计算机科学
COMPUTER SCIENCE

基于拓扑结构特征的投资组合构建研究

李瑞阳, 李庶祎, 杨越溪, 彭楚涵, 邢静雨, 乔高秀

引用本文

李瑞阳, 李庶祎, 杨越溪, 彭楚涵, 邢静雨, 乔高秀. 基于拓扑结构特征的投资组合构建研究[J]. 计算机科学, 2025, 52(10): 13-21.

LI Ruiyang, LI Shuyi, YANG Yuexi, PENG Chuhan, XING Jingyu, QIAO Gaoxiu. Research on Portfolio Construction Based on Topological Structure Features [J]. Computer Science, 2025, 52(10): 13-21.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100136

基于拓扑结构特征的投资组合构建研究

李瑞阳　李庶祎　杨越溪　彭楚涵　邢静雨　乔高秀
西南交通大学数学学院　成都 611756
(2637844500@qq.com)

摘　要　近年来,拓扑数据分析(Topological Data Analysis,TDA)在金融领域的应用逐渐显现出价值。TDA 通过持久同调等方法构建复形,能有效量化数据的形状,以便提取数据信息,为时间序列分析,特别是金融时间序列的聚类与投资组合的构建提供了独特优势。基于此,通过采用 TDA 方法对中国股票市场的时间序列数据进行深入挖掘,结合聚类算法,并将其应用于投资组合的构建,分析其有效性。通过滑动窗口法进行验证,结果表明基于 TDA(去噪)聚类的投资组合在回报风险比和稳定性方面表现良好,优于市场整体表现。研究表明,TDA 方法可以更有效地挖掘股票数据中的信息,为投资者提供科学依据,从而取得最佳收益。

关键词:拓扑数据分析;投资组合;时间序列;聚类
中图分类号　TP399

Research on Portfolio Construction Based on Topological Structure Features

LI Ruiyang,LI Shuyi,YANG Yuexi,PENG Chuhan,XING Jingyu and QIAO Gaoxiu
School of Mathematics,Southwest Jiaotong University,Chengdu 611756,China

Abstract　In recent years,the application of topological data analysis(TDA) in the financial field has gradually demonstrated its value. TDA,through methods such as persistent homology,constructing complexes that effectively quantify the shape of data,facilitating the extraction of data information. This provides unique advantages for time series analysis,particularly in the clustering of financial time series and the construction of portfolios. Based on this,by deeply mining the time series data of China's stock market using TDA methods,combined with clustering algorithms,and applying these insights to portfolio construction,the effectiveness of such approaches is analyzed. The results,validated through the sliding window method,indicate that portfolios constructed based on TDA(denoising) clustering perform well in terms of return-risk ratio and stability,outperforming the overall market. Therefore,the TDA method can more effectively mine information from stock data,providing a scientific basis for investors to optimize returns.

Keywords　Topological data analysis,Portfolio,Time series,Clustering

近年来,拓扑数据分析(TDA)的理论基础不断完善,应用范围也日益广泛,其在金融领域的应用也逐渐显现出独特的优势。David 等从理论上证明了 TDA 中的持续性图具有稳定性,说明 TDA 对所捕捉数据的异常值和缺失值具有较强的抵抗力,但并未理论证明或实例验证抵抗性的强度[1]。Gidea 等基于拓扑分析方法研究美国四大主要的股指日收益率时间序列,通过L_p范数量化持续性景观的时间变化,证明 TDA 提供了一种新的计量经济学分析类型,并提出用于早期金融危机检测的方法[2]。Fang[3]以持续同调为主的拓扑数据分析方法提取股票数据中的低维拓扑特征,并有效结合机器学习提高了股票涨跌预测的准确性,展示了其在金融预测中的实际应用价值。Karan 等[4]采用机器学习模型,将 TDA 应用到时间序列分类任务中,进一步拓展了 TDA 的应用范围。Katz 等[5]利用 Takens 嵌入和滑动窗口等技术提出基于拓扑数据分析的新的计量经济学方法,并对北美经济的多个行业的市场不稳定性进行检测,提供了市场力导致的金融崩盘临近的先行指标。

在金融时间序列分析中,聚类方法也扮演着重要角色。De Gregorio 等[6]提出了以马尔可夫算子为基础的时间序列相异性测度来实现动态聚类,并利用此方法对股票价格时间序列进行分析。Zhao 等[7]运用 PAM,agnes,diana 3 种聚类方法分别对金融时间序列进行聚类,进一步丰富了聚类方法

到稿日期:2025-01-21　返修日期:2025-05-06
基金项目:中央高校基本科研业务费专项资金(202410613071,2682025ZTPY001);国家自然科学基金(72001180)
This work was supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China(202410613071,2682025ZTPY001) and National Natural Science Foundation of China(72001180).
通信作者:乔高秀(gxqiao@home.swjtu.edu.cn)

在金融领域的应用。Li[8]采用特征因子判别法与 K-means 算法，来提高聚类分析的准确性和时间序列识别的相似度，并通过观察价差组合在样本外窗口的表现构建投资组合。Gidea 等[9]还使用了持久景观(persistence landscapes)之间的 L_2 距离来进行 k 均值聚类。Majumdar 等[10]研究金融时间序列聚类问题，证明不同行业的股票价格时间序列的拓扑特征不同，并且具有可以使用 TDA 来识别的特征，为依据拓扑特征进行聚类提供了依据。Goel 等[11]则应用了一种基于 TDA 景观范数的聚类方案来进行投资组合选择。

由于金融市场具有不稳定性，获取金融数据的渠道复杂多样，传统基于数据分析的投资组合构建方法难以总结市场规律，且处理大量数据的效率不高。Pan 等[12]在分析最优化投资组合不稳定原因的基础上，提出了一种用聚类分组法调整样本协方差阵从而得到一个更好的投资组合的新方法，但需要根据实际情况调整相关系数检验。De Luca 等[13]提出了两种基于股票价格时间序列动态聚类的投资组合选择策略，利用具有时变参数的 Copula 函数估计的低尾相关系数，推导出股票聚类的相异矩阵。Li 等[14]构建出对离群值有更好抗差性和抗干扰性的稳健夏普单指数模型。Goel 等[11]证明 TDA 可以更好地代替传统风险指标，并运用 TDA 解决了金融领域的资产配置问题，构建最优投资组合。Zhang 等[15]运用随机森林、极端梯度提升和支持向量回归 3 种机器学习方法预测股票的收益率，构建了基于机器学习预测的均值-下半方差投资组合模型。Zhang 等[16]构建了一个包含交易成本和借贷限制的，基于 BP 神经网络的三因子均值-方差投资组合模型，其在样本内外测试中均取得了优于传统模型的收益表现和更高的风险调整后收益。

