Advances in Applied Mathematics 应用数学进展, 2025, 14(1), 318-330
Published Online January 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.141032

Hans汉斯

面向集值型数据的无监督聚类方法及其应用

王　旭，马丽涛*

河北工程大学数理科学与工程学院，河北　邯郸

收稿日期：2024年12月26日；录用日期：2025年1月18日；发布日期：2025年1月29日

摘　要

分类问题是数据挖掘、机器学习等领域的基础性问题之一，然而多数分类方法仅关注向量值样本的分类问题，而对于实际中广泛存在的集值型数据样本的分类关注较少。本文提出了一种基于Wasserstein距离的无监督聚类算法(Wk-means)，利用熵正则最优传输模型度量集值型数据点之间的距离，并结合聚类的思想设计了一个可用于集值型数据的Wk-means聚类方法。为验证方法的有效性，本文首先在几个公开数据集上进行了实验，结果证实了Wk-means在多样本、多类别、多特征的集值型数据中表现优异，并且通过统计检验表明本文算法与其他算法存在显著差异。随后将本文方法实际应用于滏阳河水质数据集，结果同样表明相比传统的数据聚类算法，Wk-means能够更准确地划分水质类别，且运行效率更高。本文提出的Wk-means算法在集值型水质数据的分类任务中表现出色，能够为环境监测和管理提供有价值的决策支持。

关键词

集值型数据，分类问题，Wasserstein距离，最优传输，水质分类

Unsupervised Clustering Method for Set-Valued Data and Its Application

Xu Wang, Litao Ma*

School of Mathematics and Physics, Hebei University of Engineering, Handan Hebei

Received: Dec. 26th, 2024; accepted: Jan. 18th, 2025; published: Jan. 29th, 2025

Abstract

Classification is one of the basic problems in data mining, machine learning and other fields. However, most classification methods only focus on the vector-valued samples, while paying less attention

*通讯作者。

文章引用: 王旭, 马丽涛. 面向集值型数据的无监督聚类方法及其应用[J]. 应用数学进展, 2025, 14(1): 318-330.
DOI: 10.12677/aam.2025.141032

to the classification of set-valued data samples that are widely existed in practice. This paper proposes an unsupervised clustering algorithm (Wk-means) based on Wasserstein distance. Combined with the idea of clustering, Wk-means can be used for set-valued samples, in which the entropy-regularized optimal transport model is used to measure the distance between set-valued samples. In order to verify the effectiveness of Wk-means, experiments are conducted firstly on several public data sets. The results confirm the excellent performance of Wk-means in set-valued data with multi-sample, multi-category, and multi-feature. Moreover, the statistical test show that Wk-means is significantly different from other algorithms. Wk-means is then applied to the Fuyang River water quality data set. The results also show that Wk-means can classify water quality categories more accurately and effectively than the traditional data clustering algorithm. The Wk-means algorithm proposed in this paper performs well in the classification task of set-valued water quality data and can provide valuable decision support for environmental monitoring and management.

Keywords

Set-Valued Data, Classification Problem, Wasserstein Distance, Optimal Transport, Water Quality Classification

1. 引言

在数据挖掘、机器学习和模式识别等领域，分类技术扮演着核心角色。分类技术的核心目标在于构建一个能够根据数据集的特征将未知样本准确划分到预定义类别中的分类函数或模型，通常这个模型也被称作分类器。分类方法主要包括有监督分类、无监督分类、半监督分类。有监督分类是最常见的分类方法之一，它依赖于带有标签的训练数据来构建模型。在这种方法中，算法通过学习输入特征和输出标签之间的关系，以便能够对新的数据进行准确的预测。无监督分类是一种在没有预先标签或分类信息的情况下，将数据集中的样本划分为若干个组或类别的分类方法。半监督分类介于有监督和无监督分类之间，它使用一部分标注数据和大量未标注数据来训练模型，这种方法适用于标注数据难以获得或成本高昂的情况。

常见的分类算法有 K-近邻、决策树、支持向量机(SVM)等。K-近邻算法是一种基于实例的学习算法，其核心思想是对于一个新的输入实例，算法会在训练数据集中寻找与其最近的 K 个实例，然后根据这些最近邻的类别或输出值来预测新实例的类别或输出值。K-近邻算法简单直观，易于理解和实现，但其性能很大程度上依赖于 K 值的选择、距离度量和分类决策规则[1] [2]。决策树是一种监督学习算法，通过在特征空间中构建树状结构来进行决策。它通过递归地选择最优特征进行分割，构建出能够映射数据特征到决策结果的树模型。决策树简单直观，易于理解和解释，适用于分类和回归任务[3] [4]。然而，决策树容易过拟合，尤其是在面对复杂数据时，可能需要剪枝等技术来控制模型复杂度。支持向量机是一种强大的分类算法，可用于线性和非线性数据的分类。SVM 的核心思想是在特征空间中寻找一个最优的超平面，这个超平面能够最大化地分开不同类别的数据点，即最大化两类数据点之间的间隔。SVM 通过引入核技巧，能够处理非线性可分的数据。它在高维空间中表现良好，适用于复杂的分类问题[5] [6]。

