列长度，μ_i 表示子序列 $T_{i,m}$ 的均值，σ_i 表示子序列 $T_{i,m}$ 的标准差，$QT_{i,j}$ 表示子序列 $T_{i,m}$ 与子序列 $T_{j,m}$ 的点积[14]。特别地，当子序列数据经过 *zscore* 标准化后，即 $\mu = 0, \sigma = 1$ 时，公式（1）转变成：

$$d_{i,j} = \sqrt{2m\left(1 - \frac{QT_{i,j}}{m}\right)} \tag{2}$$

即标准化后的序列在计算距离画像时，只要计算好子序列间的点积便可快速得到该序列的距离画像。

定义 4 时间序列数据 $A = [a_1, a_2, ..., a_n]$ 和 $B = [b_1, b_2, ..., b_n]$，则 A 与 B 之间的点积 $\boldsymbol{QT}$ 计算公式为：

$$\boldsymbol{QT} = \sum_{i=1}^{n} a_i \times b_i \tag{3}$$

点积是在实现矩阵画像算法过程中会用到的重要公式，是用于计算矩阵画像中距离画像的重要部分。一个子序列与一条时间序列中所有子序列的点积的具体算法见[15]，该算法时间复杂度为 $O(n \log n)$，与传统的计算过程相比，计算效率显著提高。

定义 5 矩阵画像 $\boldsymbol{MP}$ 是时间序列 T 中每个子序列 $T_{i,m}$ 与其最相似片段（即距离最小值）之间的距离值组成的向量。距离画像相当于是每个子序列片段与其他所有子序列片段中的距离最小值。形式上，$\boldsymbol{MP} = [\min(\boldsymbol{D}_1), \min(\boldsymbol{D}_2), ..., \min(\boldsymbol{D}_{n-m+1})]$，其中 $\boldsymbol{D}_i (1 \leqslant i \leqslant n - m + 1)$ 是由子序列 $T_{i,m}$ 与其他所有子序列片段之间的距离所构成的向量。

定义 6 兴趣模式（motif）是指一条或多条时间序列中最相似的子序列片段，即在每个子序列片段所对应的已是其最近距离值（即子序列的 $\boldsymbol{MP}$ 值）的情况下，再寻找 $\boldsymbol{MP}$ 中的极小值。在寻找兴趣模式之前，需先计算出要寻找模式的 $\boldsymbol{MP}$ 值，再从 $\boldsymbol{MP}$ 中获得极小值对应的模式，进而找到兴趣模式。

定义 7 矩阵画像索引是用来记录每个子序列片段的最近子序列片段所在位置的，记为 $\boldsymbol{MPI}$，其为整数向量，即 $\boldsymbol{MPI} = [I_1, I_2, ..., I_{n-m+1}]$，$I_i = \arg\min_j (D_i(j))$，$D_i(j)$ 表示距离向量 $\boldsymbol{D}_i$ 中的第 j 个距离元素。

当子序列片段的 $\boldsymbol{MP}$ 值相同时，通过 $\boldsymbol{MPI}$ 可以快速简便地定位到 $\boldsymbol{MP}$ 相同的值，从而快速寻找序列的兴趣模式。计

算矩阵画像的算法目前有 stamp[14]、stomp[15]和 scrimp[21]等。本文使用的算法是 stomp,其具体过程见[15]。它与 stamp 最主要的区别在于子序列片段间点积的计算效率上。stomp 算法在点积处理上,遵循了以下思路,降低了算法的时间复杂度,使算法更加高效。由于

$$\begin{cases} QT_{i-1,j-1} = \sum_{k=0}^{m-1} T_{i-1+k}T_{j-1+k}; \\ QT_{i,j} = \sum_{k=0}^{m-1} T_{i+k}T_{j+k}. \end{cases}$$

因此可得下面公式:

$$QT_{i,j} = QT_{i-1,j-1} - T_{i-1,j-1} + T_{i+m-1}T_{j+m-1} \tag{4}$$

使用公式(4)可在时间复杂度为 O(n)的情况下完成 $\boldsymbol{QT}$ 的更新,提高了算法的计算效率,使矩阵画像算法更加高效。

矩阵画像的演算过程如图 1 所示,该示意图体现了求解一条长度为 n 时间序列 T 的 $\boldsymbol{MP}$ 值的过程。图中Δ指的是计算子序列片段 $T_{i,m}$ 与时间序列中其它所有长度为 m 的子序列片段的距离向量 $\boldsymbol{D}_i$,接着对每个距离向量 $\boldsymbol{D}_i$ 求最小值,即 $MP_i = \min(\boldsymbol{D}_i)$ 且 $MPI_i = \arg\min_j (D_i(j))$,便可得到所有子序列片段的距离画像 $\boldsymbol{MP} = [MP_1, MP_2, ..., MP_{n-m+1}]$。需要说明的是,由于相邻两条子序列片段重叠太多,会造成时间相近的序列片段互为最相似片段,不利于兴趣模式的发现。故在排除与 $T_{i,m}$ 重复长度超过 $m/2$ 的子序列的距离后,取距离向量 $\boldsymbol{D}_i$ 的最小值作为 $T_{i,m}$ 的 $\boldsymbol{MP}$ 值。

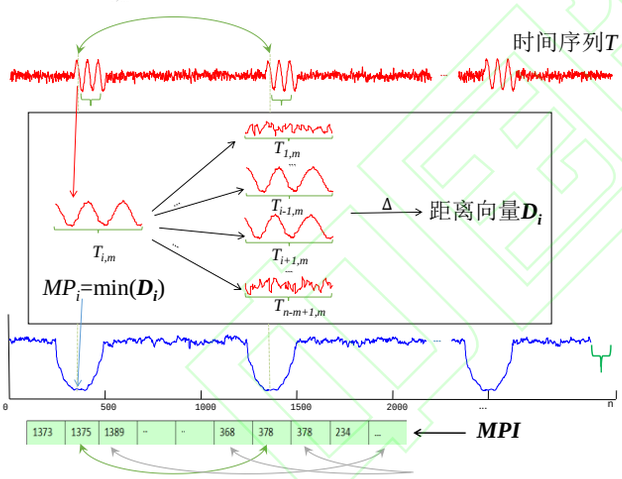

图 1 Matrix Profile 演算过程

Fig. 1 The calculation process of matrix profile

2 股票价格波动趋势预测方法

首先,使用矩阵画像方法,以金融股票的换手率数据为切入点,分别构建不同兴趣模式长度下的基于机构交易行为影响的换手率波动知识库。其次,确定待预测股票在兴趣模式长度取何值时,预测结果精确度高。最后,基于该兴趣模式长度下的知识库,预测在机构交易行为影响下的单支股票价格波动趋势。