尽管构建投资组合的策略还在不断完善，但是 TDA 的独特优越性，以及基于 TDA 对股票时间序列进行聚类，即根据提取时间序列的拓扑特征对不同行业的股票进行分类的可行性已经展现。不仅如此，TDA 还被证实具有稳定性，能够有效抵抗数据中的异常值和缺失值，还可以成功应用于金融危机检测、股票涨跌预测以及投资组合构建等多个方面。通过结合机器学习等先进技术，TDA 在金融时间序列分析中的准确性和效率得到了显著提升，为金融行业的决策提供了有力支持。但仍有部分问题待解决：1)股票时间序列在持续同调过程中存在的其他拓扑特征有待提取，以及这些特征在聚类算法中如何应用；2)在基于 TDA 进行时间序列聚类时，TDA 对数据异常值的抗干扰性有待验证，即如何有效去噪以提高拓扑特征提取的准确性；3)虽然 TDA 在聚类分析中表现出色，但其与传统投资组合构建方法的结合仍处于初步探索阶段，如何将基于 TDA 的聚类结果与传统投资组合优化模型有效结合，还有待研究。

本文深入分析了 TDA 在金融市场动态性背景下的适应性，采用了一种基于 TDA 的去噪方法，在提取数据的拓扑特征之前，对股票价格收盘价时间序列进行去噪，并将其应用于股票价格时间序列的聚类分析和投资组合构建。以拓扑理论为基础，提取数据集的多个拓扑特征，运用主成分分析法，解决了信息冗余或丢失的问题，并确定最终的聚类指标。构建投资组合时，采用滑动窗口法验证投资组合的夏普比率和累计收益两个指标，通过对比得到拓扑数据分析方法对于在聚类基础上构建投资组合的优势，以及去噪对于该方法的有效性，为金融领域的研究和实际应用提供了新的思路和方法。本文的主要贡献在于：

1)深入探讨了 TDA 方法如何捕捉到金融市场的动态变化中的拓扑特征；

2)提出了一种新的去噪方法，提高了拓扑特征提取的准确性；

3)通过实证研究验证了基于 TDA 的投资组合构建方法的稳定性和优越性，为金融领域的研究和实际应用提供了新的思路和方法。

1 TDA 理论基础

本章主要介绍 TDA 相关预备知识，明确相关术语和符号表示，理清概念定义，便于后文中的理解。

1.1 单纯复形

定义 1(单纯形) 欧氏空间中处于一般位置的 $n+1$ 个点 $\{a_0,\cdots,a_n\}(n\geqslant 0)$ 的凸包称为一个 n 维单纯形，简称 n 维单形，记作 $(a_0,a_1,\cdots,a_n)$。称 a_i 为它的顶点，$i=0,\cdots,n$。

命题：设 $n>0$，则 $A=\{a_0,\cdots,a_n\}$ 处于一般位置⇔向量组 $\{\boldsymbol{a}_1-\boldsymbol{a}_0,\cdots,\boldsymbol{a}_n-\boldsymbol{a}_0\}$ 线性无关。

定义 2(单纯复形) 单纯复形 K 是单纯形的有限集合，满足以下条件：

1)如果一个单形是 K 的单形，那么它的任意面也属于 K；

2)K 中所有单形都规则相处，即它们之间不相交或相交于公共面。

1.2 持续同调

定义 3(V-R 复形) 给定一组数据点集 X 和一个参数 $\varepsilon>0$，$V\text{-}R$ 复形 $VR(X,\varepsilon)$ 是由所有直径小于 2ε 的子集所组成的集合。

设在持续同调过程中，第 i 个洞的产生时间为 b_i，消失时间为 d_i，则该洞对应某一数对 (b_i,d_i)，点云数据的拓扑特性可从这些点中体现。某一洞持续时间越长，该洞所代表的拓扑特征就越显著；相反，则可将其视为噪声，不予考虑。条码图和持续性图是上述数对可视化的两种方式，条码图如图 1 所示，持续性图如图 2 所示。

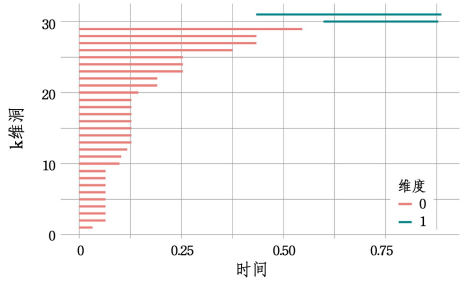


图 1 条码图

Fig. 1 Barcode image

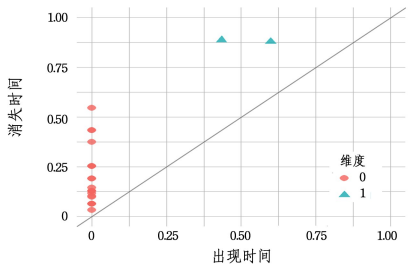


图 2 持续性图

Fig. 2 Persistence diagram

1.3 持续性景观

虽然持续性图为拓扑数据分析提供了关键信息，但其多重集特性限制了进一步处理。为解决这一问题，Bubenik[17]引入了持续性景观。持续性景观作为泛函空间中的实值函数，具备良好的稳定性，更有效地揭示了数据的拓扑特征，便于统计学和机器学习方法的应用。

对持续性景观[18]进行定义，持续性图中的点 $p=(b,d)$，b 和 d 分别为出生时间和死亡时间。持续性景观为分段线性函数 $g:N\times R\to[0,\infty]$。构造如下表达式：

$$\Lambda_p(t)=\begin{cases}0, & x\notin(b,d)\\ x-b, & x\in\left(b,\dfrac{b+d}{2}\right]\\ -x+d, & x\in\left(\dfrac{b+d}{2},d\right)\end{cases} \tag{1}$$