在实际应用中，为了降低不确定性并提高测量结果的可靠性，常常会采用多次测量或重复实验的方

法[7]。例如在水质评价领域，由于季节性波动和区域差异，如降雨量变化导致污染物浓度波动或突发性污染事件，水质数据往往带有一定的不确定性[8]。为了减少这种不确定性，通常会进行多次重复测量[9]，从而获得多个特征向量来描述单个样本，进而产生了基于集合的分类学习任务。目前，处理这类基于集合的分类问题主要有两种方法：一种是通过计算原始数据的统计量[10] [11]，如均值、中位数等，将集合转换为向量形式以进行分析，这种方法在多种算法中得到了应用[12]-[14]。但集合被转换为向量时会丢失大量的原始信息，进而产生误差，影响分类结果。另一种方法是直接开发基于集合的分类器。例如，Arandjelovic 等[15]将每个集合建模为参数分布，并引入 K-L 散度来计算分布间的相似性。这种方法要求对参数分布进行估计，但在实际应用中，由于样本量通常不足以获得准确的估计，在很多场景并不适用。

为了克服现有算法处理集值型数据时存在的不足，本文将 Wasserstein 距离引入 K-means 聚类算法，以实现对集值型数据的分类。Wasserstein 距离源自最优传输问题[16]，作为一种衡量概率分布之间差异的方法，能够准确地捕捉集值型数据之间的关系，实现分类。同时，Wasserstein 距离对极端指标不敏感，不会因为个别极端值而影响分类结果，具有较好的鲁棒性。此外，Wasserstein 距离的计算本质上是一个优化问题，这使得它可以通过优化算法进行高效求解。

本文的创新点如下：

(1) 不同于传统的分类方法只适宜处理向量值数据，本文提出的基于 Wasserstein 距离的 K-means 聚类算法能够直接对集值型数据进行聚类处理，不需要计算原始数据的统计量，避免了原始信息的遗漏。

(2) 得益于熵正则项的特殊结构，利用 Sinkhorn-Knopp 算法有效提升了算法的效率。

(3) 在 UCI 数据集及邯郸市滏阳河水质数据中验证了算法处理集值型数据的有效性。

本文结构如下：第二章为预备知识，介绍了聚类分析、最优传输理论和 Sinkhorn-Knopp 算法的相关理论。第三章设计了 Wk-means 算法，并给出了算法的实现过程和流程图。第四章为实验部分，分别在 UCI 数据集和滏阳河水质数据集上进行了实验和对比。第五章为结论与展望，总结了全文并对未来进行展望。

2. 预备知识

2.1. 聚类分析

聚类分析(Cluster Analysis)是一种重要的无监督学习方法，其核心目标是将数据集中的样本划分为若干组(簇)，使得同一簇内的样本具有较高的相似度，而不同簇间的样本相似度较低。在数据挖掘、模式识别等领域得到广泛应用。

K-means 算法是最经典的聚类算法之一，该算法于 1967 年提出，以简单直观、易于实现的特点在实际应用中占据重要地位[17]。K-means 算法的核心思想是最小化每个点到其聚类中心的距离之和。基本步骤包括：

(1) 初始化：随机选择 K 个样本点作为初始聚类中心。

(2) 分配：将每个样本分配到距离最近的聚类中心所对应的簇。

(3) 更新：重新计算每个簇的中心点(各维特征的均值)。

(4) 迭代：重复步骤(2)和(3)，直至达到终止条件。

K-means 算法原理简单，时间复杂度较低，适合处理大规模数据，且算法可解释性强，但仍存在如对异常点敏感，可能陷入局部最优等不足。

2.2. 最优传输理论

最优传输(Optimal Transport，简称 OT)理论最初由法国数学家 Gaspard Monge 于 1781 年正式提出[18]。

Monge 问题的核心在于，在给定代价函数的情况下，寻找两个概率分布之间的最优传输映射，如图 1 所示。这一问题可以直观地理解为：如何以最经济的方式将一堆沙子从一个分布形态转移到另一个分布形态，即填补一个坑洞。Monge 问题的数学表述是寻找一个映射 T，使得传输成本最小化，同时满足质量守恒的条件，具体定义如下所示：

$$\min_{T}\left\{\sum_{i} c\left(x_i, T\left(x_i\right)\right): T_{\#}\alpha=\beta\right\} \tag{1}$$

其中 c 为代价函数，α 和 β 为两个概率测度，# 为 push forward 算子。

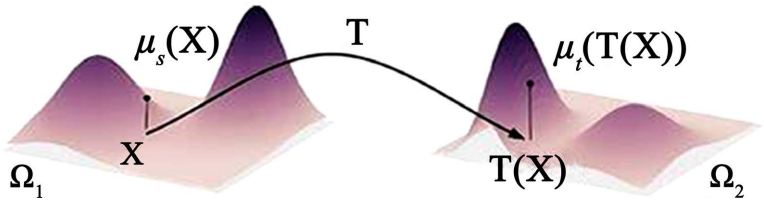

Figure 1. The optimal mapping between the two distributions
图 1. 两个分布之间的最优映射

苏联数学家 Leonid Kantorovich 在 1942 年提出了 Monge 问题的松弛形式[19]，即 Kantorovich 问题，并将其应用于经济问题的研究。Kantorovich 对 Monge 问题的主要改进在于，他将寻找最优传输映射的问题转变为寻找最优传输计划的问题。

