Advances in Applied Mathematics 应用数学进展, 2024, 13(4), 1599-1606
Published Online April 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.134150

Hans汉斯

基于隐马尔可夫模型的股票价格预测

张　政，李俊刚，李　鑫，王　然

北方工业大学理学院，北京

收稿日期：2024年3月25日；录用日期：2024年4月22日；发布日期：2024年4月29日

摘　要

本文构建隐马尔可夫模型预测比亚迪公司股票收盘价，采用K均值聚类法和AIC、BIC准则确定隐状态个数，运用EM算法进行模型参数估计，并将MSE、MAE和R^2作为评价指标评估准确性，结果显示基于模型预测结果较为准确稳定。研究结果表明HMM模型能捕捉市场因素、公司财务状况和行业趋势对价格的影响，为投资者和分析师提供深入市场洞察。本研究提供了有效的股票预测模型，同时探索了HMM模型在股票价格预测中的应用，为金融时间序列预测方法的改进和发展提供新思路和方法。

关键词

HMM，K均值聚类，股价预测

Stock Price Prediction Based on Hidden Markov Model

Zheng Zhang, Jungang Li, Xin Li, Ran Wang

College of Science, North China University of Technology, Beijing

Received: Mar. 25th, 2024; accepted: Apr. 22nd, 2024; published: Apr. 29th, 2024

Abstract

In this paper, a hidden Markov model is constructed to predict the closing price of BYD Company's stock, the number of hidden states is determined by K-means clustering method, AIC and BIC criteria, and the model parameters are estimated by EM algorithm. MSE, MAE and R^2 are used as evaluation indicators to evaluate the accuracy. The results show that the prediction results based on the model are more accurate and stable. The results show that the HMM model can capture the

文章引用: 张政, 李俊刚, 李鑫, 王然. 基于隐马尔可夫模型的股票价格预测[J]. 应用数学进展, 2024, 13(4): 1599-1606.
DOI: 10.12677/aam.2024.134150

impact of market factors, company financial conditions and industry trends on prices, providing investors and analysts with in-depth market insights. This study provides an effective stock prediction model, and explores the application of HMM model in stock price prediction, which provides new ideas and methods for the improvement and development of financial time series prediction methods.

Keywords

HMM, K-Means Clustering, Stock Price Prediction

1. 引言

1.1. 研究背景

随着金融市场的发展和投资者对于股票市场的关注度不断增加，预测股票价格成为了金融领域的一个重要问题。股票价格的预测对于投资决策、风险管理和资产配置具有重要意义。然而，股票价格的变化受到众多因素的影响，包括市场需求、公司财务状况、行业发展趋势等，这使得股票价格的预测变得复杂而具有挑战性。

在这个背景下，本论文旨在构建一个基于隐马尔可夫模型的预测框架，以预测股票收盘价。比亚迪作为一家在新能源汽车领域具有重要地位的公司，其股票价格波动受到市场、行业和政策等因素的共同影响[1]。本文以比亚迪为例，对其股票收盘价进行预测。

HMM 模型作为一种统计模型，在序列数据建模和预测中具有广泛应用。通过建立 HMM 模型，我们可以考虑比亚迪公司收盘价之间的隐藏关系和观测序列的变化规律，从而捕捉到价格波动的潜在模式。HMM 模型结合了状态转移概率、观测概率和初始状态概率，能够在给定观测序列的情况下，推断最可能的隐藏状态序列，从而实现对收盘价的预测。

通过本论文的研究，我们希望能够提供一种有效的预测框架，为投资者和分析师提供比亚迪公司收盘价的预测结果，并为他们的投资决策提供参考。此外，通过探索 HMM 模型在股票价格预测中的应用，我们也可以对金融时间序列数据的建模和预测方法进行进一步探索和改进，为金融市场的研究和实践做出贡献。

1.2. 研究意义

目前已经有许多研究探索了股票价格预测的方法和技术。传统的时间序列分析方法，如 ARIMA 模型和 GARCH 模型，被广泛应用于股票价格预测。然而，这些方法往往无法充分捕捉到非线性和动态变化的特征，限制了它们在股票市场中的应用。