2.1 机构交易行为知识库

换手率也称"周转率",指在一定时间内股票市场中股票转手交易的频率,体现了股票的流通性强弱。换手率公式为:

$$H = \frac{V}{TN} \times 100\% \tag{5}$$

其中 H 表示换手率,V 表示成交量,TN 表示发行总股数。

选择使用换手率数据代替使用成交量的主要原因是,在表示机构交易行为时,换手率数据能够反映交易的频率和交易情况,相对更能说明机构和股民在一定时期内的交易行为。一般情况下,在股票市场中针对某一支股票的散户交易量并不大,若没有机构投资者的介入,其换手率一般不高。就一般经验来说,换手率具有以下特征:(a) $H < 3\%$ 表示股票交易行为主要是散户参与;(b) $H > 7\%$ 表示股票交易行为主要是机构投资者参与。因此,本文主要根据换手率的高低来定义机构交易行为(Institutional Trading Behavior,ITB),且将存在换手率大于 8% 的股票序列片段定义为存在机构交易行为。

2.1.1 构建知识库

根据交易数据,可以构建反映机构主要交易行为的知识库,其为包含了具有典型代表意义的机构交易行为的数据库,为散户们提供有关机构交易行为的相关信息和知识,构建知识库的算法如算法 1 所示。

算法 1 BuildDB(TS,m)

输入:股票换手率序列集合 TS,兴趣模式长度 m

输出:知识库 codeDB

1 MPs=stomp(TS,m) //具体见[15]

2 $motifs$=findmotif(MPs)

3 for i=1 to length($motifs$)

4 if($motifs$[i] include ITB)

5 Save $motifs$[i] into codeDB

6 end if

7 end for

8 return(codeDB)

算法 1 中 1-2 行是将处理好的股票数据用 Matrix Profile 算法找出 motif,3-7 行是剔除不存在机构交易行为的 motif,并将剩余的 motif 前期片段、motif 片段与 motif 后续片段分别存入知识库中。

2.1.2 补充库

由于在预测过程中有可能会出现一些情况,即此时的股票数据序列是存在机构交易行为的,但其与知识库中的兴趣模式的匹配度并不高。若强行进行预测,预测结果有极大的概率会偏离实际结果,因此这种情况是不进行预测的。然而,存在机构交易行为的片段是值得注意的,其处理方式是先将该片段暂保存在其他数据库中,称该数据库为补充库(supDB)。补充库是将当前存在机构交易行为但与知识库中的兴趣模式匹配度不高的子序列片段进行保存,以备在完备知识库中使用。具体算法如算法 2 所示。

算法 2 BuildsupDB(QS,$codeDB$,ε)

输入:70 支股票中新出现且 codeDB 未包含的换手率序列集 QS,知识库 codeDB,匹配程度的阈值 ε

输出:补充库 supDB

1 for i=1 to length(QS)

2 if QS[i] include ITB then

3 Dis=dist(QS[i],$codeDB.motif$)

//计算 QS[i]与 codeDB 中所有 motif 序列的距离

4 $best$=which. min(Dis) //选出匹配的片段

5 if Dis[$best$]/$length$(QS[i]) > ε then

//QS[i]与所有 motif 序列的匹配度都不高

6 Save $QS[i]$ into supDB
7 end if
8 end if
9 end for
10 return(supDB)

算法2中2-8行表示将存在机构交易行为且其与知识库中所有motif序列的相似性程度都不高的序列片段存入补充库中。

2.1.3 完备知识库

补充库中的子序列片段即是知识库的完备项。当能在补充库找到兴趣模式时,说明该片段具有一定的代表性,则将该兴趣模式所对应的片段扩充到知识库中,具体算法如算法3所示。

算法3 perfectDB(*supDB*,*m*)
输入:补充库supDB,子序列长度*m*
输出:知识库codeDB
1 *MPs*=stomp(supDB,*m*) //具体见[15]
2 *motifs*=findmotif(*MPs*)
3 Save *motifs* into codeDB
4 return(codeDB)

算法3中1-2行中是从supDB中的序列集合中寻找到兴趣模式,第3行便是将找到的兴趣模式存入知识库codeDB中。

2.2 最佳模式匹配

对某支股票进行预测时,兴趣模式的长度不同,拟合的效果也会不同。确定兴趣模式长度与预测天数二者满足何种关系时预测效果较好是提高预测效果的手段之一,即寻找兴趣模式长度与预测天数的最佳模式匹配。本文采取的方法主要是使用历史数据进行多次训练,找出已知预测天数下拟合效果好的兴趣模式长度。首先,提出已知预测天数和兴趣模式长度下的趋势预测算法,如算法4所示。

算法4 PredictTrend(Q,TP,m,t,TS,ε)
输入:待预测片段Q,对应的股票收盘价TP,兴趣模式长度m,预测天数t,70支股票换手率序列集合TS,匹配程度的阈值ε
输出:预测出的价格趋势PT
1 codeDB=BuildDB(TS,m) //具体见算法1
2 if Q include ITB then //Q存在机构交易行为
3 Dis=dist(Q,codeDB. motif)
//计算Q与codeDB中的所有motif序列的距离
4 *best*=which. min(*Dis*) //选出与Q匹配的片段
5 if Dis[best]/length(Q)>ε then
//当不相似程度大于ε时,即匹配度不够
6 Save Q into supDB
7 end if
8 else
9 $PT = TP[(best + m):(best + m + t)]$
10 return(PT)
11 end else
12 end if
13 else //Q不存在机构交易行为
14 *Dis*=dist(Q,*codeDB. motifbefore*)
15 *best*=which. min(*Dis*) //选出与Q匹配的片段
16 if $Dis[best]$/length(Q)>ε then
//当不相似程度大于ε时,即匹配度不够
17 print("无法预测")
18 end if
19 else
20 PT=$TP[(best+m):(best+m+t)]$
21 return(PT)
22 end else
23 end else