则持续性景观$\{\lambda_i(t)\}$定义为：

$$\lambda_i(t)=\lambda(i,t)=imax,t\in[0,T],i\in N \tag{2}$$

给定 $i\in N$，$\lambda_i(t)$称为第 i 景观。由于持续性图到持续性景观的映射为单射，因此在这中间无信息损失，得到的持续性景观图如图 3 所示。

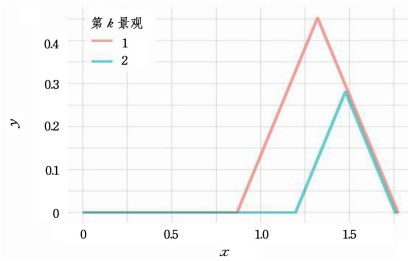


图 3 持续性景观图

Fig. 3 Persistent landscape diagram

1.4 Takens 嵌入：相空间重构

Takens 嵌入定理指出，如果有一个光滑的动态系统，并从中测量一个变量，那么通过时间延迟嵌入，将这个变量在不同时间点的值组成向量，就有可能在一个高维空间中重构出与原系统动力学等价的相空间。具体来说，这个过程是通过构造一个向量序列来实现的。

$$\boldsymbol{x}(t)=[\boldsymbol{x}(t),\boldsymbol{x}(t+\tau),\cdots,\boldsymbol{x}(t+(m-1)\tau)] \tag{3}$$

其中，$\boldsymbol{x}(t)$是观测的时间序列数据，τ 是选定的时间延迟，m 是嵌入维数。

由于 TDA 需要输入点云数据，因此对一维时间序列进行 TDA 分析前必须先进行相空间重构，将时间序列通过 Takens 嵌入方法转换为点云，利用时间延迟嵌入将其映射到高维空间。

图 4 和图 5 给出了两个时间序列 $y_1=\sin(10\pi t)$ 和 $y_2=e^{-t}\cos(10\pi t)$ 在参数 $m=2,\tau=1$ 下延迟嵌入之后的点云图像，图 4 为原始时间序列的点云图，图 5 为重构之后的点云图。

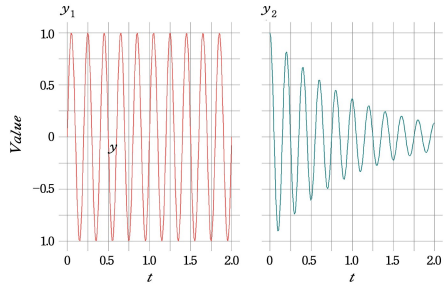


图 4 原始时间序列的点云图

Fig. 4 Point cloud diagram of the original time series

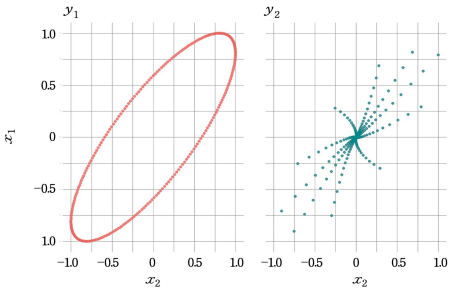


图 5 相空间重构之后的点云图

Fig. 5 Point cloud diagram after phase space reconstruction

2 数据的选择与预处理

本文所选取的股票数据为上证 180 指数成分股从 2021 年 1 月到 2023 年 12 月这 3 年的日度收盘价和收益率数据。由于上证 180 指数的成分股每年会调整两次，若存在新入选的上市股票或中途退选的股票，则对应历史数据是缺失的，不利于数据的统一处理和操作。为保证股票数据的连续性，方便后续的验证工作，本文只针对近 3 年中持续存在的成分股进行分析，共 78 支，时间序列长度为 764。数据来源为锐思金融研究数据库，所选股票的行业分布如表 1 所列。

表 1 股票行业分布

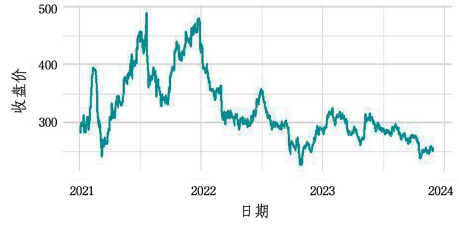
Table 1 Industry distribution of stocks

行业分布	股票数目	部分股票
农、林、牧、渔业	1	北大荒
采选业	7	中国石化、中国石油
制造业	24	宝钢股份、伊利股份
电力、热力、燃气及水的生产和供给业	2	长江电力、中国核电
建筑业	6	中国中铁、中国电建
批发和零售业	1	上海医药
运输业	2	上海机场、中国国航
信息传输、软件和信息技术服务业	3	中国联通、恒生电子
金融保险业	28	建设银行、中国平安
房地产业	3	保利发展、金地集团
租赁和商业服务业	1	中国中免

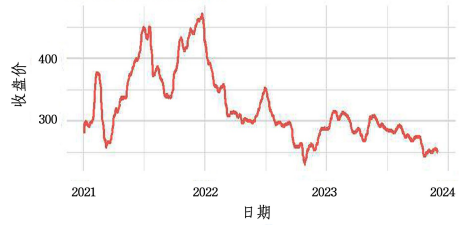
基于上述结论，本文使用均值滤波对原始的数据进行去

噪处理。均值滤波法[19-22]的原理是通过计算某一点周围多个点的平均值来作为该点滤波之后的值，可以有效地平滑信号，且选取简单，易于实现。因此，通过使用均值滤波法对这些数据点进行去噪，可研究去噪能否改进原始的拓扑数据分析方法，使其更准确地捕捉到股票价格的长期趋势和周期性变化。

在处理股票数据时，对于每一支股票的某一个时间收盘价，取前后各 7 个点与该点共 15 条数据的均值作为该时间对应的收盘价，可以试图减少突发的大幅波动，使数据更加平稳，便于在后续滑动窗口验证分析中使用。去噪后的数据与原始数据的对比图如图 6 所示。



(a)原始时间序列



(b)经过去噪处理后的时间序列

图 6 股票数据去噪前后的时间序列对比图

Fig. 6 Comparison chart of the time series of stock data before and after enoisingd

