Pure Mathematics 理论数学, 2024, 14(10), 108-116
Published Online October 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.1410350

Hans汉斯

基于不同距离度量的K-Means算法在配对交易中的应用研究

朱　军，赵　伟

西南民族大学经济学院，四川　成都

收稿日期：2024年9月4日；录用日期：2024年10月6日；发布日期：2024年10月15日

摘　要

本研究探讨了K-Means聚类算法，在不同距离度量基础上对配对交易中两种期货合约的历史价差序列进行分类的应用。本文比较了欧式距离、曼哈顿距离、切比可夫距离和余弦相似度在价差序列分类中的应用效果。研究结果表明，相较于传统的欧式距离，余弦相似度能够更好地对价差序列进行聚类，在效果评测指标上表现更加优异。

关键词

配对交易，K-Means聚类，价差序列，距离度量

Research on the Application of K-Means Algorithm Based on Different Distance Metrics in Pairing Transactions

Jun Zhu, Wei Zhao

School of Economics, Southwest University for Nationalities, Chengdu Sichuan

Received: Sep. 4th, 2024; accepted: Oct. 6th, 2024; published: Oct. 15th, 2024

Abstract

This study explores the application of K-Means clustering algorithm to classify the historical spread sequences of two futures contracts in paired trading based on different distance measures. This article compares the application effects of Euclidean distance, Manhattan distance, Chebyshev distance, and cosine similarity in price difference sequence classification. The research results indicate

文章引用: 朱军, 赵伟. 基于不同距离度量的 K-Means 算法在配对交易中的应用研究[J]. 理论数学, 2024, 14(10): 108-116. DOI: 10.12677/pm.2024.1410350

that compared to traditional Euclidean distance, cosine similarity can better cluster price difference sequences and perform better in performance evaluation indicators.

Keywords

Pair Trading, K-Means Clustering, Spread Sequence, Distance Metrics

1. 引言

配对交易是一种中性投资策略，其核心思想是利用两个相关资产价格之间的暂时偏离来获取利润。具体来说是通过买入价格相对被低估的资产，卖出价格相对被高估的资产，待价差恢复到均衡水平时反向平仓获得收益。相似资产的价格相对高低的判断是配对交易决策的核心内容，国内外学者多角度进行相关研究，主要研究方向包括筛选出配对效果更好的资产和最优化开平仓决策参数。人工智能(AI)中机器学习算法的广泛使用为配对交易开平仓决策参数的优化提供了新的研究视角和方法工具，其中聚类算法能够在无预先定义类别标签条件下识别配对交易数据的内在结构，学习配对交易历史价差序列特征，助力配对交易参数优化。具体的，对配对交易产品的历史价差序列进行聚类分析可以识别出具有相似价格行为模式的价差序列，揭示不同配对产品价格间的潜在关系，同时基于不同价差行为模式为每一类别设置更适合的配对交易参数，如开仓阈值、止损点，从而提高策略的适应性和盈利能力。

在机器学习众多聚类算法中，K-Means 聚类算法因为其简单性、高效率、适应性强和结果易解释性广泛应用于探索性的数据分析，为包括配对交易在内的交易策略提供了优化方法和工具。K-Means 聚类算法以距离来度量数据之间的相似性，数据之间的距离越小，相似度就越高，它们就越有可能在同一个类簇。K-Means 算法有四大距离测度，其中我们常用欧式距离，其他还有曼哈顿距离、切比雪夫距离和余弦相似度等距离测度。大部分学者在使用 K-Means 聚类算法处理复杂的金融时间序列数据时，往往忽略了距离测度差异这一问题，相关研究结论鲁棒性有待商榷。基于以上研究背景，本文立足于配对交易数据特征分类这一主题，从理论和实证两大视角去比较不同距离测度下的 K-Means 算法差异，寻找适用于配对交易的最优距离测度，同时也为机器学习算法的类似研究提供启发。

