使用隐马尔可夫模型预测达卡证券交易所股票价格

Raihan Tanvir*§、Md Tanvir Rouf Shawon*§、Md. Golam Rabiul Alam†*阿
赫桑努拉科技大学，达卡，孟加拉国†BRAC大学，达卡，孟加拉国{
raihantanvir.96、shawontanvir95}@gmail.com、r
abiul.alam@bracu.ac.bd

arXiv:2302.08911v1 [q-fin.ST] 26 Jan 2023

摘要——**股票市场预测是一项经典问题，已广泛运用机器学习和基于人工神经网络的工具与技术进行研究。该问题的有趣之处包括其对时间的依赖性、波动性以及其他复杂关系。为综合考虑这些因素，研究人员运用隐马尔可夫模型（HMM）预测股票价格。我们展示了最大后验（MAP）HMM 方法，该方法根据历史数据预测次日股票价格。通过分析股价的分数变化以及日内最高价和最低价，对 HMM 进行训练。随后，利用该模型计算次日所有可能股票价格的最大后验估计。我们的研究所展示的方法具有较强的通用性，只要使用某家公司股票的数据集训练 HMM，即可用于预测该公司的股价。我们采用一些广泛使用的回归问题准确度指标评估模型的准确度，结果令人满意。**

索引词——**隐马尔可夫模型、股票价格预测、时间序列分析、价格预测。**

一、引言

股票价格预测一直是人工智能界最棘手的问题之一。典型的人工智能研究主要致力于构建旨在模仿人类智能的解决方案，通常已超越预测研究的范畴。然而，由于股票价格预测具有非平稳、周期性和随机性特征，其研究仍受到严重限制。此类序列的价格波动受多种因素影响，包括股权、利率、期权、证券、认股权证，以及大型金融机构的兼并与收购等。在这样的市场中，普通投资者无法持续获利。因此，股票市场对智能预测模型的需求很高，普通投资者也对此有浓厚兴趣。

隐马尔可夫模型在分析和预测随时间变化的事件方面被认为十分有效。这项技术已应用于语音识别 [1], 手写识别 [2], 面部表情识别 [3], 心电图分析 [4], 等领域。就其与时间的内在联系而言，股票市场预测与这些问题类似。隐马尔可夫模型基于一组不可见的状态进行建模，这些状态之间可以发生状态转移，并将每种状态与可能的观测值相关联。类似地，股票市场也可以被观察。投资者通常不了解影响股价的内在因素。这些潜在状态之间的转移会受到商业策略、决策、市场环境等因素的影响。股票的价值是反映这些因素的可观测结果。因此，隐马尔可夫模型显然符合这一现实情境。

特征的选择对该方法至关重要。过去已有多种尝试采用交易量、股票动量，以及市场相关性和波动性。按照 [5] 中的建议，我们将股票价值的每日分数差异以及日内最高价和最低价的分数偏差纳入考量。要生成所需的预测，理解分数变化至关重要。日内最高价与最低价之间的分数差是预测波动性方向的良好指标。

尽管隐马尔可夫模型早已应用于这一领域，但此前没有任何研究涉及孟加拉国股票市场。因此，我们选择采用基于隐马尔可夫模型的方法来实现这一目标。为应用该技术，我们使用在达卡证券交易所（达卡证券交易所）上市公司的股票数据。针对每只股票，分别训练一个独特的隐马尔可夫模型。训练数据集唯一需要满足的条件是观测值具有显著的变异性。通过有效利用长达13年的时间跨度，这一条件得以满足；在此期间，股票价格持续波动，且波动幅度显著。

II. 先前研究

近年来，人们开展了各种研究，试图建立一个完美（或近乎完美）的股票市场预测模型。在大多数预测研究中，研究人员采用统计时间序列分析方法，例如自回归移动平均（ARMA） [6]和多元回归方法。

本文深入研究了一种基于隐马尔可夫模型的股票价格预测方法 [5] 。为了根据历史数据预测次日的股票价值，作者采用了最大后验隐马尔可夫模型方法。通过股票价值的分数差以及股票的日内最高价和最低价来训练连续隐马尔可夫模型。经过深思熟虑，我们决定采用本文所述的技术来开发我们的模型。

