Hans Journal of Data Mining 数据挖掘, 2025, 15(2), 137-150
Published Online April 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.152012

Hans汉斯

基于阴影集的共享最邻近三支DBSCAN

李志聪*，闫　昆

哈尔滨师范大学计算机科学与信息工程学院，黑龙江 哈尔滨

收稿日期：2025年3月2日；录用日期：2025年3月25日；发布日期：2025年4月1日

摘　要

传统DBSCAN算法在处理数据时，将某些不确定的数据强制划分到某一类中往往容易带来决策风险。针对此问题，提出了基于阴影集的共享最邻近三支DBSCAN算法。该算法利用三支决策思想，将核心点划分到核心域中，对于非核心点引入阴影集理论，计算样本的隶属度，将样本划分到核心域或边界域中，并通过共享最邻近算法进一步细化边界域中的样本划分，从而提升聚类的准确性和鲁棒性。该算法应用在文本分析领域，通过实验对比分析，验证了该算法具有较好的性能，提高了文本聚类的准确性。

关键词

三支决策，三支聚类，阴影集，文本聚类

Three-Way DBSCAN Text Clustering Based on Shadowed Sets and Shared Nearest Neighbor

Zhicong Li*, Kun Yan

School of Computer Science and Information Engineering, Harbin Normal University, Harbin Heilongjiang

Received: Mar. 2nd, 2025; accepted: Mar. 25th, 2025; published: Apr. 1st, 2025

Abstract

The traditional DBSCAN algorithm, when processing data, often faces decision risks by forcing certain uncertain data points into a specific cluster. A three-way DBSCAN algorithm based on shadowed sets and Shared Nearest Neighbor is proposed to address this issue. This algorithm utilizes the three-way decision-making approach to classify core points into the core region. For non-core points, the theory of shadow sets is introduced to calculate the membership degree of the samples, categorizing them into either the core region or boundary region. The Shared Nearest Neighbor algorithm

*通讯作者。

文章引用: 李志聪, 闫昆. 基于阴影集的共享最邻近三支 DBSCAN [J]. 数据挖掘, 2025, 15(2): 137-150.
DOI: 10.12677/hjdm.2025.152012

is then applied to further refine the classification of samples within the boundary region, thereby enhancing the accuracy and robustness of clustering. Applied in text analysis, experimental comparative analysis has verified that this algorithm demonstrates better performance and improves the accuracy of text clustering.

Keywords

Three-Way Decision, Three-Way Clustering, Shadowed Sets, Text Clustering

1. 引言

聚类分析是数据挖掘和机器学习中的一种常用方法，其目标是将数据集中的样本按照一定的规则分成若干个簇，使得同一簇内的样本具有较高的相似性，而不同簇之间的样本相似性较低[1]，从而揭示数据的内在结构和模式，为进一步的数据分析和决策提供依据。在当今信息爆炸的时代，文本数据的数量和复杂性呈现指数级增长，这对数据挖掘和知识发现提出了严峻挑战。文本聚类作为自然语言处理和数据挖掘的重要技术手段，能够有效地将大量无结构文本数据自动归类，从而在信息检索[2]、主题检测[3]、情感分析[4]等领域中发挥重要作用。

DBSCAN [5] [6]作为一种密度聚类算法，因其能够识别任意形状的簇且具备处理噪声数据的能力，已经成为经典的聚类算法之一。然而，DBSCAN 在处理具有模糊边界或重叠簇时的局限性显而易见。由于其硬聚类的刚性，即将每个数据点严格划分为簇或噪声，常常在处理这些复杂数据时导致次优的结果[7]。为解决这一问题，一些改进方法引入了更具灵活性的策略。例如，基于模糊理论的 FN-DBSCAN 算法通过为每个数据点分配不同的隶属度，使其同时属于多个簇，从而有效处理模糊边界问题[8]。基于共享最近邻的 SNN-DBSCAN 算法通过结合最近邻算法的思想，将共享最近邻的概念引入密度计算，增强了处理重叠簇和边界模糊的能力[9]。基于三支决策模型的 3W-DBSCAN 通过引入边界域，允许对不确定数据点采取“延迟决策”的方式，避免了传统硬分类方法中过度刚性的分类决策，从而有效处理数据的不确定性和模糊性[10]。