2. 文献综述

2.1. K-Means 聚类原理研究

K-Means 聚类算法于 1967 年首次提出，Macqueen 提出了 K-Means 算法的概念[1]，算法思想和实施步骤具体为：首先从数据集中随机选取 K 个初始聚类中心 $C_i\left(i \le 1 \le K\right)$，计算其余数据对象与聚类中心 C_i 的欧式距离，找出离目标数据对象最近的聚类中心 C_i，并将数据对象分配到聚类中心 C_i 所对应的簇；然后计算每个簇中数据对象的平均值作为新的聚类中心，替换上步中的聚类中心 C_i，进行距离计算迭代，直到聚类中心不再变化或达到最大的迭代次数时停止[2]。

为了避免聚类算法陷入局部最优的可能性，Arthur 等人于 2007 年提出了一种改进的 K-Means 初始化的方法，通过增加初始聚类中心之间的距离能够提高聚类结果的质量[3]。Sculley 等人于 2010 年提出了 MiniBatch-Kmeans 算法[4]，这种算法是一种适用于大规模数据的 K-Means 的变体，与传统 K-Means

聚类算法相比，MiniBatch-Kmeans 算法可以在每次迭代中仅使用小批量数据进行更新，减少计算时间，同时保持聚类结果的合理性。

2.2. 距离度量相关研究

在聚类算法中，相似性度量是聚类算法中的一个基础问题。相似性度量可以用来计算两个数据对象之间的相似度或距离，从而对数据进行聚类分析[5]。在绝大部分应用中，使用最广泛的距离度量是欧式距离和曼哈顿距离，常用的还有考虑样本之间属性的马氏距离[6]。针对距离度量的改进，许多学者已经做了大量研究[7]-[10]。而在 K-Means 算法中，有许多的距离度量方式：欧式距离、曼哈顿距离、余弦相似度、杰卡德相似系数、马氏距离等等。近年来，研究者们在 K-Means 算法的距离度量方面进行了大量探索，旨在提高聚类的准确性和效率。高新等人于 2020 年提出一种优化中心点选取的改进 K-Means 聚类算法，根据数据对象与数据集中其他数据对象第 t 近邻的欧式距离来确定领域参数[11]。吴建国等于 2023 年提出了距离度量学习引导的加权聚类继承算法，引入马氏距离与基聚类权重学习，旨在提高剧烈集成算法性能[12]。

2.3. 配对交易及其聚类分析研究

配对交易是一种统计套利策略，其基本思想是寻找配对资产价差稳定在固定参数内的两种资产，在其价差扩大超出固定阈值时，买入相对被低估的资产，卖空相对被高估的资产，因此，寻找最优参数具有重要的意义[13]。杨艳军通过 GARCH 模型来预测国债期货价差序列残差的方差，提出基于开平仓参数的配对交易策略[14]。胡文伟提出一种基于强化学习的配对交易模型，将模型参数的确定方法改进为自适应的动态参数优化法[15]。

通过对价差进行分析可以使得投资者更有效地规避价格风险并获得收益[16]。杨阳和马超基于协整模型，运用黄金期权合约存续期内的 5 分钟高频收盘价数据来捕捉“价差偏离”及“价差回复均值”，实证表明相较于日收盘价数据，其套利机会更多[17]。Montana 在配对交易中，根据配对股票的价差偏离历史平均水平程度来设置交易参数，实证研究表明能够提高策略的投资回报率[18]。

聚类算法在配对交易中的应用十分广泛，主要应用于划分配对资产的价差序列和选择那些具有高度相关性的金融资产。钟悦于 2020 年通过 K-Means++模型，运用主成分分析降维后的财务指标进行聚类分析，提高配对股票池的质量[19]。冯彪于 2023 年利用 K-Means 聚类及随机森林算法对配对组合价差形态进行划分及预测，不同的价差形态设置不同的交易参数，从而探究设置多组交易参数的方法，以此提高配对交易的收益[20]。

3. 配对交易价差特征的 K-Means 距离测度对比分析

3.1. 距离测度定义

欧几里得提出的欧式距离是我们最常用的距离度量方法，诸多学者在将 K-Means 算法应用到配对交易时都是使用欧式距离进行距离测度[21]。欧式距离的数学表达式为：

$$D=\sqrt{\sum_{k=1}^{p}\left(X_{ik}-X_{jk}\right)^{2}} \tag{1}$$

其中 X_i 和 X_j 为两个样本观测点，X_{ik} 和 X_{jk} 分别是 X_i 和 X_j 的第 k 个输入变量的值。

曼哈顿距离首次由闵可夫斯基提出，是一种计算坐标轴上的绝对距离总和的计算方法。与欧式距离不同，并不是计算直线距离，而是计算实际的绝对距离，比欧式距离更能反映实际情况。该测度更加适用于在幅度变化上较为剧烈，但在结构上显示出相似性的历史价差序列。也就是说，它更适用于在变化