3 基于 TDA 的时间序列聚类

3.1 拓扑特征的计算与选取

本文旨在利用拓扑数据分析股票随时间变化的特征，并将其应用于后续的聚类分析。1.2 节介绍了基础的可视化拓扑特征工具，即持续性图和条形码图，但它们仅是二维点集合，不适合直接应用于机器学习的统计方法中。本文在已有研究的基础上，以洞的稳定性为主要因素，进一步总结了部分拓扑特征，以便用于后续的聚类分析。

基于此，以能否反映时间序列的形态和发展趋势为依据，得到每个时间序列可提取出的 11 个拓扑指标，形成一个 11 维的拓扑特征，分别是 1 维洞的数量、0 维和 1 维洞的最大持续时间、0 维和 1 维显著特征的数量、0 维和 1 维洞的平均持续时间、1 维的瓦瑟斯坦距离、1 维的L_1和L_2范数以及持续性熵。接下来将应用此特征进行聚类。

3.2 聚类算法

本文提出的基于 TDA 的时间序列聚类算法实际上是一种基于特征的方法，它依照设定的规则从时间序列中提取特征，将原始时间序列转换为低维特征向量，然后利用传统的聚类方法对这些特征向量进行聚类。在拓扑特征计算与选取的基础上，本文提取了 11 维拓扑特征，接下来将对其进行聚类。

根据文献[23]，相比 K-Means 算法，K-Medoids 选择的聚类中心点是真实样本点而非均值，因此它对噪声具有更强的稳健性；且在数据集较小的情况下，K-Medoids 的计算量通常比 K-Means 更小，只需计算距离。与常用的 K-Means 聚类方法相比，采用 K-Medoids 方法对项目类别属性进行聚类，不仅解决了评分聚类可靠性不高的问题，而且具有更好的鲁棒性。因此，可认为该算法能有效地提高推荐质量。鉴于此，本文采用 K-Medoids 算法对所提取的拓扑特征进行聚类，算法流程如算法 1 所示。

算法 1 基于拓扑数据分析的时间序列聚类算法

1. 相空间重构，将时间序列数据按 d=1，τ=1 转换为点云数据；
2. 通过持续同调，提取时间序列的 11 维拓扑特征并标准化；
3. 利用主成分分析法(PCA)对拓扑特征进行降维，选取 5 个主成分作为新拓扑特征；
4. 基于提取的 5 个主成分，利用 K-Medoids 算法聚类，聚类数目未知的情况下，采用轮廓系数法确定聚类数目。

需要注意的是：在第 4 步中，本文利用提取出的 5 个主成分作为新拓扑特征，采用 K-Medoids 算法进行聚类。轮廓系数法用于确定最优的聚类数目，以确保聚类结果既紧凑又分离良好。它考虑了簇内的紧密度和簇间的分离程度，对于每个对象，计算其轮廓系数，范围为[-1，+1]，其中高值表示该对象很好地匹配其自身的簇，同时远离其他簇。在应用 K-Medoids 算法时，通过调整聚类数目并计算相应情况下的平均轮廓系数，选择使平均轮廓系数最大的聚类数目作为最终的聚类数目。

3.3 聚类效果评价指标

UCR 是时间序列数据集[24]，且每个数据集样本都带有样本类别标签，目前是时间序列挖掘领域重要的开源数据集资源。为检验所提出的 TDA 聚类的效果，选取其中的两个典型数据集进行实验分析。两个数据集的信息如表 2 所列。

表 2 实验分析的数据集

Table 2 Datasets for the experimental analysis

数据集名称	类别	样本数	时间序列长度	数据含义
BirdChicken	2	20	512	动物轮廓曲线
BeetleFly	2	20	512	甲壳虫/苍蝇轮廓曲线

对于聚类的评价，此处采用 4 个常见的评价指标，分别为：聚类纯度(Purity)、F 值、兰德系数(RI)和调整兰德系数(ARI)。

1)聚类纯度

聚类纯度的基本思想是将正确分类的样本数除以总样本数。由于在聚类结果中不知道每个簇的真实类别，因此需要选择每个簇中最大类别的样本数来计算纯度。其计算式如式(4)所示：

$$P(\Omega,C)=\frac{1}{N}\sum\max\{\omega_k\cap c_j\} \tag{4}$$

其中，N 为样本总数，$\Omega=\{\omega_1,\cdots,\omega_k\}$为聚类得到的 k 个簇，$C=\{c_1,\cdots,c_j\}$为正确的类别；ω_k表示聚类后第 k 个簇中的所有样本，c_j表示第 j 个类别中真实的样本。在这里，P 的取值范围为[0，1]，其值越大表示聚类效果越好。

2）F 值与兰德系数

混淆矩阵的定义如表 3 所列。

表 3 混淆矩阵

Table 3 Confusion matrix

	同簇	非同簇
同类	TP	FN
非同类	FP	TN

同簇：识别类别为正类。

同类：真实类别为正类。

TP：真正类，同簇且同类。

FN：假负类，非同簇但同类。

FP：假正类，同簇但非同类。

TN：真负类，非同簇且非同类。

兰德系数：用于衡量聚类结果与真实分类之间的相似度。其定义如下：

$$RI=\frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

精确度：以预测结果为判断依据，预测为正例的样本中预测正确的比例。其定义如下：

$$Precision=\frac{TP}{TP+FP} \tag{6}$$

召回率：以实际样本为判断依据，实际为正例的样本中，被预测正确的正例占总实际正例样本的比例。其定义如下：

$$Recall=\frac{TP}{TP+FN} \tag{7}$$

F 值：利用精确度和召回率评估结果好坏的指标。其定义如下：

$$F_\beta=(1+\beta^2)\frac{Precision\cdot Recall}{\beta^2\cdot Precision+Recall} \tag{8}$$

此处，RI 与F_β的取值范围均为$[0,1]$，其值越大表示聚类效果越好。在后续评价中，取 $\beta=1$。

3）调整兰德系数

为去掉随机标签对兰德系数评估结果的影响，引入调整兰德系数这一评价指标。列联表如表 4 所列。

表 4 列联表

Table 4 Contingence table

$X\backslash Y$	Y_1	Y_2	…	Y_s	sum
X_1	n_{11}	n_{12}	…	n_{1s}	a_1
X_2	n_{21}	n_{22}	…	n_{2s}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
X_r	n_{r1}	n_{r2}	…	n_{rs}	a_r
sum	b_1	b_2	…	b_s	