然而，以上聚类算法大多仅停留在对边界点的分类优化，如何进一步细化边界点的划分，并提升整体聚类算法的鲁棒性仍是一个挑战。

由 Witold Pedrycz 等[11]-[14]提出的阴影集，广泛应用于处理模糊边界和不确定性数据的领域。它通过引入α阈值将模糊集的多值逻辑演变为$\{1,[0,1],0\}$的三值逻辑，降低了隶属度带来的冗余信息。Jiang 等[15]利用阴影集提出了新的聚类集成方法 S-M3WCE。Zhang 等[16]提出了一种改进的基于参数决策理论阴影集(RFKM-DTSS)的粗糙模糊 k 均值聚类方法，来使聚类边界合理化。在阴影集中，只有隶属度在区间$[\alpha,1-\alpha]$内的元素保留了模糊性，其余的元素都被明确分类，阴影集不仅保留了模糊集的本质，还简化了处理并增强了结果的解释性。目前关于阴影集理论与 DBSCAN 算法相结合进行三支聚类的研究还相对较少，特别是应用阴影集理论优化 DBSCAN 算法以解决模糊边界问题研究领域仍处于初步阶段。

因此，本文提出了基于阴影集的共享最邻近三支 DBSCAN 算法，该算法在 DBSCAN 二支聚类结果基础上将核心点划分到核心域中，对于非核心点引入阴影集理论，通过计算样本的隶属度，将样本划分到核心域或边界域中，并通过共享最邻近(SNN)算法进一步细化边界域中的样本划分。这一方法不仅增强

了对模糊边界的处理能力，还通过细化边界域的划分，提高了聚类结果的准确性，实验结果表明，该算法在文本聚类任务中具有较好的性能。

2. 相关工作

本小节分别回顾了 DBSCAN、三支决策和阴影集的部分基础知识。

2.1. DBSCAN

DBSCAN [5] [6]是一个经典的密度基聚类算法。它将簇定义为密度相连点的最大集合，能够有效地识别数据中的类簇结构，并且能够自动处理噪声数据点。该算法能够将高密度区域划分为簇，并能在含有噪声的空间数据库中发现任意形状的聚类。

DBSCAN 通过设定邻域半径 Eps 和最小点数 MinPts，首先标记数据集中所有点为未访问状态，然后依次访问每个点，若点的邻域内包含的点数不少于 MinPts，则将其标记为核心点，并扩展其邻域内的点加入同一个簇，否则标记为噪声点。核心点的邻域内所有点会被进一步检查和扩展，重复此过程直至所有点被访问并归类，最终形成类簇和噪声点的划分。

然而，DBSCAN 在处理模糊边界时缺乏灵活性，并且对 Eps 和 MinPts 参数的选择高度敏感，可能导致聚类结果不稳定[7]。针对 DBSCAN 算法的局限性，研究者们提出了多种改进方法。主要的改进方向包括：

1) 参数自适应选择：利用不同的优化技术，来动态确定聚类所需的关键参数(Eps 和 MinPts)，以更好地适应数据集的分布特征[17]-[19]。

2) 模糊聚类：引入模糊理论，为数据点分配不同的隶属度，使每个点可以同时属于多个簇，从而有效应对模糊边界问题[8] [20]。

3) 三支聚类：该方法引入边界域，允许对不确定数据点采取“延迟决策”的方式，有效缓解了传统硬分类方法的过度刚性，改善了数据处理的灵活性[10] [21]。

2.2. 三支决策

在经典的二支决策聚类中，对象和类之间只有两种关系：对象属于类或不属于类。清晰边界的要求有利于分析结果，却不能充分显示数据集中的不确定性信息[1]。为了解决信息不确定性问题，Yao [22] [23]基于粗糙集提出的的三支决策理论，通过将决策过程分为接受、拒绝和延迟三种类型，提供了一种处理不确定性和模糊性的方法。

假设一个有限非空对象集合 X 和一个有限的标准集 C。根据标准集 C，可以将整体划分成 3 个两两互不相交的区域：POS，BND，NEG。且满足如下标准：

$$\begin{aligned}&(1)\ POS \cap BND = \varnothing, POS \cap NEG = \varnothing, BND \cap NEG = \varnothing \\ &(2)\ POS \cup BND \cup NEG = X\end{aligned} \tag{1}$$