在金融领域，HMM 模型已经被广泛研究和应用于时间序列数据的建模和预测。HMM 模型具有良好的灵活性和表达能力，能够捕捉到数据背后的潜在模式和动态特征。在股票价格预测领域，HMM 模型已经被用于建模价格波动、市场情绪和交易行为等因素，提高了预测的准确性和可靠性。

本研究的成果将为金融领域的时间序列预测方法提供新的思路和方法。通过应用 HMM 模型于股票价格预测，我们可以拓展和改进金融数据建模和预测的方法，为金融市场的研究和实践做出贡献。

2. 基于 HMM 模型的股票价格预测

2.1. HMM 模型

隐马尔可夫模型(Hidden Markov Model, HMM)是一种统计模型，用于描述具有潜在隐含状态的随机过程，由一系列离散的隐含状态和与之关联的观测序列组成，在语音识别、自然语言处理、金融市场分析等领域中被广泛应用。HMM 模型可以计算概率、解码最可能的隐含状态序列，并通过学习问题估计模型参数[2]。

连续型 HMM 模型可以被以下 5 个参数描述：

1) N 是模型隐藏状态的个数，将每个状态记录为 $S=\{s_1,s_2,\cdots,s_t,\cdots,s_N\}$，其中 s_t 表示 t 时刻的状态。

2) M 是每个隐藏状态中混合高斯概率密度函数。

3) A 是状态转移概率矩阵。$A=(a_{ij})$，其中 a_{ij} 为隐马尔可夫链中从当前状态 i 转移到下一状态 j 的概率，即 $a_{ij}=P[q_{t+1j}=j\,|\,q_t=i]$，其中 q_t 为 t 时刻所在的状态，$a_{ij}\geq 0$，$\sum a_{ij}=1$，$1\leq i,j\leq N$。

4) B 是观测概率分布矩阵。$B=\{b_j(o)\}$，其中 $b_j(o)$ 是状态 j 的随机观察值输出概率函数。

5) Π 是初始状态分布矩阵，$\Pi=\{\pi_i\}$，其中 $\pi_i=P[q_1=i]$，$\sum\pi_i=1$，$1\leq i\leq N$。

HMM 模型可表示为：$\lambda=(A,B,\Pi)$。

此外，HMM 模型还有两个基本假设[3]：

1) 齐次马尔可夫假设：任意时刻的状态只依赖于前一时刻的状态。与其他时刻的状态及观测无关。

2) 观测独立性假设：任意时刻的观测只依赖于该时刻的状态。与其他时刻的观测及状态无关。

2.2. 数据收集与处理

比亚迪在全球范围内具有广泛关注度和市场影响力，有大量的历史股票数据可供分析和建模，且其所处的汽车行业是一个重要的经济领域，通过对其股票数据进行分析，可以深入了解汽车行业的市场动态和股票价格的波动情况。故本文选取比亚迪(002594)股票为研究对象，并收集该股票 2022 年 1 月 4 日至 2023 年 9 月 28 日收盘价 X_1、当日最高价 X_2、当日最低价 X_3、成交量 X_4 等交易数据。该股票数据如图 1 所示。

对收集数据进行清洗，缺失值采用线性插值进行填充。线性插值方法可以使填充后的数据在整体上保持了原有的变化趋势。这种方法假设数据的变化是线性的，因此适用于股票数据中存在较为连续的变化模式。相比于简单地使用均值或中位数填充，线性插值方法可以提供更准确的填充结果，特别是当数据存在较大波动或趋势变化时。通过采用线性插值方法填充股票数据的缺失值，可以更好地保持数据的完整性和连续性，从而提高后续的数据分析和预测的准确性。线性插值公式为：

$$y=y_1+\frac{y_2-y_1}{t_2-t_1}*(t-t_1)$$

其中缺失值所在的时间点为 t，y 为需要填充的缺失值，其相邻数据点为 (t_1,y_1) 和 (t_2,y_2)，且 $t_1<t<t_2$。

为了对股票数据进行计算和转化，并确定隐藏状态，我们选取以下四个变量：

2.2.1. 收盘价移动平均值

收盘价的 10 日移动平均值能够平滑价格波动，捕捉价格的长期趋势，同时保留了一定的短期变动信息。选取收盘价的 10 日移动平均值作为研究变量，它可以提供有关价格走势的重要特征，帮助识别不同的市场状态或价格模式[4]。收盘价的计算公式为：