§作者贡献相同

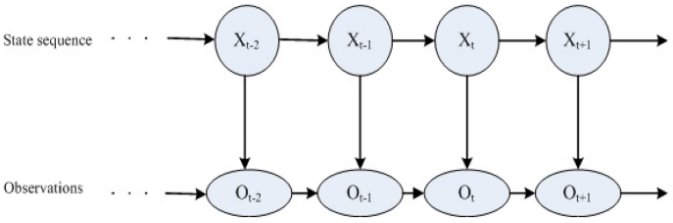

图 1. 隐马尔可夫模型

在 [7], 中，介绍了一种有向加权分块支持向量机方法，该方法将整个训练数据集划分为多个部分，并为每个部分生成支持向量。为了创建预测模型，我们在新的工作数据集上计算加权支持向量回归。

Md. Rafiul Hassan 等人提出了 [8], 一种基于隐马尔可夫模型的模型，通过对数据集中的相邻值进行插值来生成预测。该实验取得的结果令人鼓舞，并探索了一种用于股市分析的新框架。

Romahi Y 等人利用自动生成的模糊连接，提出了一种用于动态金融预测的独特技术 [9]。该方法取得了令人鼓舞的结果，但构建模糊系统需要领域专业知识。

H.Liu 等人提出了一种基于深度残差网络的预测策略，并将股价图作为输入 [10]。该模型的平均准确率为 0.40，高于随机指标的平均准确率 0.33。

此外，我们还深入研究了*hmmlearn* 软件包官方文档中介绍的一种朴素方法。在该方案中，仅考虑股票连续两个收盘价之间的差值。尽管这种方法简单，且忽略了股价预测中的若干关键因素，但它仍展现出相当令人满意的结果。

III. 方法论

隐马尔可夫模型 λ 可表示为:
$\lambda = (\pi, A, B)$,
其中， A 是转移矩阵，其各项表示某个状态转移到另一状态的似然； B 是发射矩阵，用于提供处于状态 j 时观测到 O_t 的概率 $b_j(O_t)$； π 表示时间 $t = 1$各状态的初始概率。我们认为发射概率分布是连续的，因为样本是由连续随机变量构成的向量。为简单起见，我们将其视为具有参数（μ和 Σ）的多元高斯分布。

模型的观测值是一个三维向量，表示每日的股票信息，

$$O_t = \left(\frac{\text{close} - \text{open}}{\text{open}}, \frac{\text{high} - \text{open}}{\text{open}}, \frac{\text{open} - \text{low}}{\text{open}} \right) := \left(frac_{Change}, frac_{high}, frac_{Low} \right) \quad (1)$$

在（1）中，open 表示当日的开盘价，close 表示当日的收盘价，high 和 low 分别表示当日的最高价和最低价。为了描述股票价值随时间保持恒定的方差，我们使用分数变化。

训练后，采用最大后验（MAP）方法的近似形式对其进行测试。在预测未来股价时，我们假设延迟 d 天。因此，给定隐马尔可夫模型 λ 、前d天的股票价格 $(O_1, O_2, \ldots, O_d)$，以及第$(d+1)^{st}$天的开盘价，任务是计算第 $(d+1)^{st}$ 天的收盘价，这等价于计算第 $(d+1)^{st}$ 天的分数差 $\frac{close-open}{open}$。这通过对观测向量 O_{d+1}进行最大后验概率近似来完成。

ˆ让我们考虑 O_{d+1}，即在给定前 d 天数值的情况下，第 $d+1$ 天观测值的最大后验值。

$$\hat{O}_{d+1} = \arg\max_{o_{d+1}} P(O_{d+1} \mid O_1, O_2, \ldots, O_d, \lambda) \quad (2)$$

$$= \arg\max_{o_{d+1}} \frac{P(O_1, O_2, \ldots, O_d, O_{d+1} \mid \lambda)}{P(O_1, O_2, \ldots, O_d, \lambda)} \quad (3)$$

观测向量 (O_{d+1}) 会在整个潜在值范围内进行调整。由于分母相对于 (O_{d+1}) 保持不变，最大后验概率近似可简化为（4）。

$$\hat{O}_{d+1} = \arg\max_{o_{d+1}} P(O_1, O_2, \ldots, O_d, O_{d+1} \mid \lambda) \quad (4)$$