其中 POS 域，BND 域，NEG 域分别表示正域(接受决策)、边界域(延迟决策)和负域(拒绝决策)。

Yu [24] [25]将三支决策理论结合到聚类中提出的三支聚类框架，旨在将数据点划分为核心域、边界域和琐碎域，从而更有效地处理数据中的不确定性和模糊性。近年来，三支聚类引起了大量的研究，并开发了许多三支聚类算法。

三支聚类的聚类结果用二元集合表示 $C = POS(C_i), BND(C_i)$，$POS(C_i)$ 表示第 i 类的核心域，对象肯定属于该类，$BND(C_i)$ 表示第 i 的边界域，对象不确定属于该类。三支聚类的类簇满足如下性质：

$$\begin{aligned} &POS\left(C_i\right)\neq\varnothing, i=1,2,\cdots \\ &POS\left(C_i\right)\cap POS\left(C_j\right)=\varnothing, i\neq j \\ &\bigcup_{i=1}^{k}\left(POS\left(C_i\right)\cup BND\left(C_i\right)\right)=X_i \end{aligned} \tag{2}$$

然而，目前的三支决策聚类算法大多仅停留在对边界点的分类优化，如何进一步细化边界点的划分，并提升整体聚类算法的鲁棒性仍是一个挑战。

2.3. 阴影集

阴影集[11]-[14]是一种特殊的三支决策模型。对于阴影集来说，有 3 种量化级别：$\{1,[0,1],0\}$，即将模糊集转化成阴影集的三值逻辑，依次对应隶属度为 1，0 和隶属度不确定的区域。阴影集结构如图 1 所示。

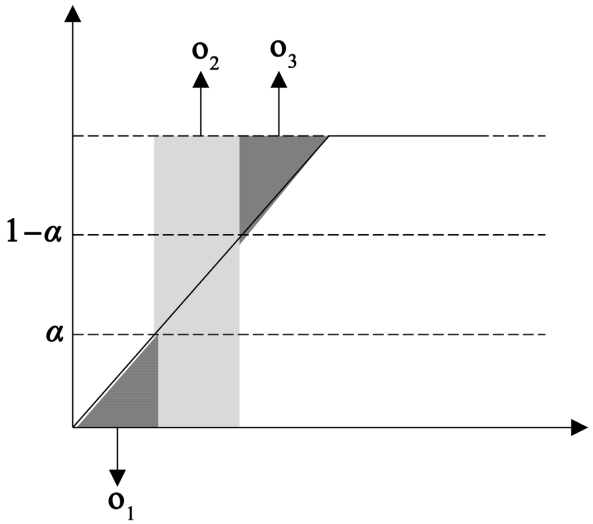


Figure 1. Shadow set structure
图 1. 阴影集结构

具体而言，如果对象 x 隶属度 $\mu_A(x)\leq\alpha$ 通过降低操作将隶属度降低为 0，该区域定义为降低区域(o_1)。如果对象 x 隶属度 $\mu_A(x)\geq 1-\alpha$ 通过提升操作将隶属度提升为 1，该区域定义为提升区域(o_3)。如果对象 x 隶属度 $\alpha<\mu_A(x)<1-\alpha$ 则变化区域定义为阴影区域(o_2)。

阴影集的一个重要问题是阈值的确定，Pedrycz 在[11]为了计算最优阈值，当隶属度值的集合是连续函数时 α 的取值应满足使式(3)取得最小值：

$$V=\left|\int_{\mu_A(x)\leq\alpha}\mu_A(x)\mathrm{d}x+\int_{\mu_A(x)\geq 1-\alpha}\left(1-\mu_A(x)\right)\mathrm{d}x-\int_{\alpha<\mu_A(x)<1-\alpha}\mathrm{d}x\right| \tag{3}$$

当处理隶属度值的集合而不是连续函数时 α 的取值应满足使式(4)取得最小值：

$$V=\left|\sum_{\mu_A(x)\leq\alpha}\mu_A(x)+\sum_{\mu_A(x)\geq 1-\alpha}\left(1-\mu_A(x)\right)-\mathrm{card}\left\{x_i\in U\mid\alpha<\mu_A(x)<1-\alpha\right\}\right| \tag{4}$$