不同于 Monge 问题中质量只能被移动不能被分割，Kantorovich 问题允许质量的分割和重组，即允许从一点到多点的质量传输。离散形式的 Kantorovich 问题可以描述为如下形式的线性优化问题：

$$L_C\left(a,b\right)=\min_{P\in U(a,b)}\langle C,P\rangle=\min_{P\in U(a,b)}\sum_{i,j}C_{i,j}P_{i,j} \tag{2}$$

其中 $U\left(a,b\right)=\left\{P\in R_{+}^{n\times m}: P1_m=a, P^{\mathrm{T}}1_n=b\right\}$，$a$ 和 b 为两个离散概率分布，C 为代价矩阵，常表示为 $C_{i,j}=\left\|x_i-y_j\right\|^2$。

Kantorovich 问题的数学表述是：给定两个概率测度 a 和 b，以及一个代价函数 C，寻找一个联合概率测度 P，使得该联合分布的边际分布分别为 a 和 b，并且传输成本最小化。

Kantorovich 问题的提出，不仅在理论上推进了最优传输理论的发展，而且在实际应用中也展现出了广泛的适用性[20]-[22]。重要的是，它使得最优传输问题的求解变得更加可行，尤其是在处理离散概率测度时，Kantorovich 问题可以转化为线性规划问题，从而可利用现有的算法进行有效求解。

Wasserstein 距离是最优传输理论中的一个重要概念，它定义了概率测度空间上的一个度量，并在数学、统计学、物理学和机器学习等多个领域中得到了应用[23]-[25]。对于两个概率分布 a 和 b，Wasserstein 距离 $W\left(a,b\right)$ 是将一个分布转换为另一个分布所需的最小成本。数学上，两个离散概率测度 a 和 b 的 p 阶 Wasserstein 距离由下式给出：

$$W_p\left(a,b\right)=L\left(a,b\right)^{\frac{1}{p}}=\min_{P\in U(a,b)}\left(\sum_{i,j}C_{i,j}P_{i,j}\right)^{\frac{1}{p}} \tag{3}$$

这里 $C_{i,j}=\left\|x_i-y_j\right\|^p$ 为代价矩阵，P 为指派矩阵。

2.3. Sinkhorn-Knopp 算法

Sinkhorn-Knopp 算法是一种用于求解最优传输问题的迭代算法，它通过在目标函数中加入熵正则化

项，将复杂的线性规划问题转化为在平滑可行域上的求解过程[26]-[28]。该算法的核心思想是在代价函数上加入熵正则化项，从而将问题转化为一个更易于求解的形式，即求解如下熵正则最优传输问题：

$$W_{Entropy}(a,b)=\min_{P\in U(a,b)}\langle C,P\rangle+\varepsilon H(P) \tag{4}$$

其中 $H(P)=\sum_i P_i\log P_i$ 。

Sinkhorn-Knopp 算法的数学表述是：给定两个概率向量 a 和 b，目标是找到一个矩阵 P，使得 P 的行和等于向量 a，P 的列和等于向量 b。算法的迭代步骤包括：

行归一化：对矩阵 P 的每一行进行缩放，使得行和接近目标向量 a；

列归一化：对矩阵 P 的每一列进行缩放，使得列和接近目标向量 b。

这个过程不断重复，逐步调整矩阵 P 的元素，使其行和和列和分别逼近 a 和 b。Sinkhorn-Knopp 算法的收敛性基于迭代过程的单调性、均匀缩放的性质和正则化项的引入，确保了算法在合理的迭代次数内收敛到一个满足约束条件的非负矩阵。

3. 基于 Wasserstein 距离的无监督聚类方法 Wk-Means

本文提出了一种基于 Wasserstein 距离的 K-means 聚类算法(Wk-means)，该算法在传统 K-means 的框架下引入 Wasserstein 度量，有效提升了算法处理概率分布数据的能力。算法 1 展示了主要计算过程，给定集值型数据集 $X\in\mathbb{R}^{N\times M\times d}$ ，参数 ε 和最大迭代次数 Maxiter，其中 N 表示样本个数，每个样本都是一个 $M\times d$ 维的矩阵。算法的核心目标是将这 N 个样本划分为 K 个类别，使得类内的 Wasserstein 距离之和最小。

算法 1：Wk-means 算法

1) 输入：集值型数据集 $X\in\mathbb{R}^{N\times M\times d}$，熵正则项参数 ε，最大迭代次数 Maxiter，类别个数 K
2) 输出：聚类结果
3) 初始化类中心 $X_k^c\ (k=1,2,\cdots,K)$；
4) while 迭代次数 < Maxiter :
5) for $i=1$ to N
6) 计算每一个样本点 X_i 与每个类中心 X_k^c 的 Wasserstein 距离： $d_k=W_{Entropy}(X_i,X_k^c)$；
7) 将样本 X_i 分配到最近的类： $Idx=\underset{k}{\operatorname{argmin}}(d_k)$；
8) end
9) 计算新的聚类中心，返回第 5 步；
10) end

接下来，为说明 Wk-means 算法的有效性，本文给出如下收敛性命题。

命题 1 Wk-means 算法收敛。

证明：给定如下损失函数

$$J=\sum_{k=1}^{K}\sum_{i=1}^{N}W^2(X_i,X_k^c) \tag{5}$$

为证明 Wk-means 算法收敛，首先证明随算法迭代损失函数 J 的函数值呈下降趋势。

Wk-means 算法主要分为两步：第一步，对每一个样本 X_i，根据其与类中心 X_i^c 的 Wasserstein 距离 $W(X_i,X_k^c)$ 来分配 X_i 的类别，第二步重新计算类中心 X_i^c 并再次分配 X_i 的类别。