趋势上相似但在变化幅度上不一致的序列特别有用。其数学表达式为：

$$D=\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right| \tag{2}$$

切比雪夫距离是向量空间中的一种度量，两个点之间的距离定义为其各座标数值的最大值。该距离的特点是取各个维度差的最大值作为距离，强调了最大偏差。在配对交易中，如果关心的是历史价差序列在某个时期内的最大波动，那么切比雪夫距离是一个更好的选择，他可以在极端的市场条件下识别出表现相似的历史价差序列。其数学表达式为：

$$D=\max\left(\left|X_{ik}-X_{jk}\right|\right) \tag{3}$$

余弦相似度是度量两个向量在方向上的相似度，忽略它们的大小。该距离能够度量高维度的数据，适用于不需要考虑向量大小的场景。余弦相似度非常适用于分析价差变化的方向或趋势，因为它衡量的是序列之间的角度而非距离。其特点就是可以忽略价差的绝对幅度差异，专注于趋势的相似性。其数学表达式为：

$$D=\frac{\sum_{k=1}^{p}\left(X_{ik}X_{jk}\right)}{\sqrt{\left(\sum_{k=1}^{p}X_{ik}^{2}\right)\left(\sum_{k=1}^{p}X_{jk}^{2}\right)}} \tag{4}$$

根据上述定义，我们在表 1 中对比了四种距离的特点。

Table 1. Comparison of different distances
表 1. 不同距离比较

距离名称	数学表达式	特点
欧式距离	$D=\sqrt{\sum_{k=1}^{p}\left(X_{ik}-X_{jk}\right)^{2}}$	直观反映价差变动幅度
曼哈顿距离	$D=\sum_{k=1}^{p}\left\|X_{ik}-X_{jk}\right\|$	反映实际距离，适用于变化趋势相似的序列
切比雪夫距离	$D=\max\left(\left\|X_{ik}-X_{jk}\right\|\right)$	强调最大偏差，适用于极端市场条件
余弦相似度	$D=\frac{\sum_{k=1}^{p}\left(X_{ik}X_{jk}\right)}{\sqrt{\left(\sum_{k=1}^{p}X_{ik}^{2}\right)\left(\sum_{k=1}^{p}X_{jk}^{2}\right)}}$	反映两种向量的方向相似性而非大小

3.2. K-Means 算法距离测度比较分析

假设存在两种金融证券 A 和 B，其分钟级别的价格序列表达式为$\{P_A\}$、$\{P_B\}$，如果两者的价格时间序列存在如下的长期均衡关系：

$$P_A=\alpha+\beta P_B+\varepsilon_t \tag{5}$$

其中，β 为协整系数，对冲比例，ε_t 为残差序列，价差序列，α 为常数项。从上式中提取价差序列为：

$$spread=P_A-\alpha-\beta P_B \tag{6}$$

我们对历史价差序列进行 K-Means 聚类分析，首先需要对价差序列的特征进行筛选，然后根据特征之间的距离来评估输入价差序列间的相似程度，最后采用成熟的聚类算法进行具体划分。在进行特征选择之前，我们需要考虑价差序列的以下特点：一是趋势变化，即两种资产的价差有可能会随着时间升高也有可能随着时间下降；第二是幅度变化，即两种资产的价差大小有可能随着时间变大或变小；三是波