$$F_t=(X_{1t-1}+X_{1t-2}+\cdots+X_{1t-n})/n$$

其中 F_t 为收盘价的移动平均值，n 在此模型中为 10。

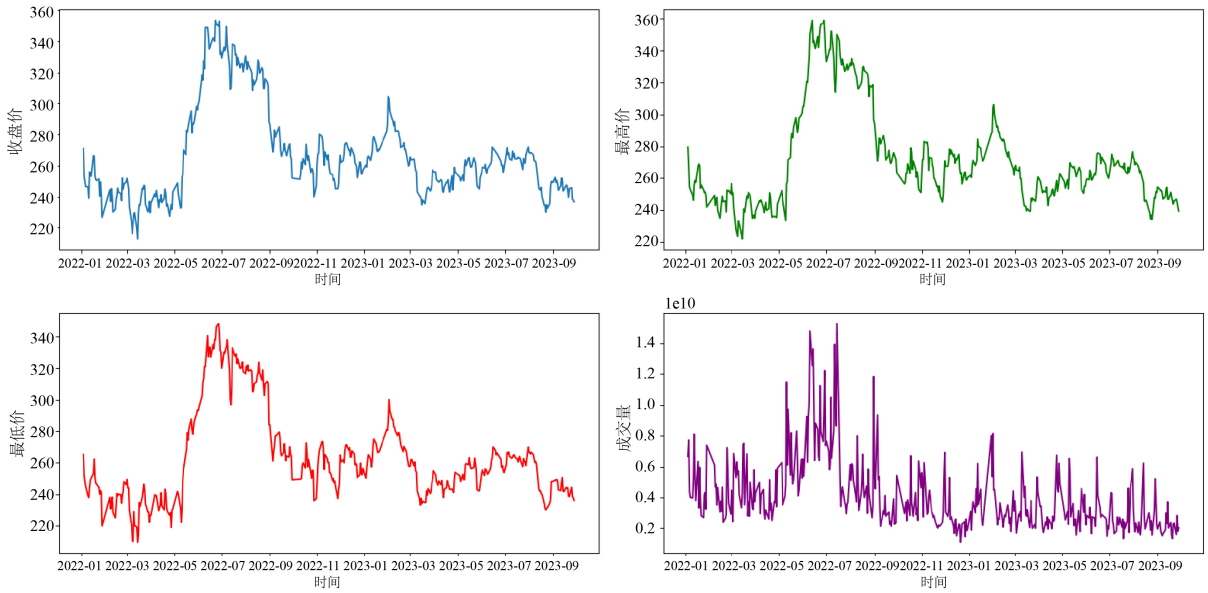

Figure 1. BYD stock data from January 22 to September 23
图 1. 比亚迪 22 年 1 月~23 年 9 月股票数据

2.2.2. 日收益率

计算日收益率，即每日收盘价相对于前一天的变化百分比。日收益率能够捕捉价格的相对变动，消除价格水平的影响，更准确地反映市场的相对涨跌情况。此外，日收益率通常具有良好的时间可比性，使得模型可以更好地处理不同时间段的数据。日收益率的计算公式为：

$$r_t = \left(X_{1t} - X_{1t-1}\right) / X_{1t-1}$$

其中 r_t 为日收益率。

2.2.3. 收盘价位置

收盘价位置，即收盘价相对于当日最高价和最低价的位置，可用于评估股票价格的相对强弱或趋势的倾向，帮助识别不同的市场阶段和价格模式。收盘价位置的计算公式为：

$$C_t = \left(X_{1t} - X_{3t}\right) / \left(X_{2t} - X_{3t}\right)$$

其中 C_t 为收盘价位置。

2.2.4. 成交量变化率

成交量是衡量市场交易活跃程度的指标，能够提供关于市场参与度、交易活跃度等重要信息。用于衡量市场交易的活跃程度和资金流动的变化。成交量变化率的计算公式为：

$$v_t = \left(X_{4t} - X_{4t-1}\right) / X_{4t-1}$$

其中 v_t 为成交量变化率。

2.3. 确定隐状态个数

为确保模型在训练集上学习到历史模式和规律，并在独立的测试集上进行评估和验证模型的预测能力，本文按照时间顺序将数据划分为训练集和测试集，2023 年 5 月 1 日之前的数据为训练集，之后的数据为测试集。