算法4中第1行是构建知识库的过程,2-12行表示待预测片段存在机构交易行为时的具体做法,第2-4行是选取知识库中与待预测片段相似度最高的兴趣模式对应片段,第5-7行对二者是否相似度高进行判断。若相似度高则获取未来股价波动趋势;反之,将待预测片段存入补充库中。第13-23行表示待预测片段不存在机构交易行为时的具体做法,第14-15行是选取知识库中与待预测片段相似度最高的兴趣模 式前序对应的片段。若相似度高,则获取未来股价波动趋势;反之,则不可预测。

其次,寻找兴趣模式长度与预测天数最佳模式匹配的算法,如算法5所示。

算法5 DeterLen(T,TP,t,num,TS,ε)
输入:待预测片段所在股票的换手率序列T,对应的股票收盘价TP,预测天数t,试验次数num,70支股票换手率序列集合TS,匹配程度的阈值ε
输出:预测效果最佳的子序列片段长度m
1 for i =1 to num //表示进行num次不同待预测片段
2 for m =(t+10) to (12×t)
3 k=random. int(1,(length(T)-m+1))
//产生在length(T)内的随机整数
4 PT=PredictTrend(Tk,m,TP,m,t,TS,ε)
//获取预测的趋势,见算法4
5 $R[m,i]$=RMSE(*real*,*PT*) //具体计算见公式(7)
6 end for
7 end for
8 $Ave[m]$=Mean($R[m,]$)
//计算确定m长度下的各个测试片段的均值
9 return(which. min(Ave))
//返回均值最小的子序列长度

算法5的目的是在已知预测天数的情况下获取基于历史数据训练下的最佳兴趣模式长度,其前提在于已知待预测片段及其所在的股票序列。第2行定义的是训练时兴趣模式的取值范围,本文使用实验验证的方法来确定最佳兴趣模式长度,故在训练实验中会尽量扩大长度的取值范围。第3-5行是随机选取一定长度的待预测片段所在股票历史数据,进行预测训练,并计算*RMSE*值来判断预测拟合程度的好坏。第8-9行是对之前做过的多次实验进行整理计算,综合选出最优的兴趣模式长度与预测天数的匹配模式。

2.3 预测算法

本文主要研究的是在机构交易行为影响下的个股价格波动,但这并不是意味着在没有存在机构交易行为的情况下,新方法就不能进行预测。由图2即可看出若待预测片段不具有机构行为,可以与知识库中保存的兴趣模式前期序列进行匹

配。若匹配度高,则有很大概率认为待预测片段可能即将迎来机构交易行为;若匹配度不高时,则表示无法预测。若存在机构交易行为,则与知识库中的兴趣模式(motif)进行匹配,匹配度高则返回未来可能的股价波动,匹配度不高则将该预测片段存入补充库中。其具体过程如图2所示:

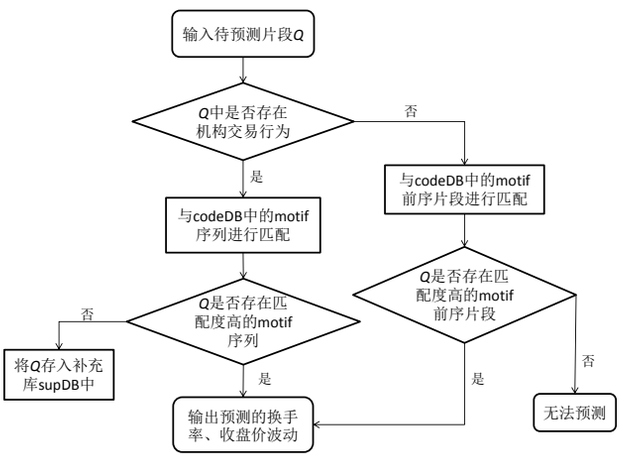

图2 基于矩阵画像的预测过程

Fig. 2 The prediction process based on matrix profile

在机构交易行为的影响下对股价波动进行预测(MP based Prediction, MPP)的具体算法如算法6所示。

算法6 MPP(Q,T,TP,t,TS,ε)

输入:待预测片段Q,Q所在股票的换手率序列T,对应的股票收盘价TP,预测天数t,70支股票换手率序列集合TS,匹配程度的阈值ε

输出:预测出的价格趋势PT

1 m=DeterLen($T,TP,t,200,TS,\varepsilon$) //见算法5

2 PT=PredictTrend(Q,TP,m,t,TS,ε) //见算法4

3 return(PT)

算法6中第1行是对Q所在的股票序列的历史数据进行训练,找出预测效果最佳的兴趣模式长度。第2行是在确定的兴趣模式长度下对Q的后续股价趋势进行预测。

3 实验分析

3.1 数据收集与处理

选取2014—2018年我国深市A股股票作为研究对象,并对这些数据进行整理:(a)剔除已经停市的股票;(b)剔除2014—2018年连续5天停止交易的股票;(c)剔除2014—2018年每年交易日期不足180的股票。整理得到70支股票样本数据,具体股票代码如表1。