动特性，价差波动的频率和幅度会受到市场条件或者资产的基本面影响。

假设我们选择价差序列的均值和方差作为聚类的特征指标，根据每日的时间点 $t=1,2,\cdots,n$，分别计算出期货 A、期货 B 每日的价格平均值：

$$\overline{P_A}=\frac{1}{n}\sum_{t=1}^{n}P_A(t) \tag{7}$$

$$\overline{P_B}=\frac{1}{n}\sum_{t=1}^{n}P_B(t) \tag{8}$$

带入价差表达式中，可得：

$$\overline{\text{spread}}=\overline{P_A}-\alpha-\beta\overline{P_B} \tag{9}$$

接下来，我们分别计算出 P_A 和 P_B 的方差以及协方差：

$$var(P_A)=\frac{1}{n-1}\sum_{t=1}^{n}\left(P_A(t)-\overline{P_A}\right)^2 \tag{10}$$

$$var(P_B)=\frac{1}{n-1}\sum_{t=1}^{n}\left(P_B(t)-\overline{P_B}\right)^2 \tag{11}$$

$$cov(P_A,P_B)=\frac{1}{n-1}\sum_{t=1}^{n}\left(P_A(t)-\overline{P_A}\right)\left(P_B(t)-\overline{P_B}\right) \tag{12}$$

假设 $Var(P_A)=\sigma_A^2, Var(P_B)=\sigma_B^2, Cov(P_A,P_B)=\sigma_{AB}$，根据方差的性质，价差序列的方差为：

$$\begin{aligned}\text{var}(\text{spread})&=\text{var}(P_A-\alpha-\beta P_B)=\text{var}(P_A)+\beta^2\,\text{var}(P_B)-2\beta\,\text{cov}(P_A,P_B)\\&=\sigma_A^2+\beta^2\sigma_B^2-2\beta\sigma_{AB}\end{aligned} \tag{13}$$

我们把每一天的价差序列的均值记为 μ，方差记为 σ，并将价差序列的均值和方差表示为一个二维向量 $X=(\mu,\sigma)$，假设 $X_t=(\mu_t,\sigma_t)$ 为第 t 天的数据，$C_t=(\mu_k,\sigma_k)$ 为第 k 个簇心。对于两个向量 X_t 和 C_t，代入各种距离公式中：

欧式距离：

$$d(X_t,C_k)=\sqrt{(\mu_t-\mu_k)^2+(\sigma_t-\sigma_k)^2} \tag{14}$$

曼哈顿距离：

$$d(X_t,C_k)=|\mu_t-\mu_k|+|\sigma_t-\sigma_k| \tag{15}$$

切比雪夫距离：

$$d(X_t,C_k)=\max(|\mu_t-\mu_k|,|\sigma_t-\sigma_k|) \tag{16}$$

余弦相似度距离：

$$d(X_t,C_k)=\frac{\mu_t\mu_k+\sigma_t\sigma_k}{\sqrt{\mu_t^2+\sigma_t^2}\sqrt{\mu_k^2+\sigma_k^2}} \tag{17}$$

推导显示，欧式距离中，每一个特征指标的平方和会被开方，因此较大的差异会被放大。如果数据中存在异常值则会显著影响距离计算，影响聚类结果偏离。曼哈顿距离简单地累加各种特征指标的绝对值差异，如果特征值范围相差较大时，某个特征的较大差异依然会主导距离计算，导致聚类结果偏离。切比雪夫距离主要关注最大的差异，会忽略其他特征指标。考虑到金融时间序列数据的特性，传统的距离度量方法如欧式距离、曼哈顿距离和切比雪夫距离虽然直观但各有局限。相比之下，余弦相似度的计算公式中，计算的是两个特征之间在方向上的相似性，而不考虑它们的大小。这种距离的优点是不受量

纲的影响，因为通过对每个特征的模型进行归一化。基于夹角的计算，余弦相似度均等地考虑了每个特征的影响，所以每个特征指标对聚类结果的影响被过分放大，这与欧式距离放大较大的特征差异、曼哈顿距离将特征差异简单相加和切比雪夫距离只关注最大的特征差异形成对比。综合比较这些距离，余弦相似度更适合用于分析和识别价格趋势的相似性，得到更好的聚类结果。

3.3. 配对交易价差序列特征的 K-Means 聚类过程

K-Means 聚类算法基于距离将相似的样本划分为一类。该算法的核心思想为：随机选择 k 个样本点作为初始簇心，然后计算每一个样本点到簇心的距离，根据最小距离原则，将样本划分到距离较近的簇里。所有样本点都划分到簇中，计算每个簇的新簇心，通常是簇类所有样本点的均值。如果两个簇心相等，那么输出聚类结果。如果两个簇心不相等，则根据新的簇心重新进行划分，再次计算簇心，直到两个簇心相等，输出聚类结果。