我们通过计算一组不同潜在 O_{d+1} 值的概率来确定最大概率。确定特定观测值的似然所需的计算成本为 $O(n^2 d)$, 其中 n 表示状态数量， d 表示延迟。对于 O_{d+1}的所有不同潜在值集合，都会重复此过程。我们有$n = 4, d = 30$，并且 O_{d+1}的可能值数量为 $50 \times 10 \times 10$ （见表II）。给定某一天的收盘价可以通过取该天的开盘价，并将其与该天预期的分数差相结合来确定，如（5）所示。

$$closing_price = opening_price \times (1 + frac_{change}) \quad (5)$$

我们还采用了一种朴素方法，只考虑股票连续两日收盘价之间的差值。该方法中的观测值为连续两日的收盘价差值和股票成交量。模型训练完成后，通过计算转移矩阵与观测均值的内积进行预测。最后，将预测的变化量与前一日的收盘价相结合，得出最终预测。

IV. 数据集

为了实践所选技术，我们从Kaggle选择了DSEBD[1] 数据集。该数据集包含在达卡证券交易所注册的公司的年度股票数据

[1]https://www.kaggle.com/mahmudulhaque/dsebd

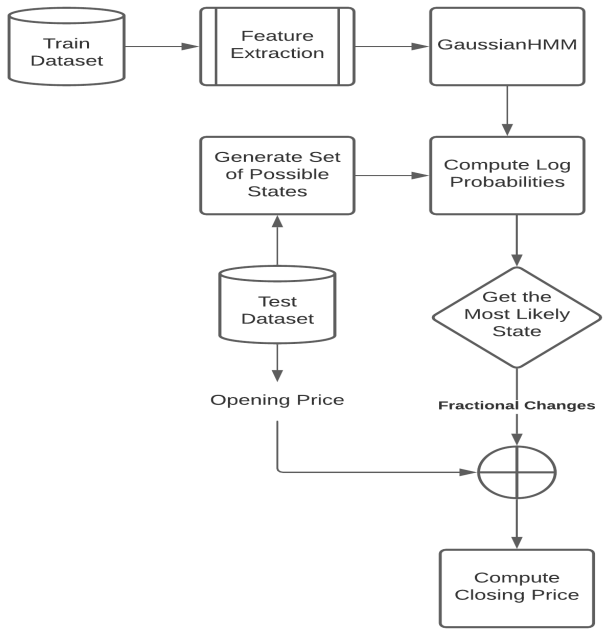

图 2. 使用分数变化预测收盘价

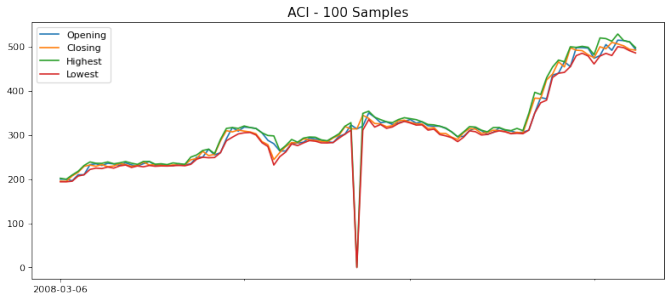

图 3. ACI股票开盘价、收盘价、最高价和最低价的趋势（100 个样本）

（DSE），涵盖2008年1月至2020年12月，并分别存储在JSON文件中。该数据集包含来自22个不同行业的589家公司的股票记录，涉及5种不同的金融工具类型。数据集中的观测值数量为15,75,134。

为将数据集纳入我们的方案，我们执行了以下处理步骤。

- 将 JSON 文件加载到 pandas数据框中，并存储到数据框列表中。将列表中的各项合并为一个数据框。将合并后的数据框保存为 CSV 文件。。通过从合并后的数据框中提取各个个股的数据，创建单独的数据集。

使用 sklearn.model_selection模块提供的*train_test_split*函数，将数据集按 0.8 : 0.2 的比例进行划分，其中 80% 的数据用于训练集，其余 20% 用于测试集。由于我们处理的是时间序列数据，因此需要保留数据的时间顺序。为避免