实验主要任务是将70支股票从2014年1月2日到2018年2月1日为止共70万条换手率数据用于创建知识库codeDB,预测这70支股票2018年4月10日以后的股价趋势波动。在兴趣模式长度与预测天数的模式匹配中,模拟预测用到的训练数据均从对应待预测股票中随机选取。所使用的数据主要是股票的换手率数据与收盘价数据,实验之前要对股票的收盘价数据根据以下公式进行标准化处理:

$$TP_i = \frac{TP_i - \mu_{TP}}{\sigma_{TP}} \tag{6}$$

其中TP_i指的是第i个收盘价,μ_{TP}指的是整条收盘价序列的均值,σ_{TP}指的是整条收盘价序列的标准差。

表1 70只深市A股股票代码表

Tab. 1 the stock code table of 70 A-shares in ShenZhen

代码	名称	代码	名称	代码	名称
000001	平安银行	000572	海马汽车	000809	铁岭新城
000011	深物业A	000573	粤宏远A	000819	岳阳兴长
000014	沙河股份	000596	古井贡酒	000822	山东海化
000027	深圳能源	000598	兴蓉环境	000823	超声电子
000039	中集集团	000623	吉林敖东	000830	鲁西化工
000049	德赛电池	000631	顺发恒业	000848	承德露露
000059	华锦股份	000632	三木集团	000869	张裕A
000088	盐田港	000637	茂化实华	000880	潍柴重机
000089	深圳机场	000652	泰达股份	000886	海南高速
000096	广聚能源	000680	山推股份	000888	峨眉山A
000151	中成股份	000702	正虹科技	000895	双汇发展
000157	中联重科	000722	湖南发展	000898	鞍钢股份
000338	潍柴动力	000725	京东方A	000915	山大华特
000402	金融街	000726	鲁泰A	000919	金陵药业
000419	通程控股	000729	燕京啤酒	000921	海信家电
000423	东阿阿胶	000731	四川美丰	000937	冀中能源
000528	柳工	000758	中色股份	000951	中国重汽
000532	华金资本	000759	中百集团	000965	天保基建
000548	湖南投资	000767	漳泽电力	000966	长源电力
000550	江铃汽车	000777	中核科技	000983	西山煤电
000554	泰山石油	000780	平庄能源	000985	大庆华科
000559	万向钱潮	000785	武汉中商	000988	华工科技
000561	烽火电子	000789	万年青		
000570	苏常柴A	000800	一汽轿车		

3.2 预测结果评测标准

为了对不同方法的预测结果进行比较,引入了均方根误差与平均绝对百分比误差来对预测结果进行评估,从而比较不同预测方法之间的优劣性。

(1)均方根误差(Root-Mean-Square Error,RMSE)。

均方根误差是用来衡量实际值与预测值之间的偏差。具体公式为:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{real,i} - x_{predict,i})^2}{n}} \tag{7}$$

其中$x_{predict,i}$指的是第i个预测值,$x_{real,i}$指的是第i个真实值,n指的是预测值或真实值的个数。$RMSE$的值越小,说明预测效果越好,预测值与实际值之间的偏差越小。

(2)平均绝对百分误差(Mean Absolute Percentage Error,MAPE)。

平均绝对百分误差可以用来衡量一个模型预测结果的好坏,通过比较不同方法的MAPE值才能知道对应模型和方法预测的准确性或者优劣性,MAPE值越小,说明模型预测的准确性较高。具体公式为:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_{real,i} - x_{predict,i}}{x_{real.i}}\right| \times 100\% \tag{8}$$

其中$x_{predict,i}$指的是第i个预测值,$x_{real,i}$指的是第i个真实值,n指的是预测值或真实值的个数。

3.3 实例分析

在使用MPP方法时,需先确定兴趣模式motif的长度,即m值。由于不同的m值会造成不同的预测结果,其拟合效果差异性较大,故选好合适的m值有利于得出拟合效果好的预测结果。为了确定兴趣模式的长度,根据不同motif长度设定对待预测片段所在股票的历史数据进行训练,选择对应预测

效果最佳的长度为兴趣模式的长度。如图3所示，在不同m值下进行多次训练得到的$RMSE$值所构成的盒图。

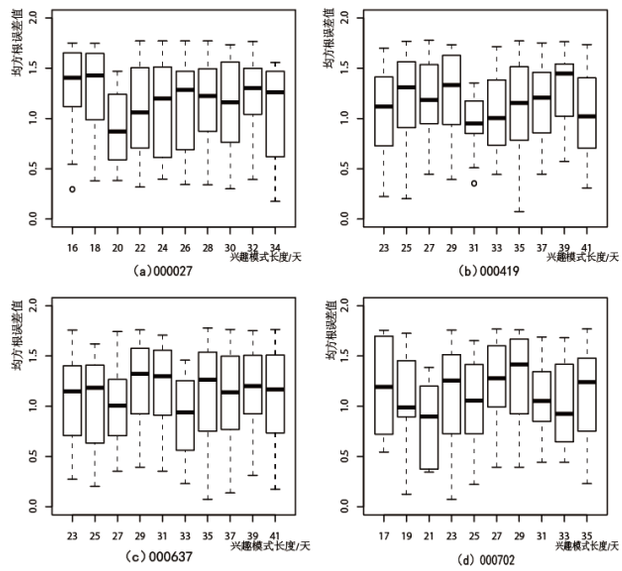

图3 不同m值下的RMSE值比较

Fig. 3 Comparison of RMSE values at different m values

根据算法5来确定motif长度，以股票代码为000027、000419、000637和000702的换手率和收盘价数据为例，通过预测得到了如图3所示的$RMSE$值分布，通过误差分析可以获得对应股票片段进行MPP预测时可选取的合适的兴趣模式长度。为验证本文提出方法MPP的性能，将MPP与ARMA模型和LSTM网络预测方法作对比，同时预测70支股票自2018年4月10日起未来5个交易日的股价趋势波动。根据代码为000027、000419、000637和000702的股票的换手率数据和收盘价数据，由图3中盒图的中位数可选得四只股票较好的兴趣模式长度分别为20、31、33和21。使用算法4预测自2018年4月10日起未来5个交易日的趋势波动，预测所得的价格波动与实际价格波动的拟合情况具体如图4所示。

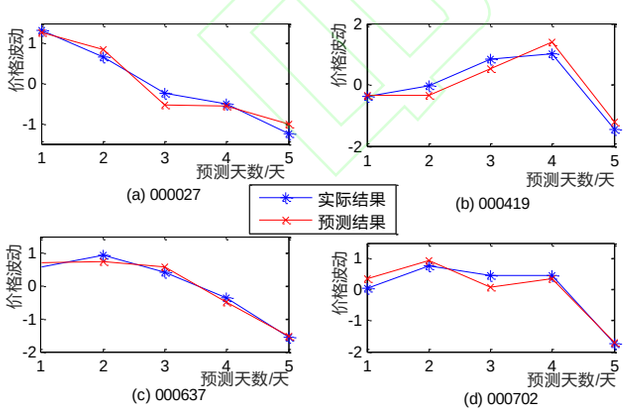

图4 MPP预测效果

Fig. 4 Prediction effect of MPP

图4中带*的蓝线部分表示实际的价格波动，带×的红线部分表示预测的价格波动，从图4可以看出这四支股票未来五个交易日的股价波动涨幅趋势基本相同，且涨幅程度差异不大，预测的效果较好。