第一步，计算每个样本与其类中心的 Wasserstein 距离，可得当前的损失函数值 J。

第二步，对每个样本而言，若 X_i 的类别不变，显然其与类中心 X_i^c 的 Wasserstein 距离 $W\left(X_i, X_k^c\right)$ 不变；若 X_i 类别改变，记新类别为 $X_k^{c^*}$，由于类中心应为距离所有同类样本点最近的样本，此时显然有

$$\begin{aligned} W^2\left(X_1, X_k^c\right) &\geq W^2\left(X_1, X_k^{c^*}\right) \\ W^2\left(X_2, X_k^c\right) &\geq W^2\left(X_2, X_k^{c^*}\right) \\ &\cdots \\ W^2\left(X_N, X_k^c\right) &\geq W^2\left(X_N, X_k^{c^*}\right) \end{aligned} \tag{6}$$

此时新的损失函数

$$\begin{aligned} J^* &= \sum_{k=1}^{K} \sum_{X_i \in R} W^2\left(X_i, X_k^{c^*}\right) \\ &= W^2\left(X_1, X_k^{c^*}\right) + W^2\left(X_2, X_k^{c^*}\right) + \cdots + W^2\left(X_N, X_k^{c^*}\right) \\ &\leq W^2\left(X_1, X_k^c\right) + W^2\left(X_2, X_k^c\right) + \cdots + W^2\left(X_N, X_k^c\right) \\ &= \sum_{k=1}^{K} \sum_{X_i \in R} W^2\left(X_i, X_k^c\right) \\ &= J \end{aligned} \tag{7}$$

故损失函数 J 单调不增。此外，将 N 个样本分配到 K 个类别的划分方式有限(K^N)，即 J 一定存在最小值。综上可知损失函数 J 单调不增且有下界，由单调有界定理可知损失函数 J 一定收敛，故 Wk-means 算法一定收敛，命题得证。

图 2 给出了 Wk-means 算法的具体流程，可以看出，Wk-means 算法的核心思想在于利用计算的 Wasserstein 距离矩阵进行聚类，从而实现对集值型数据的分类。

结合流程图，Wk-means 聚类算法主要包含以下 4 个步骤：

1) 在初始化阶段，首先输入集值型数据集 X 以及 Sinkhorn-Knopp 算法中的熵正则项参数 ε，最大迭代次数 Maxiter，类别个数 K，并确定算法的初始化中心点选取策略：随机选取(random)、均匀选取(uniform)和均值选取(mean)。

2) 在迭代优化阶段，算法交替执行样本分配和中心点更新两个步骤：首先计算每个样本到各类中心的 Wasserstein 距离，将样本分配给距离最近的类；然后重新计算每个类的中心点，通过不断迭代，实现聚类。

3) 算法的收敛基于目标函数值的变化，具体终止于以下情况：目标函数值不再显著下降；达到预设最大迭代次数。

4) 空聚类处理策略。针对实际应用中可能出现的空聚类问题，提供两种处理方案：直接删除(drop)或创建单点聚类(singleton)。前者简单地忽略空聚类，这可能导致最终得到的类别少于预期类别个数。而后者则从现有聚类中选择最远的点作为新的聚类中心。在后续实验中，本文采取了单点聚类策略，以保证聚类完整性。

为提高算法的计算效率，本方法采用 Sinkhorn-Knopp 算法，通过引入熵正则项，将原本的最优传输问题转化为一个可以快速求解的近似问题，显著提高算法计算效率。与传统的基于欧氏距离的 K-means 相比，Wk-means 通过考虑数据的概率分布特性，能够更好地捕获数据间的本质差别，对具有分布特征的数据具有更强的分类适应性，更重要的是，Wk-means 能够直接对集值型数据进行聚类处理，而不必从原始数据中提取统计量再进行聚类分析。