图 4. ACI股票从 2008 年 1 月至 2020 年 12 月的收盘价趋势。

表 I ACI股票数据集中的样本数据

date	open	high	low	收盘价	成交量	前收盘价
2008年3月6日	200.0	202.0	194.0	195.5	266850	198.8
2008年3月9日	199.8	199.8	194.0	195.0	333600	195.5
2008年3月9日	199.8	199.8	194.0	195.0	333600	195.5
2008年3月10日	196.5	209.5	195.4	207.3	381650	195.0
2008年3月11日	209.9	217.9	207.0	215.5	509550	207.3

将数据随机划分为训练集和测试集，我们将*shuffle=False*作为参数传入。

ACI 有限公司的股价趋势如图3和图4所示。表I展示了**ACI**股票数据集中的一些样本数据。

五、实现

作为我们的隐马尔可夫模型，我们采用了hmmlearn[2] 包中的GaussianHMM类，并使用其提供的*fit* 方法进行参数估计。由于我们研究了两种方法，下面将介绍第一种技术的具体实现，该技术使用股价的分数变化。随后，我们将讨论另一种策略的实现，该策略考虑股票价格的连续波动。

A. 基于分数变化的隐马尔可夫模型

1) 初始化*::* 根据以下配置设置参数，以完成隐马尔可夫模型的初始化：

- 隐状态数量，$n = 4$。观测值维度，$D = 3$
- 延迟，$d = 30$ 天。最大迭代次数，$n\ iter = 10000$。收敛阈值 $tol = 0.001$

这些值取自 [5]，，我们将在此基础上构建模型。此外，[11]建议使用 4 个底层状态，因为观测值的维度同样为 4。其余模型参数采用*hmmlearn* 包提供的GaussianHMM类默认值进行初始化。

[2]https://hmmlearn.readthedocs.io/en/latest/

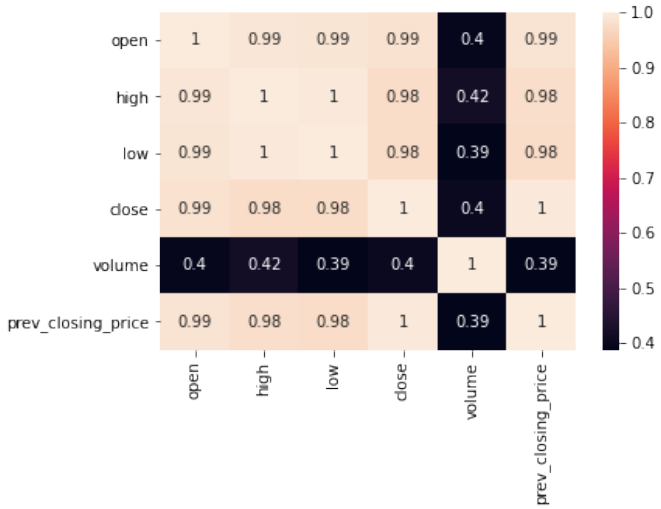

图5. ACI股票数据的相关矩阵

2) 训练： 尽管我们的数据集包含许多特征，但并非所有特征都对股票的收盘价有显著影响。从图 5 的相关矩阵中可以看出，开盘价、最高价和最低价与收盘价呈强正相关关系（恰好为 1 或非常接近 1）。因此，我们将这四个特征视为感兴趣的属性。

现在，我们每天拥有的特征相对较少，主要是当天股票的开盘价和收盘价，以及股票的最高价和最低价。因此，我们利用它们来计算股票价格。我们并非直接使用这些数值，而是提取了它们各自的分数差异，用于训练我们的隐马尔可夫模型。

在特征提取之后，某些行可能包含NaN、无穷大，或者对于dtype('float64')而言过大的数值。通过数据清洗程序移除这类数据。完成上述两项操作后，使用获得的特征向量，通过*fit* 方法对**GaussianHMM** 类进行隐马尔可夫模型训练。尽管$n\ \ iter$被设为10000，但由于收敛阈值 tol，训练在35次迭代后提前停止。