HMM 模型涉及计算概率转移矩阵和观测概率矩阵，需要考虑多个可能的隐藏状态和观测状态。为简化模型，提高计算效率，本文采用 K 均值聚类将连续观测序列转换为离散隐藏状态序列，减少模型中需要建模的参数数量。

本文使用训练集的观测变量进行 K 均值聚类。K 均值聚类算法是一种迭代性聚类算法，对一个 n 维向量的数据点集 $D=\{x_i \mid i=1,\cdots,N\}$ 进行聚类，其中 x_i 表示第 i 个数据点，最终将集合 D 划分维 k 个类簇。组内对象越相似、组间差距越大越好。通常以欧式距离作为相似度度量，以误差平方和(SSE)作为度量聚类效果的目标函数，通过最小化目标函数，按照数据点到聚类中心的远近划分为 k 个簇[5]。

数据点与聚类中心的欧式距离公式为：

$$d(x,C_i)=\sqrt{\sum_{j=1}^{m}\left(x_j-C_{ij}\right)^2}$$

其中，x 为数据点，C_i 为第 i 个聚类中心，m 为数据对象的维度，x_j 为 x 的第 j 个值，C_{ij} 为 C_i 的第 j 个值。

SSE 计算公式为：

$$\mathrm{SSE}=\sum_{i=1}^{k}\sum_{x\in C_i}\left|d(x,C_i)\right|^2$$

在 python 中编写代码，对训练集数据进行 K 均值聚类，将每个观测数据点分配到最接近的聚类簇，对应的聚类簇标签作为隐状态，得到不同隐状态数的初始分组。针对每个分组，通过 HMM 模型的学习和参数估计，利用 AIC 准则和 BIC 准则对不同隐状态数模型的复杂度进行评估和比较。最终，选择 AIC 值和 BIC 值最小的模型对应的隐状态数作为最优的 HMM 模型隐状态数。在 python 上编写代码进行计算，得到不同隐状态数下模型的 AIC 值和 BIC 值，数据如表 1 所示。

Table 1. Comparison of model AIC value and BIC value under different number of hidden states
表 1. 不同隐状态数下模型 AIC 值和 BIC 值比较

隐状态数	AIC	BIC
2	3520.078492	3588.516622
3	3386.887549	3500.951100
4	3180.835413	3348.128622
5	3290.194093	3518.321195
6	2966.871623	3274.436856
7	2977.907064	3339.514665
8	2997.399509	3453.653714

由表中数据可以得出，当隐状态数为 6 时，AIC 值和 BIC 值均最小，故确定模型的隐状态数为 6，即 K 均值聚类中心个数应为 6。K 均值聚类结果如表 2 所示：

Table 2. Clustering results
表 2. 聚类结果

聚类中心	1	2	3	4	5	6
样本数量	50	38	88	16	112	27

可以看到，聚类中心 5 是最大的簇，包含 112 个数据点，代表了该股票价格的普遍情况；而聚类中

心 4 仅有 16 个数据点，代表着该股票的一些特殊或罕见情况。从聚类结果的簇大小平衡性来看，该聚类结果存在一定的不平衡性，但考虑到股票价格的实际情况，认为此次聚类结果合理，可以进行下一步的参数估计。

2.4. 参数估计

确定隐状态后，计算每个隐状态的频率，即为初始状态概率 Π 的估计值，得到结果如下：

$$\Pi = (0.151, 0.115, 0.266, 0.048, 0.338, 0.082)$$

之后采用期望最大化算法(Expectation-Maximization algorithm，EM 算法)进行状态转移概率和观测概率分布的估计。EM 算法是一种迭代优化算法，用于在存在隐变量或缺失数据的情况下，通过最大似然估计来估计模型的参数[3]。