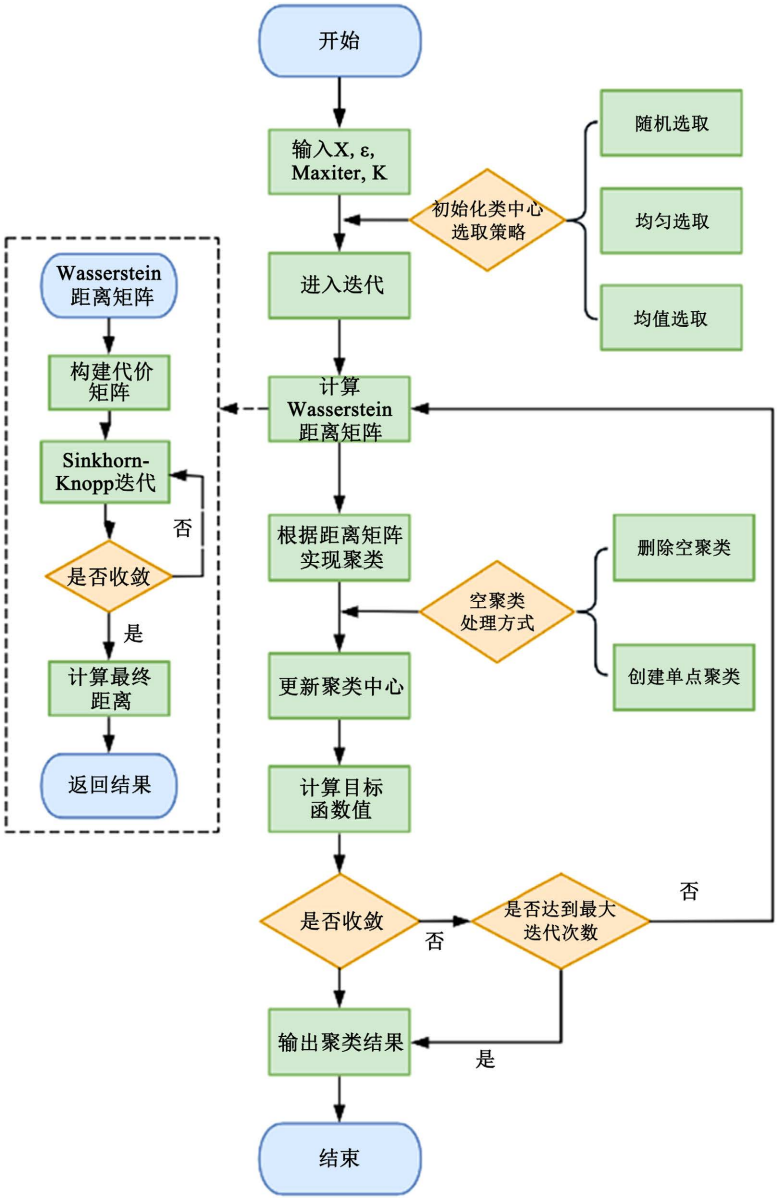

Figure 2. Flowchart of the Wk-means algorithm
图 2. Wk-means 算法流程图

4. 实验

本节将在 UCI 数据集和河北省邯郸市滏阳河水质数据中验证算法的有效性，实验是在一台拥有英特尔酷睿 i5-13500HX 处理器，16GB 内存的计算机上运行的，在 MATLAB R2023b 中进行仿真。算法中熵正则项参数 ε 见表 1，最大迭代次数 Maxiter 为 10,000，聚类个数 K 选取与数据集有关，在 UCI 数据集中，K 的个数与实际相同，在滏阳河数据中 K 的个数为水质等级个数。

4.1. UCI 数据集

为了对 Wk-means 算法进行检验，本文首先利用 UCI 数据集来生成新的集值型数据进行模型测试。具体生成方式为：向原始数据中的样本添加高斯噪声，使得每个样本生成 10 个包含高斯噪声的样本，从

而形成一个集值样本。所有的样本汇总就得到了新的集值型数据集。以 Yeast 数据集为例，该数据集包含 1484 个样本，即可生成 1484 个数据 A_i, $i=1,2,\cdots,1484$，每个数据 A_i 为一个集值样本，汇总起来就能得到一个包含 1484 个数据 $A_i=\{x_{ik},k=1,2,\cdots,10\}$, $i=1,2,\cdots,1484$ 的集值型数据集。数据集特征如表 1 所示，本文选取的数据集覆盖不同样本个数、特征个数及类别个数。

Table 1. UCI datasets and characteristics
表 1. UCI 数据集及特征

数据集	样本个数	特征个数	类别个数	ε
Glass	214	10	6	0.01
Iris	150	4	3	1
Seeds	210	7	3	0.01
Yeast	1484	8	10	0.01
Wine	178	13	3	1

我们对比了 Wk-means 方法与其他方法的表现，包括用于处理不确定数据的二阶锥规划(SOCP)方法[29]、基于正则化外壳的图像集协同表示和分类(RHISCRC) [30]和用于集合分类的稀疏近似最近点(SANP) [31]，以及基于核特征选择的广义预测集方法(GPS) [32]，结果如表 2~4 所示。从表 2 中可以看出，Wk-means 在多数的数据集上表现良好，尤其是在大样本数据集，如 Yeast 和 Glass 数据集上，Wk-means 预测精度要显著优于其他四个方法。而在运行效率方面，Wk-means 虽略低于 SOCP 方法，但与 RHISCRC 时间相近，且远优于 SANP 和 GPS 方法。表 4 则展示了 5 种方法预测的标准差，Wk-means 在多数数据集上标准差更小，即预测更稳定，波动更小。

为了能够更直观地展示不同方法之间的差异，我们进行了十折交叉验证，将数据分为 10 份分别计算准确率，然后将得到的结果进行 t 检验：

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{8}$$

其中 $\bar{x}_1$， s_1 分别为使用 Wk-means 方法计算的准确率的均值和标准差，$\bar{x}_2$， s_2 为其他方法计算的准确率的均值和标准差，$n_1=n_2=10$。

Table 2. Comparison of the accuracy of different methods on UCI datasets
表 2. 不同方法在 UCI 数据集上的准确率对比

数据集	Wk-means	SOCP	RHISCRC	SANP	GPS
Glass	**74.75**	68.3	69.5	56.1	63.2
Iris	93.3	**96.6**	82	68.7	71.4
Seeds	**96.44**	93.2	92.3	51.7	79.3
Yeast	**73.45**	59.2	52.9	67.9	55.8
Wine	95.46	**98.8**	98.2	59.4	76.8

Table 3. Comparison of running time of different methods on UCI datasets
表 3. 不同方法在 UCI 数据集上的运行时间对比