3）预测： 模型训练完成后，我们便估计股票的收盘价。给定某一天的股票开盘价以及前 d 天的数据，我们可以计算该日的股票收盘价。我们的预测器将具有 d 天的延迟。这意味着，如果我们能够估计某一特定日期的 $frac_{change}$，就可以利用公式（5）计算收盘价。

如果将 $frac_{change}$ 视为连续变量，优化该问题在计算上会非常昂贵。因此，我们将它们离散化为两个有限变量边界内的取值（如表II所示），并通过求这三个变量的笛卡尔积，生成一组分数变化：$< frac_{change}$、$frac_{high}$、$frac_{low} >$。对于 $frac_{change}$，我们采用了更多步数，因为它们最终会用于

表II 分数变化划分范围

观测值	最小值	最大值	步数
$frac_{change}$	-0.1	0.1	50
$frac_{high}$	0	0.1	10
$frac_{low}$	0	0.1	10

股票预测。随后，使用GaussianHMM类的*score* 方法，计算模型对前 d 天观测值序列的对数概率，并找出其中的最大值。这样便能找到概率最高的结果，即：$frac_{change}$，并使用公式5计算收盘价。预测 d 天后的股价时，会针对这些天的索引迭代调用此方法。图6和7分别展示了使用训练好的模型对**ACI Limited** 和 **Grameenphone Limited**的实际股价及预测股价。我们采用的方法具有通用性，只要隐马尔可夫模型是在相应公司的股票数据集上训练的，就可以用于构建该公司的预测器。因此，我们可以采用相同的方法，为达卡证券交易所上市的每家公司分别建立预测模型。

B. 连续波动隐马尔可夫模型

此方法的初始化设置与前文所述方法相同。不同之处在于该方法所采用的特征集。



图 8. ACI股票收盘价的波动。

1) 训练*::* 采用此方法训练模型时，我们首先提取股票的收盘价和成交量。然后计算并存储股票价格的连续差值。最后，将日期、连续差值和成交量组成的元组纵向堆叠，形成特征向量。最后，使用这些特征向量训练模型。模型经过 240 次迭代后收敛。

2) 预测*::* 由于我们的模型经过训练，能够识别收盘价的变化模式，因此在给定前一天收盘价的情况下，它已准备好预测次日收盘价的变化。为实现这一目标，我们计算了转移矩阵A 与数据分布均值的点积。随后，我们将预期变化的数值与前一天收盘价的数值相加