其中 card{·}代表集合中对象的数量。

目前关于阴影集理论与 DBSCAN 算法相结合进行三支聚类的研究还相对较少。在这项工作中，我们引入将阴影集理论优化了现有的 DBSCAN 算法进行三支聚类。

3. 基于阴影集的共享最邻近三支 DBSCAN

3.1. 隶属度和相似度设计

为解决传统 DBSCAN 对模糊和不确定数据处理不足的问题，设计隶属度函数，用于量化每个数据点对不同区域的归属程度，通过计算隶属度，数据点不再被简单地划分为核心点或噪声点，而是进一步细化为核心域、边界域和噪声域。

图 2 为带有一个噪声点的数据分布，由图 2 可以看出，噪声点和 x_1 到类中心的距离相等，如果隶属度函数只按照样本点到类中心的距离作为评价指标，将会赋予 x_1 和噪声点相同的隶属度，显然不符合实际。因此，为了度量图 2 中对象隶属度之间的差异，本文借鉴文献[26]结合样本到类中心的距离和样本周围的紧密度的隶属度函数设计得到每个簇中样本点的隶属度。本文隶属函数定义为：

$$D_i = \frac{1}{K}\sum_{x_j \in N_k(x_i)} \left\| x_i - x_j \right\| \tag{5}$$

$$u(x_i) = \left(1-\beta\frac{d_i}{\max\limits_j(d_j)+0.0001} - (1-\beta)\frac{D_i - \min\limits_j(D_j)}{\max\limits_j(D_j) - \min\limits_j(D_j)+0.0001}\right)^m \tag{6}$$

其中，$\beta \in (0,1)$，m 的取值范围为 $\{0.1,0.2,0.3,\cdots,1.0\}$，$N_k(x_i)$ 为 x_i 的 k 个邻近样本的集合，d_i 为 x_i 到类中心之间的距离。本文 $\beta = 0.7, m = 0.2$。

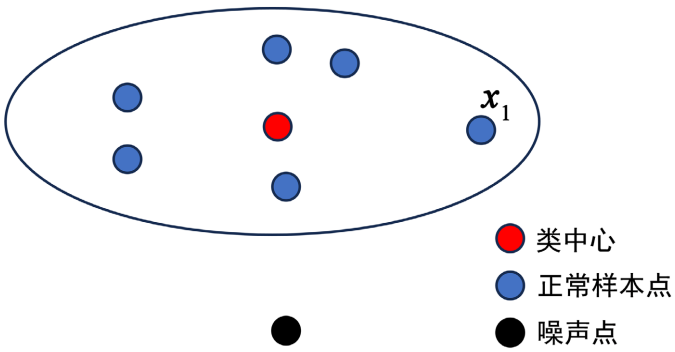


Figure 2. Data distribution chart
图 2. 数据分布图

为了更好地细化对文本数据边界点的划分，共享最邻近相似度通过计算每个点与其他点共享的最近邻个数，增强了对点与点之间关系的度量。本文相似度定义为：

$$SNN(P,Q) = \left| N_k(P) \cap N_k(Q) \right| \tag{7}$$

其中，$N_k(P)$ 和 $N_k(Q)$ 是点 P 和 Q 的 k 最近邻集合，本文 $k = 6$。

3.2. 算法思想

本文提出了一种基于阴影集的共享最邻近三支 DBSCAN 聚类算法(SS3W-DBSCAN)。该算法的核心思想在 DBSCAN 二支聚类结果基础上将核心点划分到核心域中，对于非核心点引入阴影集理论，通过计算样本的隶属度，将样本划分到核心域或边界域中。本文通过引入阴影集理论，允许不确定的数据点保留在边界域中，避免强制归类。之后，利用共享最近邻(SNN)方法对这些边界点进行进一步细化，确保文本的正确归类。算法具体步骤如下：