在此模型中，E 步(Expectation step)使用前向－后向算法计算每个时间步处于每个隐藏状态的后验概率，后验概率表示在给定观测序列下，每个时间步处于每个隐藏状态的概率。M 步(Maximization step)根据 E 步中计算得到的后验概率，更新 HMM 模型的参数，将第一个时间步的后验概率作为新的初始状态概率向量，每个时间步的后验概率作为计算新的状态转移概率矩阵，使用后验概率计算新的观测概率矩阵，根据更新后的参数计算新的模型的对数似然函数值。重复 E 步和 M 步，使其逼近最大似然估计，迭代 100 次停止。

在 python 中编写代码进行模型的训练，将训练集数据带入，参数估计结果如下。

状态转移概率矩阵 A 的估计值为：

$$A = \begin{pmatrix} 0.152 & 0.000 & 0.000 & 0.000 & 0.750 & 0.098 \\ 0.431 & 0.000 & 0.135 & 0.000 & 0.434 & 0.000 \\ 0.091 & 0.837 & 0.072 & 0.000 & 0.000 & 0.000 \\ 0.007 & 0.000 & 0.000 & 0.272 & 0.000 & 0.721 \\ 0.092 & 0.407 & 0.192 & 0.000 & 0.309 & 0.000 \\ 0.000 & 0.000 & 0.010 & 0.696 & 0.000 & 0.294 \end{pmatrix}$$

观测概率分布 B 的估计值如表 3 所示。

Table 3. Observational probability distribution
表 3. 观测概率分布

隐状态	估计值
1	$\begin{pmatrix} 1.270 & -1.252 & -0.879 & 0.768 \\ 0.553 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.502 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.384 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.784 \end{pmatrix}$
2	$\begin{pmatrix} 1.237 & -0.383 & -0.423 & -0.657 \\ 0.461 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.075 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.193 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.784 \end{pmatrix}$

续表

3	$\begin{pmatrix} 1.278 & 1.680 & 1.231 & 0.790 \\ 0.485 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.435 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.342 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.324 \end{pmatrix}$
4	$\begin{pmatrix} -0.596 & 0.321 & 0.214 & 0.722 \\ 0.156 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.598 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.157 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.033 \end{pmatrix}$
5	$\begin{pmatrix} 1.348 & 0.161 & 0.376 & -0.313 \\ 0.505 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.176 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.581 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.242 \end{pmatrix}$
6	$\begin{pmatrix} -0.596 & -0.246 & -0.220 & -0.660 \\ 0.148 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.271 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.870 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.132 \end{pmatrix}$