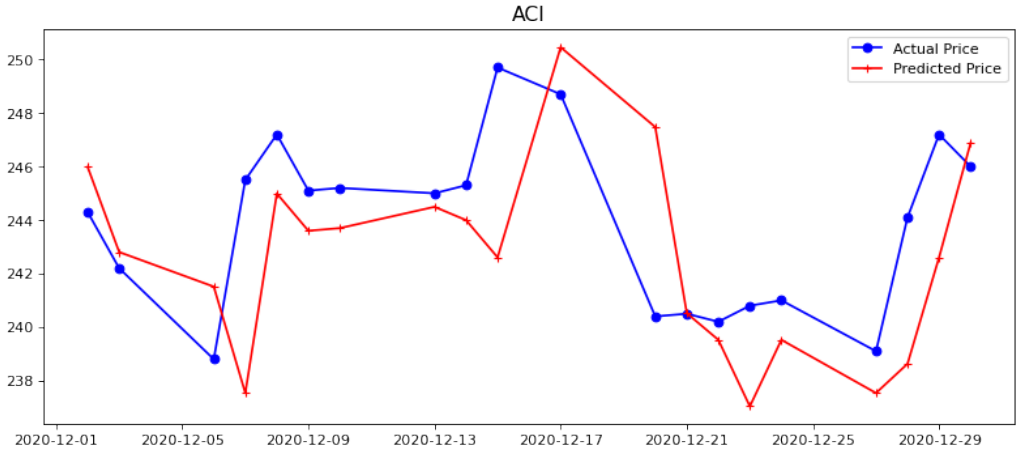

图6. 使用分数变化的隐马尔可夫模型预测2020年12月2日至2020年12月30日ACI股票的实际值与预测值。

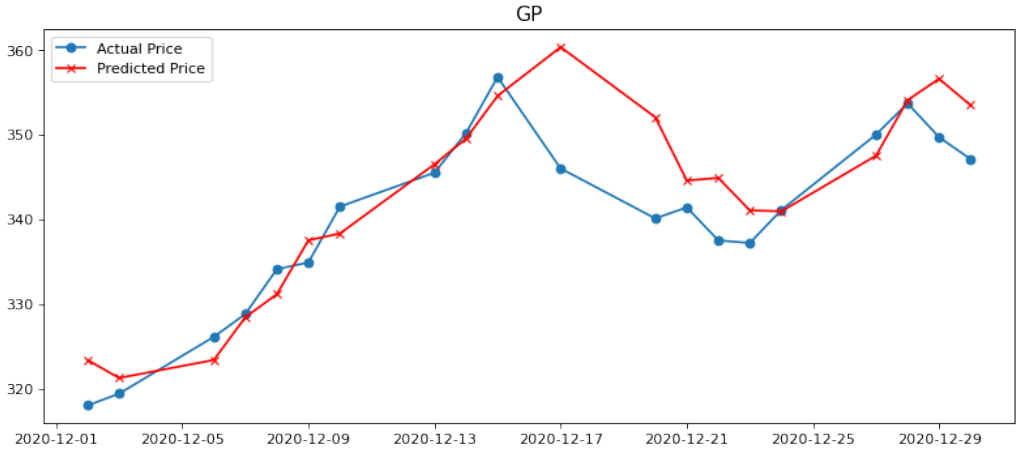

图7. 实际值与预测值　　从2020年12月2日至Dec，GrameenPhone股票的值　　使用分数变化的隐马尔可夫模型预测2020年12月30日的值。

以得出当天收盘价的最终预测。图9展示了*ACI Limited*的实际股价和预测股价。

C. 评估

为了评估我们模型的性能，我们采用了三种广受认可的回归任务性能指标。

平均绝对误差（MAE）是一种直接的度量方法，用于计算实际值与预测值之间的绝对差异。它是对离群值最稳健的指标。

均方根误差（RMSE）顾名思义，是均方误差（MSE）的平方根。MSE是真实值与预测值之间差异平方的均值。

平均绝对百分比误差（MAPE）是实际值与预测值之间以百分比表示的平均绝对误差。

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - y_i|}{|y_i|} \times 100\% \qquad (6)$$

表III ACI股票预测结果的评估

指标	$HMM_{frac_{change}}$	$HMM_{succ_{fluctuation}}$
MAE	2.5064	3.3382
RMSE	3.4003	5.8141
MAPE	1.0265	1.3672

其中，y_i 和 p_i 分别为实际值和预测值，n 为评估数据所涵盖的天数。

表III展示了使用我们训练好的模型对ACI Limited˙股票进行预测所得的指标得分。我们可以看到，模型的*MAPE* 值分别为1.0265和1.3672。根据 [12]，MAPE得分低于5表明预测准确度可以接受。因此，暂且不考虑后续章节所述的局限性，我们可以得出结论：预测器的性能相当出色。