具体地，将 K-Means 聚类算法应用到配对交易价差序列划分的主要流程为：

1) 将配对资产的价差序列数据划分为多段等时间段的长度价差序列，形成样本集 $S=\{\text{Spread}_1,\text{Spread}_2,\text{Spread}_3,\cdots,\text{Spread}_t\}$。

2) 针对样本集 S 中的每个列向量，假设以均值和方差作为特征指标。构建价差序列形态的特征指标，生成价差形态的特征指标样本集 $S_t=\{A_1,A_2,A_3,\cdots,A_t\}$。

3) 从特征指标样本集 S_t 中随机选取 k 个样本作为簇心，记为 $C=\{C_1,C_2,C_3,\cdots,C_t\}$。利用距离公式计算 $D_{tk}=\text{distance}(A_t,C_k)$，并根据最小距离原则将 S_t 中各样本划分至距离最小的簇心。最终将 S_t 划分为 k 个簇，并计算各簇的新簇心 $D=\{D_1,D_2,D_3,\cdots,D_k\}$。

4) 若新簇心 D 与之前的簇心 C 相等(即 $D=C$)，则输出价差划分结果，即 k 个分类。若新簇心 D 与之前的簇心 C 不相等(即 $D\neq C$)，则以 D 为簇心重新计算 S_t 到簇心的距离，再划分出 k 个簇，计算新簇的簇心 E。比较 E 与 D 是否相等，相等则输出划分好的 k 个簇，不相等则循环上述操作，直至划分簇心与新簇心相同，最终输出聚类结果 k 个簇。

我们得到聚类结果后，需要对聚类划分结果进行有效性评价，聚类评价指标是用来评估聚类结果质量的指标，轮廓系数为评价 K-Means 聚类算法较为常见的评价指标。轮廓系数的计算公式为：

$$S(i)=\frac{B(i)-A(i)}{\max\left(A(i),B(i)\right)} \tag{18}$$

$A(i)$：表示样本 i 与其同一类的样本中所有点之间距离的平均距离，即样本与其同一类样本的相似度。

$B(i)$：表示样本 i 与另一个最近的一类中所有样本点之间的平均距离，即样本与最近另一个类中样本点的相似度。

根据上述轮廓系数的公式，轮廓系数范围在−1 与 1 之间。如果轮廓系数越大且越接近于 1，说明样本与自己所在的类的样本越相似，与其他类的样本越不相似。轮廓系数为负数时，这说明样本与其他类的样本更相似。当轮廓系数为 0 时，说明样本与其所在类与其他类样本相似度相同，属于同一个类别。因此，轮廓系数越接近 1 越好，越接近−1 越差。

4. 基于蒙特卡罗模拟的不同距离测度比较分析

4.1. 蒙特卡罗模拟

我们通过蒙特卡罗模拟方法比较配对价差序列在不同的距离度量下的 K-Means 聚类差异，计算平均

正确率来评估每种距离度量的聚类效果。具体步骤如下：

一、模拟生成价差序列。利用 Python 语言生成模拟的 100 天的价差数据，其中包含三个不同波动率阶段的价差序列：前 30 天为低波动率，接下来的 30 天为中波动率，最后 40 天为高波动率。

二、价差序列特征计算以及定义真实标签。计算出生成的价差数据的平均值和方差，并根据原始波动率分段生成真实标签：前 30 天为标签 0，接下来的 30 天为标签 1，最后 40 天为标签 2。

三、K-Means 聚类。对模拟的价差数据进行 K-Means 聚类，通过不同的距离进行聚类，并计算每种距离下的聚类的准确率。

四、分类结果比较分析。我们通过平均正确率来评估不同距离度量的聚类结果，平均准确率越高表示聚类的结果越准确。最后，为了统计验证聚类算法的一致性和效果，我们重复以上整个过程 100 次，并计算不同距离度量下正确率的平均值。下图 1 为蒙特卡罗模拟流程图。