自回归滑动平均模型(Auto-Regressive Moving Average Model，ARMA模型)是研究时间序列的重要方法，是目前常用的用于拟合平稳序列的模型。它可以细分为自回归模型(Auto-Regressive Model，AR模型)、移动平均模型(Moving Average Model，MA模型)和ARMA模型。在ARMA模型进行对时间序列数据进行建模分析时，通常用AIC信息准则(Akaike Information Criterion)与BIC准则(Bayesian Information Criterion)对模型的优劣进行评估。AIC准则与BIC准则的具体公式如下：$AIC = -2\ln(MLV) + 2NUP$和$BIC = -2\ln(MLV) + \ln(n) \times NUP$，其中$MLV$表示模型的极大似然函数值，$n$表示时间序列的长度，$NUP$表示模型中未知参数的个数。当AIC准则与BIC准则的值最小时，认为此时的模型达到最优。

使用AIC准则与BIC准则确定ARMA模型中的参数，构建好模型后，通过实验可以得到四支股票000027、000419、000637和000702从2018年4月10日起未来5个交易日的价格趋势波动预测与实际的个股价格波动趋势的拟合效果，如图5所示。

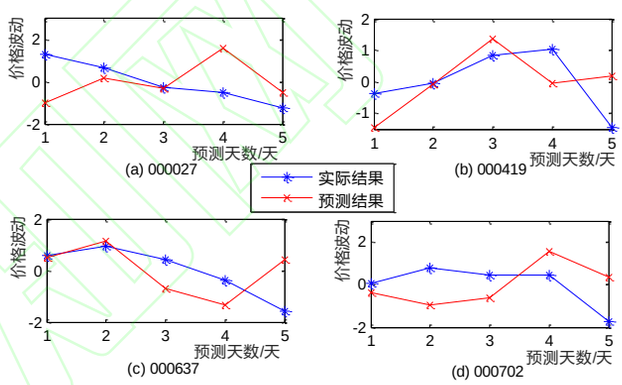

图5 ARMA预测效果

Fig. 5 Prediction effect of ARMA

图5中带*的蓝线部分表示实际的价格波动，带×的红线部分表示预测的价格波动，由图5中四幅图可以看出ARMA模型的预测效果不太理想，预测结果的拟合效果并不好。

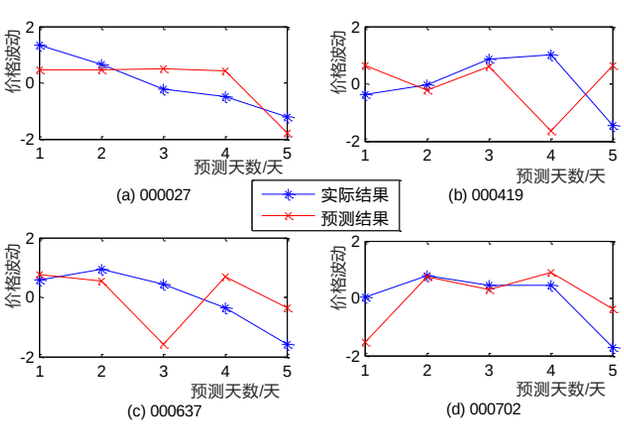

图6 LSTM网络预测效果

Fig. 6 Prediction effect of LSTM

长短时记忆网络(Long Short Term Memory，LSTM)是一种特殊的循环网络(Recurrent Neural Network，RNN)类型，解决了RNN存在的长期依赖问题，对传统的RNN进行了隐层中结构上的改进，具有长期记忆能力，LSTM引入“门”的结构来去除或者增加信息到细胞状态的能力，LSTM网络中有输入门、输出门和遗忘门。通过利用LSTM模型预测相同股票的价格趋势波动，其价格趋势波动预测与实际的个股价格波动趋势

的拟合效果程度如图6所示。

图6中带*的蓝线部分表示实际的价格波动,带×的红线部分表示预测的价格波动。图6中(a)、(b)和(c)三幅图的涨幅趋势的预测效果不太理想,(d)中可看出预测到的涨幅趋势与实际的涨幅趋势大致相同,只是涨幅程度差异较大。

3.4 实验评估

使用ARMA模型和LSTM网络以及基于机构交易行为下的趋势预测MPP这三种方法对70支深圳A股进行预测分析,即预测自2018年4月10日起未来5天(A时间段)的股价趋势波动,且使用RMSE和MAPE这两种评价指标对三种方法的预测结果进行评价。

在使用基于机构交易行为下的趋势预测方法MPP进行价格波动预测,所定的匹配程度的阈值$\varepsilon = 1.2$,在预测过程中000014(沙河股份)和000554(泰山石油)这两支股票的匹配程度不够,将其剔除,终预测数据是68支股票。由于000632(三木集团)、000767(漳泽电力)和000809(铁岭新城)这三支股票在2018年4月10日至4月16日中5个交易日的收盘价经过标准化处理后存在0值,*MAPE*值无法计算。因此,进行预测结果RMSE评价指标比较的股票总数为68支,MAPE评价指标比较的股票总数为65支,三种方法对不同股票A时段的预测误差如表2所示(所有数据均保留小数点后两位)。表2中黑体的数值表示股票在RMSE与MAPE评价下的预测误差最小值,可以得出共有62支股票的*RMSE*最小值与56支股票*MAPE*最小值来自基于机构交易行为下的趋势预测方法MPP,且RMSE与MAPE评价下的均值最小值和标准差小值均来自MPP,由此可知MPP方法的拟合结果优于其他两种方法。