表 4 中，$X=\{X_1,X_2,\cdots,X_r\}$表示聚类得到的 r 个簇的集合，而 $Y=\{Y_1,Y_2,\cdots,Y_s\}$表示根据正确标签对应结果修正后的集合，n_{ij} 表示X_i与Y_j相交部分的样本数量，即$n_{ij}=|X_i\cap Y_j|$。

由此可得调整兰德系数的计算式：

$$ARI=\frac{\sum_i\sum_j\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}} \tag{9}$$

其中，i 的取值范围为$[1,r]$，j 的取值范围为$[1,s]$，r 表示聚类结果的簇数，s 表示真实分类的类别数，它们均为常数；ARI 的取值范围为$[-1,1]$，其值越大表示聚类效果越好。

3.4 聚类效果分析

为验证 TDA 聚类的效果，将另选取 3 种传统的时间序列聚类方法与之相比，分别为基于欧几里得距离（Euclidean Distance，ED）的 K-Medoids 聚类[25]、基于 DTW[26] 的 K-Medoids 聚类和 K-Shape 聚类[27]。这 3 种方法各有优劣，下文将展示各聚类方法在数据集上的效果，并对结果进行比较。

1）BirdChicken 数据集

对数据集中的两类时间序列按类别求平均，得到两类时间序列的平均走向。该数据集的两时间序列如图 7 所示。

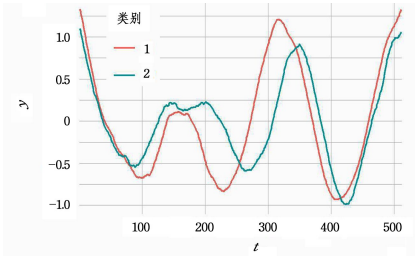


图 7 BirdChicken 数据集的两类时间序列

Fig. 7 Two types of time series of the BirdChicken dataset

表 5 列出了在 BirdChicken 数据集上的最终聚类结果。其中，以主成分作为拓扑指标的 TDA 聚类各方面表现均优于其他 3 种聚类方法。

表 5 BirdChicken 数据集上聚类结果的比较

Table 5 Comparison of clustering results on the BirdChicken dataset

	Purity	*F1*	*RI*	*ARI*
TDA	0.771	0.638	0.638	0.277
ED	0.605	0.521	0.512	0.029
DTW	0.562	0.509	0.498	0.017
K-Shape	0.647	0.566	0.546	0.102

本文共提取 11 个拓扑指标，其选取组合共有 $\sum_{i=1}^{11}C_{11}^i=2047$ 种，遍历这 2047 种组合，统计各组合的 ARI，绘制 ARI 频数分布直方图，如图 8 所示。

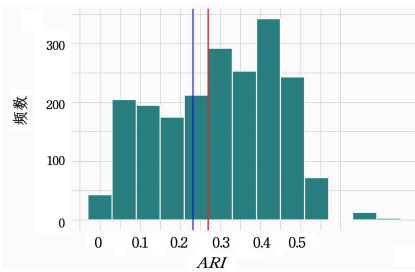


图 8 BirdChicken 数据集基于不同拓扑指标聚类的 ARI 分布（电子版为彩图）

Fig. 8 ARI distribution of clustering based on different topological indicators for the BirdChicken dataset

图 8 中，红线标注的是本文提出的基于主成分的聚类方法的 ARI，蓝线标注的是使用全部 11 个指标进行聚类的 ARI。由图可见，主成分聚类优于全指标聚类，且主成分聚类优于 40.5%的指标选择。

2)BeetleFly 数据集

对数据集中的两类时间序列按类别求平均,得到两类时间序列的平均走向。该数据集的两时间序列如图 9 所示。

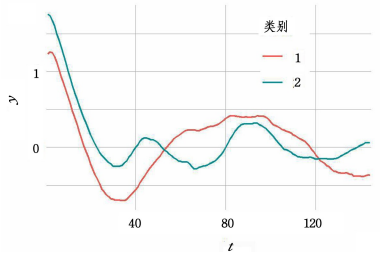


图 9 BeetleFly 数据集的两类时间序列

Fig. 9 Two types of time series of the BeetleFly dataset

表 6 列出了在 BeetleFly 数据集上的最终聚类结果。其中,根据主成分进行 TDA 聚类的效果最佳。

表 6 BeetleFly 数据集上聚类结果的比较

Table 6 Two types of time series of the BeetleFly dataset

	Purity	*F1*	*RI*	*ARI*
TDA	0.697	0.583	0.567	0.136
ED	0.618	0.513	0.520	0.045
DTW	0.649	0.537	0.541	0.087
K-Shape	0.600	0.513	0.518	0.051

统计各组合的 ARI,绘制 ARI 频数分布直方图,如图 10 所示。

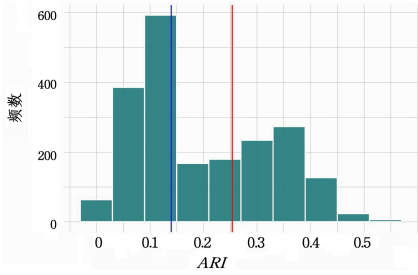


图 10 BeetleFly 数据集基于不同拓扑指标聚类的 ARI 分布

Fig. 10 ARI distribution of clustering based on different topological indicators for the BeetleFly dataset

由图 10 可以得出,基于主成分的聚类优于全指标聚类,且优于 41.3%的指标选择。

综上所述,得到如下结论:

1)基于拓扑特征主成分的聚类虽然无法总是达到利用拓扑指标聚类可达到的最佳效果,但其稳定性较好。在实际聚类时,数据集多为无标注数据集,无法准确找到合适的拓扑指标。因此,TDA 聚类方法对大多数数据集均适用,适用范围更加广泛。

2)TDA 聚类在不同数据集上的表现略有浮动,但相较于其他 3 种方法具有一定的稳定性,在各个数据集上均表现良好。

4 投资组合的构建与验证

4.1 投资组合的构建

根据马科维茨投资组合理论[28],投资时尽量选择弱相关的股票有利于规避风险,提高收益。而聚类可以使不同类簇之间的相关性变小,故从不同类簇中分别选取股票更容易优化投资组合,实现其多样化。