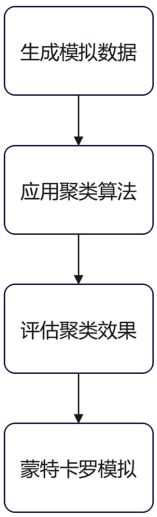

Figure 1. Monte Carlo simulation flowchart
图 1. 蒙特卡罗模拟流程图

4.2. 实证结果分析

表 2 汇总了 K-Means 聚类算法在各种距离度量下，对模拟的价差序列进行 100 次聚类的平均聚类效果评估。观察这些结果，我们可以发现“余弦相似度”能够达到最好的聚类效果，平均正确率为 0.650。表现最差的是曼哈顿距离，平均正确率为 0.597。相对而言，切比雪夫距离和欧式距离表现相近，平均正确率分别为 0.600 和 0.609。

Table 2. Clustering effects at different distances
表 2. 不同距离的聚类效果

不同距离	平均准确率
欧式距离	0.609
曼哈顿距离	0.597
切比雪夫距离	0.600
余弦相似度	0.650

5. 实证比较分析

为了验证余弦相似度在实际金融数据中的表现，基于米筐平台，我们选取大连期货交易所的黄大豆(B)和菜籽粕(RM)作为实证研究对象，相关系数为0.92，高相关系数的选择确保了配对资产在统计意义上的联动性，提高了配对交易策略的有效性。选择2023年5月28日至2024年2月28日的分钟级别数据，总计9个月的数据。选用这段时间内的数据是为了覆盖一个完整的市场周期，能够更全面地反映市场变化的特征。采样频率为每分钟一次。高频数据能够更精细地捕捉价差序列的波动特性，有助于提高聚类结果的准确性和精确度。

通过对配对资产的收盘价进行ADF检验，收盘价序列不平稳，进行一阶差分后，收盘价价格序列满足一阶单整。然后使用OLS对配对资产的收盘价格序列进行回归分析，以黄大豆B收盘价序列做被解释变量，以菜籽粕RM收盘价序列做解释变量，OLS回归结果如图2所示。

OLS Regression Results

Dep. Variable:	close	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	3.263e+05
Date:	Sat, 11 May 2024	Prob (F-statistic):	0.00
Time:	18:12:21	Log-Likelihood:	-3.6520e+05
No. Observations:	62655	AIC:	7.304e+05
Df Residuals:	62653	BIC:	7.304e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1000.1244	5.111	195.693	0.000	990.107	1010.141
close	1.0356	0.002	571.234	0.000	1.032	1.039

Omnibus:	3800.786	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4445.079
Skew:	0.643	Prob(JB):	0.00
Kurtosis:	2.779	Cond. No.	4.38e+04

Figure 2. OLS regression results
图2. OLS回归结果

得到价差序列的一元线性回归方程为：

$$\text{spread} = P_B - 1.03P_{RM} - 1000.12 \tag{19}$$

为保证实证的严谨性，我们依然选择方差和均值作为特征指标，使用四种不同距离进行K-Means聚类，设定K值为3，最后使用轮廓系数指标来进行评价聚类结果，轮廓系数结果如表3所示。

Table 3. Clustering effect of different distances in actual futures
表3. 不同距离在实际期货中的聚类效果

不同距离	聚类效果
欧式距离	0.454
曼哈顿距离	0.432
切比雪夫距离	0.437
余弦相似度	0.761

在实际数据的不同距离聚类比较研究中，可以看到余弦相似度对历史价差序列能够取得最好的聚类

效果，其轮廓系数为 0.761，其次是欧式距离，其轮廓系数为 0.454。曼哈顿距离与切比雪夫距离的轮廓系数相差不大，其中曼哈顿距离最差，其轮廓系数仅为 0.432。实际数据的聚类结果与蒙特卡罗模拟结果相差不大。由此可见，余弦相似度对比传统的欧式距离、曼哈顿距离以及切比雪夫距离，具有最优的聚类效果，特别是在历史价差序列的数据集上。

6. 结论与展望

本文研究了金融时间序列分析中的价差序列聚类问题，并提出了使用余弦相似度作为距离度量的方法。通过蒙特卡洛模拟和实际数据应用，我们验证了余弦相似度在 K-Means 聚类算法中的有效性。实验结果表明，余弦相似度能够更准确地捕捉到金融数据的特征，提高聚类的准确性。

然而，本文的研究仍有一定的局限性。例如，我们只考虑了单一的聚类算法(K-Means)和距离度量方法，未来可以尝试其他聚类算法和距离度量方法，以找到更适合金融数据的方法。此外，本文只关注了价差序列的均值和方差，未来还可以研究如何结合其他特征指标进行综合分析。

参考文献

[1] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Le Cam, L.M. and Neyman, J., Eds., *Proceedings of the* 5*th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 281-297.