VI. 局限性与未来工作

我们考虑了一个包含2008年1月至2020年12月股票数据的数据集。有许多开源

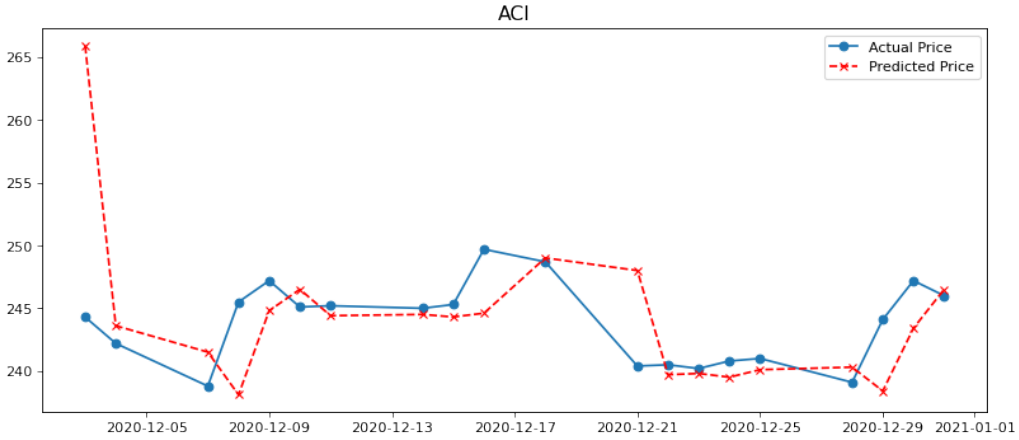

图9. 2020年12月2日至2020年12月30日期间，采用连续波动隐马尔可夫模型对*ACI*有限公司股价实际值与预测值的比较。

抓取工具可以帮助获取最新的股票信息。由于时间有限，我们无法使用这些工具。股票市场具有波动性，股票价格可能大幅波动。对于这种情况，我们的模型无法给出准确的预测。未来，我们希望通过采用各种平滑技术来克服这一局限。

七、结论

人工智能和机器学习技术常用于预测孟加拉国股票市场的股票价值。据我们所知，尚无人利用隐马尔可夫模型的优势开展相关研究。本文展示了隐马尔可夫模型利用历史数据预测股票价格的能力。我们研究了两种方法，二者在典型情况下均取得了令人满意的结果。如果能够收集更多领域知识，并应用处理股票价值突变的技术，我们或许就能进一步提高模型的预测准确性。最后，我们相信，本文将拓展使用隐马尔可夫模型进行预测的研究空间。

参考文献

[1] N. Najkar、F. Razzazi 和 H. Sameti，《一种利用粒子群优化的隐马尔可夫模型语音识别系统新方法》，*2009*年第四届生物启发计算国际会议，第1–6页，2009年。[2] P. Kumawat、A. Khatri 和 B. Nagaria，《利用不变矩结合隐马尔可夫模型与支持向量机-隐马尔可夫模型分类器进行离线手写识别的比较分析》。[3] X. Jiang，《一种基于隐马尔可夫模型的面部表情识别模型》，载于*2011*年电子与机械工程及信息技术国际会议论文集，第6卷，2011年，第3054–3057页。[4] R. Andreao、B. Dorizzi 和 J. Boudy，《通过隐马尔可夫模型进行心电信号分析》，*IEEE*生物医学工程汇刊，第53卷，第8期，第1541–1549页，2006年。[5] A. Gupta 和 B. Dhingra，《利用隐马尔可夫模型进行股票市场预测》，载于*2012*年学生工程与系统会议，2012年，第1–4页。

[6] T. Kimoto、K. Asakawa、M. Yoda 和 M. Takeoka，《采用模块化神经网络的股票市场预测系统》，载于*1990*年*IJCNN*神经网络国际联合会议，1990年，第1–6页，第1卷。[7] L. Cao 和 F. Tay，《使用支持向量机进行金融预测》，神经计算与应用，第10卷，192，2001年5月，184页。[8] M. Hassan 和 B. Nath，《利用隐马尔可夫模型进行股票市场预测：一种新方法》，载于第五届智能系统设计与应用国际会议（*ISDA'05*），2005年，第192–196页。[9] Y. Romahi 和 Q. Shen，《利用自动生成的模糊关联进行动态金融预测》，载于第九届*IEEE*模糊系统国际会议。*FUZZ-IEEE 2000*（目录号*00CH37063*），第1卷，2000年，第493–498页，第1卷。[10] H. Liu 和 B. Song，《基于深度残差网络和股价图的股价趋势预测模型》，载于*2018*年第十一届计算智能与设计国际研讨会（*ISCID*），第02卷，2018年，第328–331页。[11] M. R. Hassan，《隐马尔可夫模型与模糊模型相结合的股票市场预测》，神经计算，第72卷，第16期，第3439–3446页，2009年，金融工程、计算与环境智能（IWANN 2007）。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S0925231209001805 [12] D. A. Swanson，《同一误差汇总度量在同一时间点跨多个特征使用时各数值之间的关系：对MALPE和MAPE的考察》，《经济与金融评论》，第5卷，第1–14页，2015年8月。 [在线]。可访问：https://ideas.repec.org/a/bap/journl/150301. html