算法 SS3W-DBSCAN

输入：数据集 Data，Eps，MinPts，K 最邻近
输出 $C_j=\left(C_j^p,C_j^B\right)$

步骤 1：将数据集输入到 DBSCAN 算法中进行初始聚类，得到初始聚类结果 C_i， $i=1,2,3,\cdots,k$ 。

步骤 2：对于噪声点，将噪声点归类到离它最近的核心点的类簇中的边界域。

步骤 3：根据式(6)和式(4)分别计算各类簇的样本点隶属度和最优阈值 α_i， $i=1,2,3,\cdots,k$ 。

步骤 4：对于类簇 C_j 中任意点 p，如果点 p 是核心点，则将点 p 划分到 C_j 的核心域 C_j^p 之中。否则，获得 p 点的隶属度 $\mu_j(p)$，若 $\mu_j(p)\geq 1-\alpha_j$ 则把点 p 划分到 C_j 的核心域 C_j^p 之中。若 $\mu_j(p)<1-\alpha_j$ 则把点划分边界域 C_j^B 之中。

步骤 5：对于划分到类 C_j 中边界域中的样本点，由于可能是其他类簇的边界域中的对象，需要进一步处理，处理策略如下：对于 $p\in C_j^B$，得到点 p 的 *eps* 邻域 $N_{eps}p$，若点 x 为 p 的 *eps* 邻域中的点，即 $x\in N_{eps}p$，有 $x\in C_i, j\neq i$，则把 p 也划分到 C_i 的边界域中，即 $p\in C_i^B$ 。

步骤 6：根据式(7)计算边界点与核心点之间的 SNN 相似度。如果某个边界点与簇内某个核心点的相似度大于 $\frac{K}{2}$，则该边界点归入该核心点所在簇的核心域中；反之，该点将保留在边界域中。细化处理完成后，生成最终的三支聚类结果。

4. 实验和结果分析

为了验证本文算法的有效性和先进性，首先，将 SS3W-DBSCAN 算法与 3W-DBSCAN 算法[10]进行对比，使用三支聚类评价指标进行评估，三支聚类评价指标在 4.2 小节中已给出详细定义。然后，将 SS3W-DBSCAN 算法对文本集进行文本聚类，使用 ARI、NMI、ACC 指标进行评估。

4.1. 实验设置

本文在常用的人工数据集和 UCI 数据集上将本文所提算法与其他聚类算法进行对比，数据集的信息如表 1 所示，其中 1~5 为人工数据集，6~10 为 UCI 数据集。

Table 1. Data set
表 1. 数据集

序号	数据集	类个数	维度	样本数
1	Flame	2	2	240
2	Aggregation	7	2	788
3	Square	4	2	1000
4	R15	15	2	600
5	Cure	3	2	2000
6	Iris	3	4	150
7	Wine	3	13	178

续表

8	Seeds	3	7	197
9	Ecoli	5	7	210
10	Heart	4	13	303

4.2. 评价指标

4.2.1. 硬聚类评价指标

1) 调整兰德系数(ARI)是一种用于评估聚类结果的外部评价指标。其取值范围在[−1, 1]之间，其中 1 表示完全一致的聚类结果，值越接近 1 表示聚类效果越好。

$$\mathrm{ARI}=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left(\sum_{i}\binom{a_i}{2}\times\sum_{j}\binom{b_j}{2}\right)\Big/\binom{n}{2}}{\frac{1}{2}\left(\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}\right)-\left(\sum_{i}\binom{a_i}{2}\times\sum_{j}\binom{b_j}{2}\right)\Big/\binom{n}{2}} \tag{8}$$

其中，n 表示数据集中的样本总数，n_{ij} 表示真实标签为 α_i 且聚类结果为 b_j 的样本数。

2) 标准化互信息(NMI)是度量聚类结果和真实标签的相似程度的外部评价指标。其取值范围在[0, 1]之间，值越高表示越相似，聚类效果越好。

$$\mathrm{NMI}=\frac{2\times I(U,V)}{H(U)+H(V)} \tag{9}$$

其中，$I(U,V)$ 是聚类结果 U 和真实标签 V 之间的互信息，$H(U)$ 是聚类结果 U 的熵，$H(V)$ 是真实标签 V 的熵。

3) 准确度(ACC)可以衡量聚类结果的准确性。其取值范围[0, 1]之间，1 表示聚类结果与真实标签完全一致，值越高聚类效果越好。

$$\mathrm{ACC}=\frac{1}{n}\sum_{i=1}^{k}C_i \tag{10}$$

其中，n 表示数据集中的样本总数，C_i 表示类簇 i 与真实标签一致的对象个数，k 表示类簇的个数。

4.2.2. 三支聚类评价指标

由于硬聚类指标无法评估类簇的核心域和边界域之间的关系，而三支聚类在处理模糊边界以及提供更丰富的样本与簇关系信息方面具有显著优势。因此本文也使用了三支聚类指标进行评估。