数据集	Wk-means	SOCP	RHISCRC	SANP	GPS
Glass	**6.03**	10.58	11.63	84.52	66.76
Iris	1.47	1.85	**1.46**	32.32	16.58
Seeds	6.82	**2.72**	8.2	46.47	33.61
Yeast	221.5	**110.2**	363.3	1831.66	986.52
Wine	5.85	**2.76**	13.76	31.19	14.19

Table 4. Comparison of standard deviations of different methods on UCI datasets
表 4. 不同方法在 UCI 数据集上的标准差对比

数据集	Wk-means	SOCP	RHISCRC	SANP	GPS
Glass	**0.0158**	0.0741	0.1012	0.0988	0.1435
Iris	0.0569	**0.0331**	0.0889	0.1525	0.1192
Seeds	0.0412	**0.0383**	0.0385	0.2089	0.0958
Yeast	**0.0153**	0.0177	0.0244	0.0416	0.0763
Wine	**0.0175**	0.0210	0.0290	0.1193	0.0847

Table 5. Comparison of different methods at 0.05 significance level
表 5. 0.05 显著性水平下的不同方法对比

数据集	Wk-means	SOCP	RHISCRC	SANP	GPS
Glass		1	0	1	1
Iris		0	1	1	1
Seeds	基准组	0	0	1	1
Yeast		1	1	1	1
Wine		0	0	1	1

表 5 展示了在显著性水平为 0.05 的情况下，不同方法间的对比结果。其中 1 表示有显著差异，0 表示没有显著差异。可以看出，在数据集 Glass、Iris 和 Yeast 上，Wk-means 不仅准确率远优于其他数据集，且与大多数方法均有显著性差异。而在部分小样本数据集如 Wine 上，Wk-means 虽然准确性没有达到最好，但与其差距很小，仍具有一定竞争性。

4.2. 滏阳河水质数据

滏阳河，如图 3 所示，源出河北省邯郸市峰峰矿区滏山南麓，流经邯郸、邢台、衡水，最终在沧州地区的献县与滹沱河汇合，形成子牙河，汇入渤海。全长 415 公里，流域面积达 2.8 万平方公里。滏阳河不仅是邯郸市的主要水源，也是该地区农业灌溉的重要依托，对于维持当地生态平衡和促进农业发展具有重要意义。

滏阳河的水质状况直接关系到邯郸市及其下游地区的水资源安全。近年来，随着工业化和城市化的快速发展，对滏阳河水质进行分类研究，不仅有助于监测和评估水质状况，还能为水资源管理和污染控制提供科学依据。

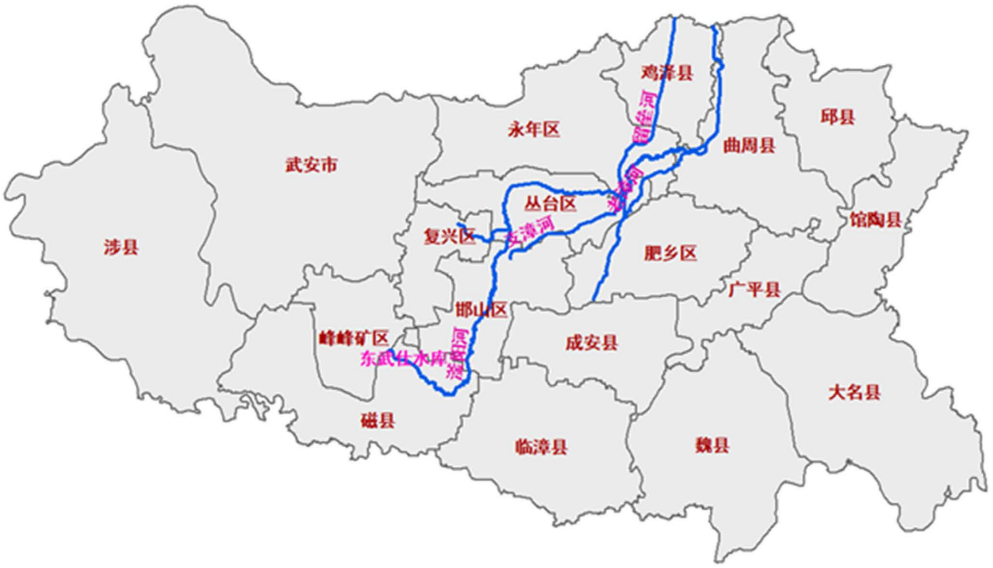

Figure 3. Fuyang river basin map
图 3. 滏阳河流域图

为对滏阳河水质进行分类，共设置 84 个采样点，且为了降低水质数据的不确定性，每个采样点测量 10 次，即得到 84 个数据 A_i，$i=1,2,\cdots,84$，分别属于 5 个不同等级(I, II, III, IV, V)。每个数据 A_i 由 10 个向量 $x_{i1},x_{i2},\cdots,x_{i10}$ 组成，即 $A_i=\{x_{i1},x_{i2},\cdots,x_{i10}\}$，$i=1,2,\cdots,84$。每个向量 x_{ik}，$k=1,2,\cdots,10$，包含 11 个特征：pH，DO，CODcr，NH-N，TN，N03，N02，PO4，D0，TP，CODmn。其中 DO、NH-N、TN、N03、N02、PO4、TP 和 CODmn 的测量单位均为“mg/L”，如表 6 所示，表中的数据已经过归一化处理。