夏普比率能够同时对收益与风险加以综合考虑,常用于衡量基金绩效情况,与本文评价投资组合的需求相近,故选择夏普比率作为评价投资组合优劣的指标。

夏普比率 S 定义为:

$$S=\frac{E(R_p)-R_f}{\sigma_p} \tag{10}$$

其中,$E(R_p)$为投资组合年化利率的期望值,R_f为年化无风险利率,σ_p为投资组合年化利率的标准差。夏普比率越高,认为投资组合就越好,且在相同的时间内累计收益越高,投资组合效果越好。

为验证使用基于主成分分析的 TDA 方法进行聚类对于构建投资组合的应用效果,以及对原始数据去噪对于应用效果的改进效果,本文选取 4 个基准策略作为对比,分别构建投资组合。

1)基于随机选择的投资组合。直接从 78 支股票中随机选取 5 支股票作为成员股,利用均值-方差模型和蒙特卡洛模拟,确定投资组合的夏普比率最高时的各股权重,最终构成投资组合方案。

2)基于行业分类的投资组合。首先将 78 支股票根据所属的不同行业进行分类,随机选取其中的 5 个行业,并从上述 5 个行业中各随机选取一支股票作为投资组合的成员股,利用均值-方差模型和蒙特卡洛模拟,确定夏普比率最高时各支股票的权重,最终构成投资组合方案。

3)基于原始数据进行 TDA 聚类的投资组合。应用 TDA 聚类将股票聚为 5 簇,簇内的股票都有相似的收益模式。从各簇中随机选取一支股票作为投资组合的成员股,运用均值-方差模型和蒙特卡罗模拟计算组合达到最高夏普比率时的各股权重,形成最终的投资组合。

4)先对原始股票收盘价时间序列进行去噪,并根据去噪后的结果计算得出股票的收盘价时间序列,采用与方案 3)相同的方法形成最终的投资组合。

5)选取沪市 A 股中规模大、流动性好的 180 只股票组成的投资组合,即上证 180 指数。

通过比较 TDA、随机划分、行业划分和 TDA(去噪)4 种策略,可以对比不同投资策略下的投资组合效果。通过比较 TDA、TDA(去噪)和上证 180 指数 3 种策略,可以进一步评估基于 TDA 聚类及改进的 TDA 聚类的投资组合策略相对于市场整体的表现。

4.2 投资组合的验证方案

为考查实际投资组合的效果,利用滑动窗口法检验投资组合在样本外窗口上的表现。

首先,将时间序列划分为多段窗口,每段窗口算作一个独立的数据集,窗口内又分为样本内窗口和样本外窗口,分别对应训练集和测试集。具体到本文对投资组合策略的验证中,即在样本内窗口上构建投资组合,在样本外窗口上观察对应投资组合的效益,其中方案 4)构建的投资组合仍然根据实际数据即原始数据对应的样本外窗口进行验证。

设时间序列长度为 n,窗口长度为 $d=d_1+d_2<n$,d_1为样

本内窗口的长度，d_2 为样本外窗口的长度。根据数据的频率和实际交易日数量来确定滑动窗口的长度，并用半年的历史数据构建投资组合，以未来 3 个月考察投资组合效果。共划分 10 个窗口，最后一个窗口的样本外窗口长度为 35。具体算法步骤如算法 2 所示。

算法 2 滑动窗口法检验投资组合效益步骤

1. 输入股票收盘价数据，以 $d_1=128$，$d_2=64$，划分为 10 个窗口；
2. 在每个窗口上检验四种投资组合策略：利用样本内窗口确定投资组合，利用样本外窗口计算对应投资组合的每日对数收益率；
3. 将 10 个样本外窗口拼接，得到不同投资策略在较长一段时间内的收益率情况，计算累计收益率曲线；
4. 重复上述步骤 10 次，计算各策略累计收益率曲线的平均值，将其作为最终表现。

4.3 风险收益分析

本文利用聚类分析结果选择股票，构建基于 TDA 与基于 TDA 去噪的投资组合，与随机选择、行业划分两种方法进行对比，计算不同策略下的风险与收益。

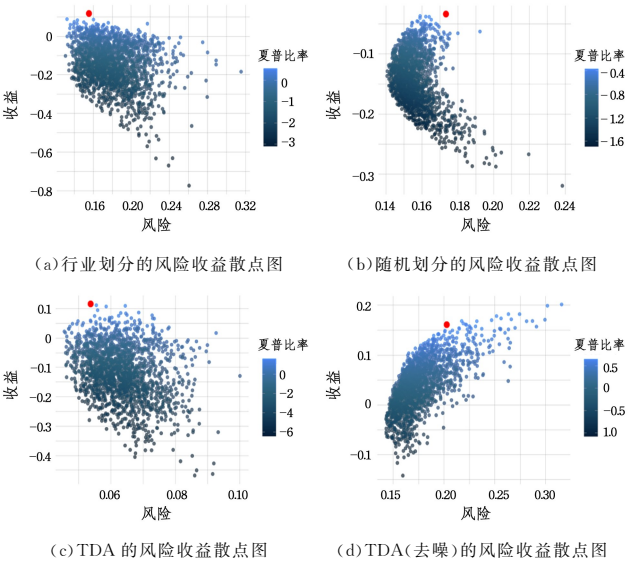
依据马科维茨均值方差模型，使用均值度量收益，方差度量风险。此处，收益使用历史收益率均值来衡量，风险使用历史收益率方差来衡量。

$$E(r_p)=\sum_{i=1}^{5}\omega_i E(r_i) \tag{11}$$

$$Var(r_p)=\sum_{i=1}^{5}\omega_i^2 Var(r_i)+\sum_{i\neq j}\omega_i\omega_j Cov(r_i,r_j) \tag{12}$$