表2 三种方法的预测误差

Tab. 2 Forecast error of three methods

股票	RMSE			MAPE		
代码	*ARMA*	*LSTM*	*MPP*	*ARMA*	*LSTM*	*MPP*
000001	1.34	0.90	**0.50**	325.81	184.40	**163.08**
000011	1.67	1.38	**0.44**	218.61	190.21	**86.71**
000027	1.47	0.70	**0.19**	151.76	124.00	**36.62**
000039	1.34	1.20	**0.86**	951.06	638.14	**199.50**
000049	1.34	1.16	**0.49**	162.89	153.04	**77.65**
000059	1.44	1.46	**1.05**	130.67	157.73	**100.15**
000088	1.04	0.73	**0.25**	106.78	62.53	**36.65**
000089	0.79	1.68	**0.73**	**63.05**	162.87	70.92
000096	1.74	1.02	**0.91**	224.09	129.39	**124.46**
000151	1.30	0.69	**0.30**	128.37	71.40	**69.49**
000157	0.47	1.02	**0.40**	48.10	118.46	**41.63**
000338	1.29	1.21	**1.06**	1047.13	1089.70	**265.42**
000402	1.53	1.15	**0.53**	228.85	214.77	**116.45**
000419	1.01	1.60	**0.28**	**126.36**	253.18	199.00
000423	1.37	1.14	**0.85**	428.29	348.64	**307.60**
000528	1.53	1.28	**0.68**	299.35	738.07	**183.48**
000532	**0.49**	1.39	0.61	**209.53**	308.24	339.29
000548	1.57	1.50	**0.27**	373.99	229.84	**42.56**
000550	1.43	0.59	**0.45**	531.70	337.78	**190.67**
000559	1.71	1.42	**0.88**	218.62	189.05	**129.25**
000561	1.63	1.59	**1.05**	319.62	**179.04**	308.26
000570	1.02	1.27	**0.44**	106.93	177.36	**48.80**
000572	1.18	1.10	**0.26**	679.26	299.96	**100.74**
000573	1.11	1.29	**1.06**	188.27	257.38	**146.89**
000596	1.63	1.26	**0.91**	231.26	144.38	**94.77**
000598	0.73	1.47	**0.57**	76.47	115.16	**56.45**
000623	0.94	1.34	**0.49**	198.39	173.26	**98.89**
000631	1.53	1.72	**1.06**	201.58	255.85	**142.12**
000632	1.05	1.12	**0.53**	/	/	/
000637	1.12	1.17	**0.13**	140.98	183.66	**22.29**
000652	1.44	1.62	**0.91**	135.31	167.37	**112.91**
000680	1.34	0.94	**0.30**	160.19	109.36	**34.78**
000702	1.41	0.95	**0.23**	370.25	805.59	**168.29**
000722	0.47	0.92	**0.30**	47.63	115.05	**30.05**
000725	0.61	**0.32**	0.56	94.45	**70.97**	95.40
000726	1.25	1.40	**0.42**	506.04	368.97	**134.05**
000729	1.66	1.08	**0.66**	431.10	713.33	**296.20**
000731	1.64	0.71	**0.43**	241.75	126.32	**65.48**
000758	1.57	0.90	**0.39**	175.09	162.28	**54.49**
000759	1.45	1.63	**0.97**	346.47	204.87	**115.66**
000767	1.26	1.47	**0.62**	/	/	/
000777	1.17	1.47	**0.88**	188.89	154.82	**117.67**
000780	0.39	0.89	**0.32**	92.14	162.74	**75.11**
000785	0.86	1.36	**0.77**	150.11	255.93	**94.40**
000789	0.83	1.12	**0.77**	1172.88	785.29	**438.57**
000800	1.48	**0.59**	0.66	394.80	289.24	**89.59**
000809	1.50	1.39	**0.63**	/	/	/
000819	1.27	1.64	**0.95**	127.34	175.90	**78.21**
000822	1.21	1.66	**0.98**	94.66	136.28	**85.25**
000823	0.59	1.58	**0.58**	90.97	166.17	**54.63**
000830	1.74	1.12	**0.99**	181.87	129.81	**82.59**
000848	1.74	0.87	**0.40**	338.85	355.40	**181.45**
000869	0.76	1.67	**0.70**	127.95	218.37	**117.12**
000880	0.91	1.30	**0.78**	86.06	139.62	**68.44**
000886	0.96	1.61	**0.93**	238.55	244.64	**168.38**
000888	1.17	1.20	**0.42**	361.08	327.54	**198.10**
000895	1.03	0.82	**0.68**	247.78	**200.50**	251.69
000898	1.33	1.29	**0.59**	138.44	167.68	**84.55**
000915	0.63	1.62	**0.61**	161.90	136.88	**91.27**
000919	1.39	1.42	**0.65**	162.74	220.20	**118.16**
000921	1.67	1.23	**0.47**	197.27	158.23	**77.84**
000937	1.38	**0.57**	0.69	125.94	70.93	**69.52**
000951	0.55	1.58	**0.52**	**283.59**	285.33	367.56
000965	**0.18**	0.60	0.37	**409.38**	1259.03	578.53
000966	0.98	1.22	**0.40**	588.93	392.68	**297.43**
000983	1.23	1.43	**0.63**	304.30	271.71	**190.65**
000985	1.74	0.77	**0.52**	216.84	166.29	**100.87**
000988	**0.65**	0.74	0.67	96.33	**55.81**	97.98
均值	1.20	1.20	**0.61**	264.72	265.52	**139.72**
标准差	0.39	0.34	**0.25**	221.59	232.78	**105.48**

图7是将表2数据可视化后的结果,图7(a)表示基于机构交易行为下的趋势预测方法MPP与ARMA模型的*RMSE*值比较,其中纵轴表示ARMA方法的*RMSE*值且记为Ra,横轴表示MPP方法的*RMSE*值且记为Rm。由散点图易知只有3个点在下三角区域,即$Ra < Rm$,说明ARMA模型只有3支股票的预测结果优于MPP。相反,MPP方法在65支股票数据中取得比ARMA更好的预测结果。图7(b)表示MPP与ARMA模型的*MAPE*值比较,由于*MAPE*值是百分后的数值,为了使图像直观好看且坐标轴不用设置太大,故将*MAPE*值均除以100后再将其可视化,其中纵轴表示ARMA方法的*MAPE*值且记