Table 6. Some water quality data and characteristics
表 6. 部分水质数据及特征

样本		pH	DO	CODcr	NH-N	TN	N03	N02	PO4	D0	TP	CODmn	等级
A_1	x_{11}	0.61	0.64	0.29	0.04	0.16	0.31	0.03	0.04	0.50	0.17	0.28	I
	x_{12}	0.65	0.61	0.27	0.03	0.17	0.29	0.03	0.04	0.49	0.17	0.29	
	x_{13}	0.63	0.62	0.28	0.02	0.18	0.29	0.06	0.03	0.49	0.19	0.29	
	x_{14}	0.62	0.64	0.27	0.04	0.18	0.30	0.04	0.05	0.48	0.18	0.29	
	x_{15}	0.63	0.61	0.28	0.00	0.16	0.29	0.04	0.04	0.50	0.18	0.28	
	x_{16}	0.64	0.64	0.29	0.03	0.16	0.32	0.03	0.03	0.49	0.18	0.26	
	x_{17}	0.64	0.63	0.31	0.03	0.17	0.29	0.06	0.06	0.48	0.17	0.29	
	x_{18}	0.61	0.63	0.29	0.03	0.14	0.30	0.04	0.01	0.48	0.19	0.27	
	x_{19}	0.62	0.63	0.28	0.04	0.18	0.30	0.04	0.05	0.49	0.18	0.29	
	x_{110}	0.64	0.61	0.28	0.02	0.15	0.29	0.04	0.04	0.50	0.18	0.27	

续表

A_2	x_{21}	0.68	0.07	0.25	0.29	0.57	0.55	0.06	0.12	0.03	0.19	0.24	II
	x_{22}	0.67	0.06	0.24	0.28	0.57	0.56	0.06	0.11	0.03	0.20	0.25	
	x_{23}	0.68	0.06	0.25	0.26	0.58	0.57	0.05	0.12	0.03	0.20	0.24	
	x_{24}	0.67	0.07	0.27	0.29	0.58	0.58	0.06	0.13	0.06	0.18	0.25	
	x_{25}	0.67	0.06	0.22	0.28	0.57	0.56	0.05	0.13	0.04	0.17	0.25	
	x_{26}	0.67	0.06	0.25	0.29	0.59	0.58	0.04	0.14	0.02	0.20	0.26	
	x_{27}	0.67	0.04	0.25	0.30	0.56	0.56	0.04	0.10	0.03	0.18	0.23	
	x_{28}	0.68	0.06	0.26	0.28	0.56	0.57	0.06	0.14	0.02	0.16	0.24	
	x_{29}	0.66	0.06	0.25	0.28	0.56	0.56	0.06	0.13	0.03	0.20	0.27	
	x_{210}	0.69	0.05	0.23	0.29	0.57	0.55	0.04	0.12	0.05	0.19	0.25	
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 7. Comparison of results of different methods on water quality data
表 7. 不同方法在水质数据上的结果对比

	Wk-means	SOCP	RHISCRC	SANP	GPS
准确率	**69**	66	41	7	46
时间	**1.94**	7.74	3.46	4.4	8.73

与上一个实验相同，我们将所提的 Wk-means 方法与其他分类方法进行比较，表 7 展示了预测准确率和运行所需时间。从表 7 中可以看出，Wk-means 方法在预测精度和运行效率均达到了最优，具体表现在以下两个个方面：分类准确率有明显的提升，Wk-means 方法达到 69%的准确率，比次优的 SOCP 方法高出 3 个百分点，相比 RHISCRC、SANP 和 GPS 方法，准确率提升更为显著，分别高出 28、62、23 个百分点；计算效率显著提升，Wk-means 方法耗时 1.94 秒，相比 SOCP 方法(7.74 秒)和 GPS 方法(8.73)分别节省了约 75%和 78%的计算时间，较 RHISCRC、SANP 和 GPS 方法分别节省了 44%和 56%的运行时间。

5. 结论与展望

本文提出了一种基于 Wasserstein 距离的无监督聚类方法，用于集值型数据的分类问题，克服了传统欧氏距离无法处理集值型数据的局限性。在公开数据集的交叉验证实验中，尤其在大样本数据集(如 Yeast 和 Glass)上，Wk-means 分类准确率显著优于 SOCP、RHISCRC 和 SANP 等传统的集值数据处理方法，与新型集值数据处理方法 GPS 相比也有明显的提升，统计显著性检验进一步验证了算法性能的可靠性和稳定性。在滏阳河水质数据集上的实验表明，Wk-means 方法相比传统方法精度更高，运行时间更短。Wk-means 算法展现出优异的分类准确率，同时兼顾了计算效率，为处理大样本、复杂的集值型数据提供了一种新的技术解决方案。

在未来研究中，我们考虑在算法优化方面进行进一步深入研究，并将该方法推广到其他类型的集值数据中，如空气质量分类、土壤污染评估等。

基金项目

河北省中央引导地方科技专项项目(246Z1825G)；河北省高等学校科学研究项目(QN2020188,

ZD2020185)；河北省“三三三人才工程”资助项目(C20221026)。

参考文献

[1] 李久生, 盛姣, 纪鉴航, 等. 基于 KNN 算法研究遥感图像地块分割与提取[C]//国家新闻出版广电总局中国新闻文化促进会学术期刊专业委员会. 2021 年创新人才培养与可持续发展国际学术会议论文集(中文). 2021: 69-72.