其中，r_i 为投资组合内第 i 支股票的收益率，ω_i 为投资组合内第 i 支股票的权重。

构建投资组合过程中的风险和收益的散点图如图 11 所示。其中，风险指非市场性风险，可以通过分散投资进行适当规避，红色标记的是夏普比率最大的点。



(a)行业划分的风险收益散点图　(b)随机划分的风险收益散点图

(c)TDA 的风险收益散点图　(d)TDA(去噪)的风险收益散点图

图 11 部分投资组合的风险-收益图(电子版为彩图)

Fig. 11 Risk-return graph of some investment portfolios

由图 11 可得，通常在金融或投资领域，风险和收益是成正比的。根据散点的分布情况，可以看到数据点在某些区域密集，而在其他区域稀疏。密集区域代表了更常见的风险-收益组合，而稀疏区域则代表了较少见的情况。

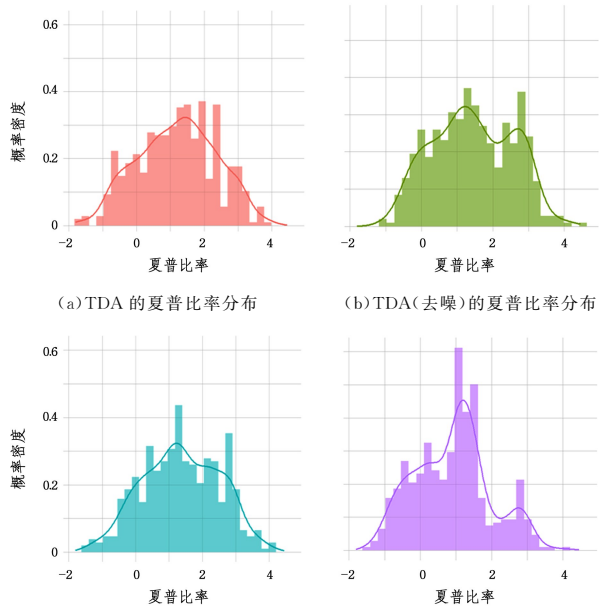
可见，基于 TDA(去噪)的投资组合回报风险最高；基于 TDA 的投资组合次之，稳定性较好；基于行业划分的投资组合回报风险比位列第三；基于随机选取的投资组合回报风险比最低，为负数，风险大于收益。基于此，可推断 TDA 聚类具有一定的优越性，且对数据进行预处理会使效果更佳。

通过滑动窗口比较 4 种策略所构建投资组合的优劣，模拟 500 次，投资组合夏普比率的均值、标准差如表 7 所列。夏普比率的分布如图 12 所示。

表 7 3 种策略下投资组合的夏普比率的均值、标准差

Table 7 Mean and standard deviation of the Sharpe ratio of investment portfolios under three strategies

	TDA(去噪)	TDA	行业划分	随机选择
均值	1.432	1.205	0.898	1.169
标准差	1.125	1.153	1.059	1.346



(a)TDA 的夏普比率分布　(b)TDA(去噪)的夏普比率分布

(c)随机选择的夏普比率分布　(d)行业划分的夏普比率分布

图 12 4 种策略下投资组合的夏普比率分布

Fig. 12 Sharpe ratio distribution of investment portfolios under four strategies

可见，基于 TDA 去噪的投资组合的回报风险比更高，且最为稳定；基于 TDA 的投资组合次之，效果较稳定；基于随机选择的投资组合位列第三，效果最不稳定；基于行业划分的投资组合虽效果稳定，但回报风险比最低。这说明 TDA 去噪聚类产生了较好的效果，依据聚类类簇构建的投资组合具有一定的优越性。

4.4 累计收益曲线

通过滑动窗口法，计算 5 种投资组合的累计收益，并重复多次求平均值，得到如下结论。

1)通过滑动窗口法，对基于原始数据的 TDA 聚类的投资组合、基于原始数据去噪所构建的投资组合、基于行业划分的投资组合以及基于随机划分的投资组合进行比较，结果如图 13 所示。

通过比较基于不同聚类方法得到的投资组合方案的累计收益可以看出，拓扑特征视角下的两种基于股票时间序列的聚类方法所对应的效果更好，始终优于其他两种方案；并且，

对于对原始数据集进行去噪处理得到的数据，基于 TDA 构建的投资组合方案在验证环节的表现更好，有利于提高最终投资组合的效果。

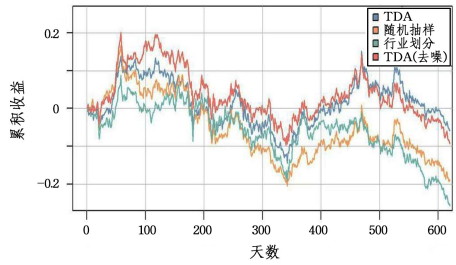


图 13 4 种投资方案下的累计收益对比曲线

Fig. 13 Comparative curves of cumulative returns under four investment schemes

2)通过滑动窗口法，比较基于原始数据的 TDA 聚类的投资组合、基于原始数据去噪所构建的投资组合以及上证 180 指数，结果如图 14 所示。

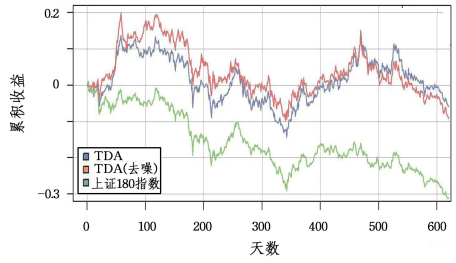


图 14 3 种投资方案下的累计收益对比曲线

Fig. 14 Comparative curves of cumulative returns under three investment schemes

通过比较基于不同聚类方法得到的投资组合方案的累计收益可以看出，基于去噪处理后的 TDA 时间序列聚类方法效果最好，其次是基于 TDA 时间序列的聚类方法，且这两种方法均始终优于上证 180 指数的累计收益。这说明基于 TDA 去噪及聚类的效果优于市场整体表现。

结束语 本文提出了一种基于 TDA 的股票聚类方法，即在去噪后的时间序列上应用 TDA，进而构建更优的投资组合。通过与其他较为传统的投资组合构建方法的效果和整体股票市场形势进行对比分析，证明所提方法具有良好且稳定的收益效果。

目前，TDA 仍属于较新兴的研究领域，将 TDA 应用于股票时间序列分析，是对传统金融分析技术的一种补充和扩展。相比传统方法，基于 TDA 去噪的方法不仅提高了数据分析的效率和准确性，而且有助于优化投资组合构建并减小风险。本研究不仅为 TDA 在金融市场的应用奠定了基础，也为探索其在金融市场动态性背景下的适应性提供了初步见解。

尽管本文已介绍 TDA 的基本理论，但其在金融领域应用的理论深度仍有待进一步拓展，尤其是结合金融市场动态性对 TDA 方法的适应性进行深入分析。未来，可以将本文方法应用于绿色金融、可持续金融等新兴且至关重要的领域，以有效预防金融危机或减轻其蔓延，为金融市场的繁荣与可持续发展贡献更多的智慧与力量。

参 考 文 献

[1] DAVID C S, EDELSBRUNNER H, JOHN H. Stability of Persistence Diagrams[J]. Discrete & Computational Geometry, 2007, 37(1): 103-120.