2.5. 模型结果及评价

在训练好 HMM 模型后，将测试集输入模型预测股票收盘价，并与实际收盘价进行对比，如图 2 所示：

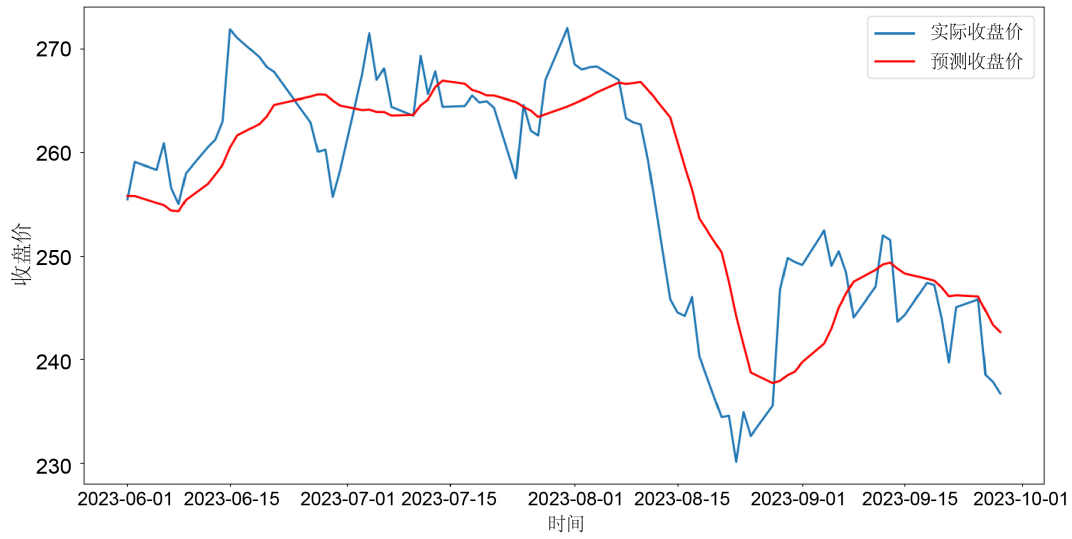

Figure 2. Comparison between the predicted value and the actual value of the stock closing price
图 2. 股票收盘价预测值与实际值比较

由图 2 可以看出，测试集股票收盘价的预测值与实际值之间显示出较为良好的一致性和准确性。预测模型成功地捕捉到了股票收盘价的趋势和波动，并且在大部分情况下，预测值与实际值较为接近。

为量化评价模型的预测效果，选取均方误差(MSE)、平均绝对误差(MAE)和拟合优度(R^2)作为评价指标。在 python 上编写相关代码，计算指标值，结果为表 4。

Table 4. Model fitting effect
表 4. 模型拟合效果

指标	MSE	MAE	R^2
数值	42.641	5.032	0.663

MSE 约为 42.641，数值较低，说明此模型的预测值与实际观测值之间的平均偏差较小，说明模型能够较准确地捕捉到收盘价的变化趋势和波动性；MAE 约为 5.032，也相对较小，说明模型对于异常值的敏感度较低，整体上能够提供较为准确的预测结果[6]；R^2 为 0.663，表明模型能够解释约 66.3%的收盘价方差，说明模型对观测数据的拟合较好，能够较好地解释收盘价的变异性。

综上所述，该模型在预测股票收盘价上表现出较高的准确性和解释能力，预测结果可为投资者提供有价值的信息。

3. 结论

本文对比亚迪公司历史股票数据进行分析，提取了收盘价移动平均值、日收益率、收盘价位置和成交量变化率这四个变量，采用 K 均值聚类法和 AIC、BIC 准则结合的方式确定模型隐状态个数，运用 EM 算法估计模型参数，成功地训练了一个 HMM 模型，并利用该模型进行了测试集收盘价的预测。

将预测结果与真实股票收盘价数据进行比对，发现模型预测结果能捕捉到价格变动的趋势，采用 MSE、MAE 和 R^2 作为评价指标进行量化分析，结果显示预测结果对异常值的敏感度较低，较为准确稳定。

总的来说，基于隐马尔可夫模型的股票收盘价预测结果显示出良好的性能，具有实际应用的潜力。HMM 在预测金融时间序列方面具有潜力，未来的研究可以进一步探索、改进和优化 HMM 模型，在更广泛的金融市场和股票预测领域中创造价值。

基金项目

本论文工作由大学生创新创业项目和北京市属高校基本科研业务费(No. 11005297192/103)资助。

参考文献

[1] 方诚铭. 基于市盈率与剩余收益的比亚迪估值分析[J]. 全国流通经济, 2023(1): 92-95.

[2] 李方圆, 张涛. 基于 HMM-XGBoost 的股价预测[J]. 桂林航天工业学院学报, 2021, 26(4): 484-488.

[3] 冷寒雨, 王胜烽, 肖井华. 基于 EM 算法对报警时间序列的分析预测[J]. 中国高新科技, 2022(12): 98-101.

[4] 富瑶, 王立柱. 基于移动平均线的股票买入时机算法[J]. 牡丹江师范学院学报(自然科学版), 2022(1): 6-8+35.

[5] 王森, 刘琛, 邢帅杰. K-Means 聚类算法研究综述[J]. 华东交通大学学报, 2022, 39(5): 119-126.

[6] 余文利, 廖建平, 马文龙. 一种新的基于隐马尔可夫模型的股票价格时间序列预测方法[J]. 计算机应用与软件, 2010, 27(6): 186-190.