[2] 张炎亮, 张超, 李静. 基于动态用户画像标签的 KNN 分类推荐算法研究[J]. 情报科学, 2020, 38(8): 11-15.

[3] 陈婷, 谢志龙. 基于改进决策树的不平衡数据集分类算法研究[J]. 计算机仿真, 2024, 41(8): 497-501.

[4] 韩彩娟. 基于决策树的制冷设备电子电路故障智能检测方法[J]. 电工技术, 2024(15): 140-142.

[5] 刘生富, 张鹏程, 周广宇, 等. 基于支持向量机与改进分水岭的红细胞识别算法研究[J]. 测试技术学报, 2022, 36(1): 48-53.

[6] 陶佳慧, 别雨轩, 顾约翰, 等. 基于多特征融合的 SVM 图像分类算法研究[J]. 上海航天(中英文), 2021, 38(S1): 98-102.

[7] DeSanto, J.B. and Sandwell, D.T. (2019) Meter-Scale Seafloor Geodetic Measurements Obtained from Repeated Multibeam Sidescan Surveys. *Marine Geodesy*, **42**, 491-506. https://doi.org/10.1080/01490419.2019.1661887

[8] 郝勇敢, 尚圆圆. 总磷总氮水质在线分析仪不确定度评定[J]. 广东化工, 2022, 49(1): 173-176.

[9] 陶莉. 水质监测中影响水质采样质量的因素及控制对策[J]. 清洗世界, 2022, 38(10): 109-111.

[10] Ashino, K., Kamiya, N., Zhou, X., Kato, H., Hara, T. and Fujita, H. (2024) Joint Segmentation of Sternocleidomastoid and Skeletal Muscles in Computed Tomography Images Using a Multiclass Learning Approach. *Radiological Physics and Technology*, **17**, 854-861. https://doi.org/10.1007/s12194-024-00839-1

[11] Yoneyama, J. (2012) Robust Sampled-Data Stabilization of Uncertain Fuzzy Systems via Input Delay Approach. *Information Sciences*, **198**, 169-176. https://doi.org/10.1016/j.ins.2012.02.007

[12] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/bf00994018

[13] Li, Y.F., Tsang, I.W., Kwok, J.T., *et al*. (2013) Convex and Scalable Weakly Labeled SVMs. *Machine Learning*, **14**, 2151-2188.

[14] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/bf00116251

[15] Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R. and Darrell, T. (2005) Face Recognition with Image Sets Using Manifold Density Divergence. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, 20-25 June 2005, 581-588. https://doi.org/10.1109/cvpr.2005.151

[16] Villani, C. (2009) Optimal Transport: Old and New. Springer.

[17] 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.

[18] Monge, G. (1781) Mémoire sur la théorie des déblais et des remblais. *Académie Royale des Sciences* (*France*), 666-704.

[19] Kantorovitch, L. (1958) On the Translocation of Masses. *Management Science*, **5**, 1-4. https://doi.org/10.1287/mnsc.5.1.1

[20] 范启哲. 基于最优传输与领域自适应的语义分割研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2023.

[21] 张浩. 基于深度学习和最优传输的地震数据重构与全波形反演[D]: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2021.

[22] 张琪. 基于最优运输理论的环境智适应无线网络研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2022.

[23] 张沙沙, 刘小弟, 张世涛. 基于 Wasserstein 测度的概率犹豫模糊聚类方法[J]. 模糊系统与数学, 2023, 37(6): 41-54.

[24] 晏远翔, 曹国, 张友强. 基于 Wasserstein 距离与生成对抗网络的高光谱图像分类[J]. 计算机系统应用, 2024, 33(2): 13-22.

[25] 苏连成, 朱娇娇, 郭高鑫, 等. 基于 XGBoost 和 Wasserstein 距离的风电机组塔架振动监测研究[J]. 太阳能学报, 2023, 44(1): 306-312.

[26] Altschuler, J., Niles-Weed, J. and Rigollet, P. (2017) Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. *Neural Information Processing Systems*, **2017**, 1964-1974.

[27] Schmitzer, B. (2019) Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *SIAM Journal on Scientific Computing*, **41**, A1443-A1481. https://doi.org/10.1137/16m1106018

[28] Lin, T., Ho, N. and Jordan, M. (2019) On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms. 2019 *International Conference on Machine Learning*, Long Beach, 9-15 June 2019, 3982-3991.

[29] Shivaswamy, P., Bhattacharyya, C. and Smola, A. (2006) Second Order Cone Programming Approaches for Handling Missing and Uncertain Data. *Machine Learning*, **7**, 1283-1314.

[30] Zhu, P., Zuo, W., Zhang, L., Shiu, S.C. and Zhang, D. (2014) Image Set-Based Collaborative Representation for Face Recognition. *IEEE Transactions on Information Forensics and Security*, **9**, 1120-1132. https://doi.org/10.1109/tifs.2014.2324277

[31] Hu, Y., Mian, A.S. and Owens, R. (2011) Sparse Approximated Nearest Points for Image Set Classification. 2011 *Conference on Computer Vision and Pattern Recognition*, Colorado Springs, 20-25 June 2011, 121-128. https://doi.org/10.1109/cvpr.2011.5995500

[32] Wang, Z. and Qiao, X. (2023) Set-Valued Classification with Out-of-Distribution Detection for Many Classes. *Journal of Machine Learning Research*, **24**, 1-39.