[2] 王紫涵. 聚类分析中 K-Means 聚类算法的改进与新聚类有效性指标研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2022.

[3] Arthur, D. and Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, 7-9 January 2007, 1027-1035.

[4] Sculley, D. (2010) Web-Scale K-Means Clustering. *Proceedings of the* 19*th International Conference on World Wide Web*, Raleigh, 26-30 April 2010, 1177-1178. https://doi.org/10.1145/1772690.1772862

[5] 栗庆杰. 启发式 k-means 聚类算法的改进与应用研究[D]: [硕士学位论文]. 大连: 大连交通大学, 2023.

[6] 邵俊健, 王士同. 高维数据的增量式聚类算法的距离度量选择研究[J]. 计算机工程与科学, 2019, 41(2): 214-223.

[7] Wu, B., Wang, L. and Xu, C. (2009) Possibilistic Clustering Using Non-Euclidean Distance. 2009 *Chinese Control and Decision Conference*, Guilin, 17-19 June 2009, 938-940. https://doi.org/10.1109/ccdc.2009.5191912

[8] 熊拥军, 刘卫国, 欧鹏杰. 模糊 C-均值聚类算法的优化[J]. 计算机工程与应用, 2015, 51(11): 124-128.

[9] Liu, W.-Y., Chen, Z.-W., Bai, P., Fang, S.-F. and Shi, Y. (2005) A Kind of Improved Method of Fuzzy Clustering. 2005 *International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August 2005, 2646-2649. https://doi.org/10.1109/icmlc.2005.1527391

[10] 朱兴晨. 距离测度优化的模糊聚类分析及应用[D]: [硕士学位论文]. 镇江: 江苏大学, 2023.

[11] 高新. 一种改进 K-Means 聚类算法与新的聚类有效性指标研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2020.

[12] 吴建国. 基于马氏距离度量的聚类集成算法研究[D]: [硕士学位论文]. 太原: 山西大学, 2023.

[13] 于晓雨, 毕秀春, 张曙光. 配对交易的最优阈值[J]. 中国科学技术大学学报, 2020, 50(6): 784-792.

[14] 杨艳军, 陈思岑. 基于高频数据的我国国债期货市场套利研究[J]. 财务与金融, 2018(2): 1-6.

[15] 胡文伟, 胡建强, 李湛, 等. 基于强化学习算法的自适应配对交易模型[J]. 管理科学, 2017, 30(2): 148-160.

[16] 吴丰. 基于多尺度与关系特征挖掘的期货价差预测方法研究[D]: [硕士学位论文]. 长沙: 中南大学, 2023.

[17] 杨阳, 马超. 基于配对交易的上期所黄金期权套利策略研究[J]. 投资研究, 2024, 43(4): 145-159.

[18] Montana, G., Triantafyllopoulos, K. and Tsagaris, T. (2009) Flexible Least Squares for Temporal Data Mining and Statistical Arbitrage. *Expert Systems with Applications*, **36**, 2819-2830. https://doi.org/10.1016/j.eswa.2008.01.062

[19] 钟锐. 基于 K-Means++与 Adaboost 弹性网络的多股票配对交易策略设计[D]: [硕士学位论文]. 上海: 上海师范大学, 2020.

[20] 冯彪. 基于机器学习的商品期货配对交易参数优化研究[D]: [硕士学位论文]. 成都: 西南民族大学, 2023.

[21] 吴胜义, 王义贵, 王飞, 等. 基于多距离度量 kNN 模型的森林蓄积量反演[J]. 中南林业科技大学学报, 2023, 43(2): 10-18.