为*Ma*，横轴表示MPP方法的*MAPE*值且记为*Mm*。图7(b)中共有8个点在下三角区域，即*Ma* < *Mm*，说明ARMA模型在8支股票数据中的MAPE指标优于MPP，而MPP在57支股票中取得比ARMA更好的预测结果。图7(c)和(d)分别表示了MPP与LSTM网络*RMSE*值和*MAPE*值的比较。图7(c)中表示通过RMSE评价指标得出MPP的趋势预测方法共有65支股票的预测结果优于LSTM网络。图7(d)中表示通过MAPE评价指标得出MPP的趋势预测方法共有59支股票的预测结果优于LSTM网络。

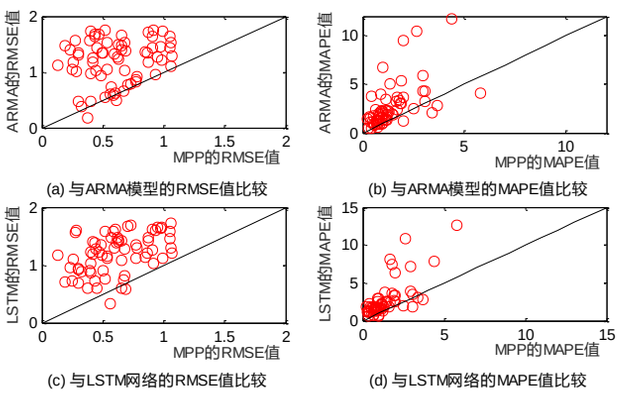

图7 三种方法对A时间段的预测结果比较

Fig. 7 Comparison of the prediction results in time A with three methods

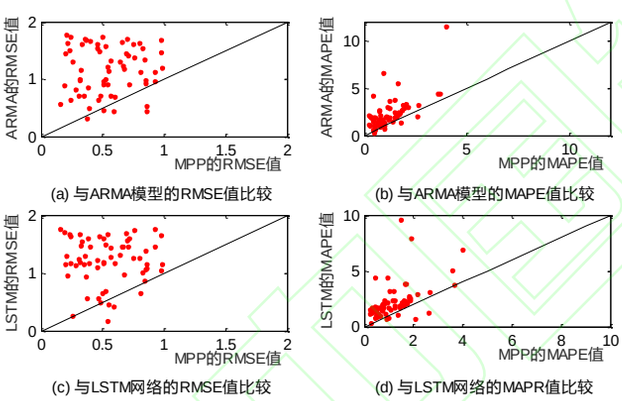

图8 三种方法对B时间段的预测结果比较

Fig. 8 Comparison of the prediction results in time B with three methods

为保证上述比较时间段不具有偶然性，另选取了B时间段(预测自2018年4月24日起未来5个交易日的股价波动)进行相同的实验步骤，同样使用RMSE和MAPE这两种评价指标评价MPP、ARMA、LSTM这三种方法拟合结果。如图8所示。

由于在预测过程中000014(沙河股份)、000151(中成股份)、000532(华金资本)、000789(万年青)、000819(岳阳兴长)、000830(鲁西化工)和000886(海南高速)这七支股票的匹配程度不够导致不可以预测，故最终进行预测结果比较的总共是63支股票。图8(a)中表示通过RMSE评价指标得出基于MPP方法获得的预测结果优于ARMA模型的股票共有58支；图8(b)中表示通过MAPE评价指标得出基于MPP方法获得的预测结果优于ARMA模型的共有59支。同理，图8(c)中MPP的趋势预测方法共有59支股票的预测结果优于LSTM网络，图8(d)中MPP的趋势预测方法共有58支股票的预测结果优于LSTM网络。

4 结论

根据深市A股股票的换手率数据，使用stomp算法获取具有机构交易行为的兴趣模式片段，构建完备知识库，进而提出基于知识库中兴趣模式的单支股票的金融股票价格趋势波动预测方法。针对某支股票，根据股票的历史换手率数据，收盘价数据以及待预测天数，筛选出基于历史数据具有最佳预测效果的兴趣模式长度，从而进行未来几天的股价趋势预测。在时间效率方面，由于前期要对70支股票数据使用Matrix Profile算法建立不同兴趣模式长度的知识库，数据量大，算法的时间复杂度也不低，且后期预测时需进行多次的模拟训练，故耗费时间较长。在应用方面，在已构建好知识库的情况下，对在构建知识库的过程中应用到的所有股票都可使用MPP方法进行股价趋势预测，说明该方法具有相对的普遍应用价值。新方法MPP与ARMA模型和LSTM网络的预测结果相比较，实验结果表明，基于矩阵画像的金融股价波动预测效果较好。本研究中获得的贡献性表现为：(1)使用了矩阵画像算法与股票预测相结合，利用矩阵画像算法，构建了基于机构交易行为下的知识库，并根据该知识库可对股票的未来趋势进行较为准确的预测。(2)将待预测股票的历史数据作为训练集，测试在确定预测时间内兴趣模式序列长度为何值时最佳，进一步优化了预测模型，提高了预测方法的拟合效果。另外，通过研究获得的信息和知识可以降低机构交易行为对散户的影响，帮助散户们在市场中获取较稳定的收益。同时，帮助金融市场监管部门对股价进行监控预测，防范可能出现的股价波动异常。此外，在确定兴趣模式的最佳长度时，主要通过进行多次模拟预测实验，取多次预测结果拟合值的最小均值所对应的兴趣模式长度。该过程并不能保证每次所取的兴趣模式长度是最佳的，故针对兴趣模式长度的分析是未来值得研究的问题。

参考文献 (References)

[1] 何佳，何基报，王霞，等. 机构投资者一定能够稳定股市吗？——来自中国的经验证据[J]. 管理世界，2007(8)：35-42. (HE J，HE J B，WANG X，et al. Can institutional investors be able to stabilize the stock market? ——Evidence from China [J]. Management World，2007，(8)：35-42.)

[2] 王咏梅，王亚平. 机构投资者如何影响市场的信息效率——来自中国的经验证据[J]. 金融研究，2011(10)：112-126. (WANG Y M，WANG Y P. How institutional investors affect the information efficiency of markets--evidence from china[J]. Financial Research，2011，(10)：112-126.)