[2] GIDEA M, KATZ Y. Topological Data Analysis of Financial Time Series: Landscapes of Crashes[J]. Physica A, 2018, 491: 820-834.

[3] FANG X S. Research on Stock Market Crashes and Stock Price Volatility Based on Topological Data Analysis and Machine Learning[D]. Hangzhou: Zhejiang University, 2021.

[4] KARAN A, KAYGUN A. Time Series Classification Via Topological Data Analysis[J]. Expert Systems With Applications, 2021, 183, 115326.

[5] KATZ Y A, BIEM A. Time-Resolved Topological Data Analysis of Market Instabilities[J]. Physica A: Statistical Mechanics and Its Applications, 2021, 571: 125816.

[6] DE GREGORIO A, IACUS S M. Clustering of Discretely Observed Diffusion Process[J]. Computational Statistics & Data Analysis, 2010, 54(2): 598-606.

[7] ZHAO C. Comparative Study on Clustering Analysis Methods of Financial Time Series: An Empirical Analysis Based on Stock Return Rates of Listed Companies[J]. Times Finance, 2013 (17): 56-59.

[8] LI X F. Research on Cluster Analysis Based on Time Series Characteristics In Margin Trading and Short Selling as Well as A-Share Transactions[D]. Jinan: Shandong University, 2017.

[9] GIDEA M, KATZ Y. Topological Data Analysis of Financial Time Series: Landscapes of Crashes[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 491: 820-834.

[10] MAJUMDAR S, LAHA A K. Clustering and Classification of Time Series Using Topological Data Analysis with Applications to Finance[J]. Expert Systems With Applications, 2020, 162: 113868.

[11] GOEL A, PASRICHA P, MEHRA A. Topological Data Analysis in Investment Decisions[J]. Expert Systems With Applications, 2020, 147: 113222.

[12] PAN Y, CHENG X J. Optimal Portfolio Plan Based on Clustering Grouping Method to Modify Covariance Matrix[J]. Journal of the University of Science and Technology of China, 2014, 44(3): 244-247, 256.

[13] DE LUCA G, ZUCCOLOTTO P. Dynamic Tail Dependence Clustering of Financial Time Series[J]. Statistical Papers, 2017, 58: 641-657.

[14] LI X Y, WANG B H. Construction and Comparative Study of the Robust Sharpe Single-Index Model[J]. Journal of Applied Statistics and Management, 2020, 39(3): 544-555.

[15] ZHANG P, LI J X, CUI S L. Research on M-SV Portfolio Decision-Making Based on Machine Learning Prediction[J]. Mathematics in Practice and Theory, 2024, 54(5): 70-82.

[16] ZHANG P, ZHU S H, CUI S L. Three-Factor Mean-Variance Portfolio Optimization Based on BP Neural Network [J/OL].

[2025-02-20]. http://kns.cnki.net/kcms/detail/11.2018.o1.20250106.1749.032.html.

[17] BUBENIK P. Statistical Topological Data Analysis Using Persistence Landscapes[J]. Journal of Machine Learning Research, 2015,16(1):77-102.

[18] BUBENIK P,DLOTKO P. A Persistence Landscapes Toolbox for Topological Statistic[J]. Journal of Symbolic Computation, 2017,78:91-144.

[19] WANG R. Research on Portfolio Optimization Based on Data Denoising and Stock Price Prediction[D]. Dalian:Dongbei University of Finance and Economics,2023.

[20] MEI J. Research on Denoising of Financial Time Series Based on Multiscale Threshold Method[J]. Financial Economy, 2013(12):155-156.

[21] CAI X S,DAI J B,LI X N. Application of Median Filtering and Mean Filtering in Barcode Denoising[J]. Journal of Changchun Normal University(Natural Science Edition),2008(8):40-42.

[22] ZHENG J H. Research on Denoising Method of Ghost Imaging Based on Mean and Median Filtering[D]. Nanjing:Nanjing University of Posts and Telecommunications,2023.

[23] WANG Y,WAN X Y,TAO Y Z,et al. Collaborative Filtering Recommendation Algorithm Based on K-Medoids Item Clustering[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition),2017,29(4):521-526.

[24] CHEN Y,KEOGH E,HU B,et al. The UCR Time Series Classification Archive[EB/OL]. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[25] ZHANG Y,GE P S,CUI F L,et al. Obstacle Detection Based on Improved Euclidean Clustering for Lidar[J]. Journal of Dalian Nationalities University,2021,23(3):223-227.

[26] TAN Z L,YUAN H,WANG F H. Similarity Search Algorithm Based on DWT[J]. Statistics & Information Forum, 2023, 38(1):3-15.

[27] HE L,CHEN L,JI S S,et al. Anomaly Detection Method for Continuous Liquid Level Monitoring Data Based on K-Shape Clustering[J]. China Water & Wastewater,2023,39(11):56-61.

[28] JIANG W F. Research on Optimal Portfolio Based on China's Carbon Trading Sector in The Stock Market — An Empirical Study and Optimization of Markowitz Portfolio Theory Under Principal Component Analysis[J]. China Collective Economy, 2023(25):83-86.

LI Ruiyang, born in 2003, is a member of CCF (No. O3787G). His main research interests include data mining and deep learning.

QIAO Gaoxiu, born in 1982, Ph.D, associate professor. Her main research interests include time series forecasting, derivatives pricing, machine learning and energy economics.

(责任编辑:喻藜)