[3] 刘京军，徐浩萍. 机构投资者：长期投资者还是短期机会主义者？[J]. 金融研究，2012(9)：141-154. (LIU J J，XU H P. Institutional investors：long-term investors or short-term opportunists?[J]. Financial Research，2012(9)：141-154.)

[4] 史永东，王谨乐. 中国机构投资者真的稳定市场了吗?[J]. 经济研究，2014，49(12)：100-112. (SHI Y D，WANG J L. Is Chinese institutional investors really stabilizing the market? [J]. Economic Research，2014，49(12)：100-112.)

[5] 王强，吕政，王霖青，等. 基于深度去噪核映射的长期预测模型[J]. 控制与决策，2019，34(5)：989-996. (WANG Q，LV Z，WANG L Q，et al. Deep denoising kernel mapping-based long-term prediction model[J]. Control & Decision，2019，34(5)：989-996.)

[6] 张贵生，张信东. 基于微分信息的ARMAD-GARCH股价预测模型[J]. 系统工程理论与实践，2016，36(5)：1136-1145. (ZHANG

G S, ZHANG X D. ARMAD-GARCH stock price prediction model based on differential information[J]. Systems Engineering--Theory & Practice, 2016, 36(5):1136-1145.)

[7] 吴少聪. 基于混合模型的股票趋势预测方法研究[D]. 哈尔滨:哈尔滨工业大学, 2017:61.(WU S C. Research on stock trend forecasting method based on hybrid model[D]. Harbin: Harbin Institute of Technology, 2017:61.)

[8] 宋刚,张云峰,包芳勋,等. 基于粒子群优化LSTM的股票预测模型[J]. 北京航空航天大学学报, 2019, 45(12): 2533-2542.(SONG G, ZHANG Y F, BAO F X, et al. Stock prediction model based on particle swarm optimization LSTM[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2533-2542.)

[9] 石浩. 基于递归神经网络的股票趋势预测研究[D]. 北京:北京邮电大学, 2018:45.(SHI H. Research on stock trend forecasting based on recurrent neural network [D]. Beijing: Beijing University of Posts and Telecommunications, 2018:45.)

[10] 谢琪,程耕国,徐旭. 基于神经网络集成学习股票预测模型的研究[J]. 计算机工程与应用, 2019,55(8):238-243.(XIE Q, CHENG G G, XU X. Study on stock prediction model based on neural network integrated learning[J]. Computer Engineering and Applications, 2019, 55(8):238-243.)

[11] NAKAGAWA K, IMAMURA M, YOSHIDA K. Stock price prediction using k-medoids clustering with indexing dynamic time warping[J]. Electronics and Communications in Japan, 2019, 102(2): 3-8.

[12] 李海林,梁叶,王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2018,33(8):1345-1353.(LI H L, LIANG Y, WANG S C. Review on dynamic time warping in time series data mining[J]. Control and Decision, 2018, 33(8):1345-1353.)

[13] CHEN Y, KEOGH E, HU B, et al. The UCR time series classification archivee [EB/OL]. (2019-07-01) [2019-09-01], http://www.cs.ucr.edu/eamonn/time(~series data/.)

[14] YEH C C M, ZHU Y, ULANOVA L, et al. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets[C]// Proceedings of the 2016 IEEE 16th International Conference on Data Mining. Piscataway: IEEE, 2016: 1317-1322.

[15] ZHU Y, ZIMMERMAN Z, SENOBARI N S, et al. Matrix profile ii: Exploiting a novel algorithm and GPUs to break the one hundred million barrier for time series motifs and joins [C]// Proceedings of the 2016 IEEE 16th International Conference on Data Mining. Piscataway: IEEE, 2016: 739-748.

[16] YEH C C M, KAVANTZAS N, KEOGH E. Matrix profile iv: using weakly labeled time series to predict outcomes [J]. Proceedings of the VLDB Endowment, 2017, 10(12):1802-1812.

[17] DAU H A, KEOGH E. Matrix profile V: A generic technique to incorporate domain knowledge into motif discovery [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 125-134.

[18] YEH C C M, KAVANTZAS N, KEOGH E. Matrix profile VI: Meaningful multidimensional motif discovery[C]// Proceedings of the 2017 IEEE International Conference on Data Mining. Piscataway: IEEE, 2017: 565-574.

[19] ZHU Y, IMAMURA M, NIKOVSKI D, et al. Matrix profile vii: time series chains: a new primitive for time series data mining [C]// Proceedings of the 2017 IEEE International Conference on Data Mining. Piscataway: IEEE, 2017: 695-704.

[20] ZHU Y, MUEEN A, KEOGH E. Matrix profile ix: admissible time series motif discovery with missing data [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(8): 1-11.

[21] ZHU Y, YEH C C M, ZIMMERMAN Z, et al. Matrix profile xi: scrimp++: time series motif discovery at interactive speeds[C]// Proceedings of the 2018 IEEE International Conference on Data Mining. Piscataway: IEEE, 2018:837-846.

[22] GHARGHABI S, IMANI S, BAGNALL A, et al. Matrix Profile xii: mpdist: a novel time series distance measure to allow data mining in more challenging scenarios[C]// Proceedings of the 2018 IEEE International Conference on Data Mining. Piscataway: IEEE, 2018: 965-970.

[23] IMANI S, MADRID F, DING W, et al. Matrix profile xiii: time series snippets: a new primitive for time series data mining[C]// Proceedings of the 2018 IEEE International Conference on Big Knowledge. Piscataway: IEEE, 2018: 382-389.

[24] YEH C C M, ZHU Y, ULANOVA L, et al. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile[J]. Data Mining and Knowledge Discovery, 2018, 32(1): 83-123.

This work is partially supported by the National Natural Science Foundation of China (71771094), Project of Science and Technology Plan of Fujian Province of China (2019J01067).

GAO Shile, born in 1972, Ph. D., Assistant. His research interests include data mining, complex network.

WANG Ying, born in 1997, Undergraduate, Her research interests include time series data mining.

LI Hailin, born in 1982, Ph. D., Professor. His research interests include data mining, business analysis;

WAN Xiaoji, born in 1984, Ph. D., Assistant. His research interests include data mining, financial analysis.