第一个评价指标是 Maji 等[27]于 2007 年提出，其公式定义为：

$$\gamma=\frac{1}{n}\sum_{j=1}^{k}\left|C_j^p\right| \tag{11}$$

其中，n 表示数据集中的样本总数，k 表示类簇的个数，$\left|C_j^p\right|$ 表示 C_j 核心域中对象个数。γ 表示所有类簇的核心域的平均质量，其值越高表示聚类效果越好。

第二个评价指标是 Zhang 等[28]于 2019 年提出，其公式定义为：

$$\alpha = \frac{1}{k}\sum_{j=1}^{k}\frac{\left|C_j^p\right|}{\left|C_j^p\right|+\left|C_j^b\right|} \tag{12}$$

$$\alpha^* = \frac{\sum_{j=1}^{k}\left|C_j^p\right|}{\sum_{j=1}^{k}\left(\left|C_j^p\right|+\left|C_j^b\right|\right)} \tag{13}$$

其中，$\left|C_j^b\right|$表示C_j边界域中对象个数，α和α^*分别表示所有类簇的平均质量和整体质量，其值越高表示聚类效果越好。

4.3. 三支聚类实验

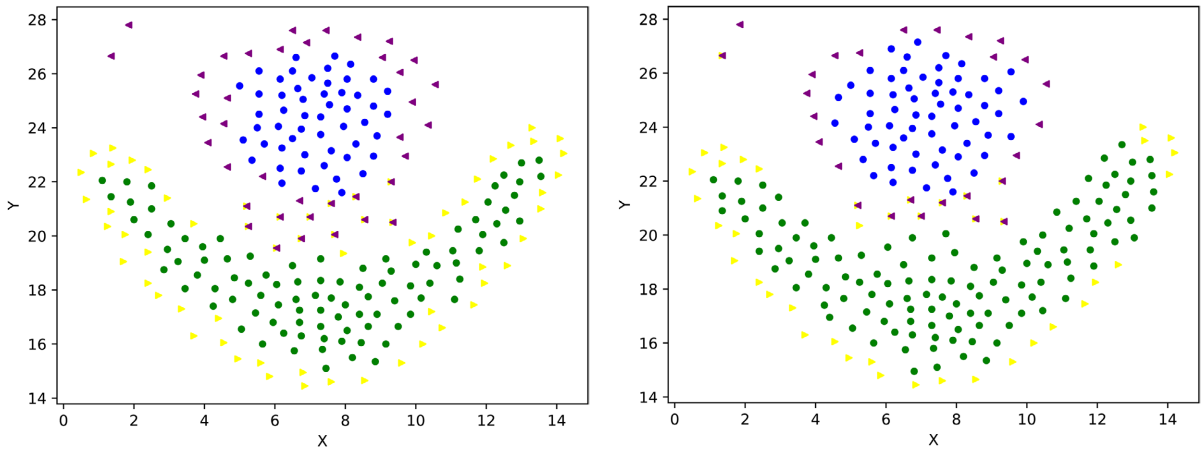
本小节使用3W-DBSCAN算法与本文所提出的SS3W-DBSCAN算法对10个常用数据集进行三支聚类，其中，在人工数据集上的参数设置如表2所示，聚类效果如图3~7所示。在UCI数据集和人工数据集进行聚类得到的三支聚类结果评价指标评估得分如表3所示。

图3~7为SS3W-DBSCAN算法3W-DBSCAN算法在人工数据集中的聚类结果，从图中可以看出，与3W-DBSCAN聚类结果相比，SS3W-DBSCAN算法进一步细化边界点的划分，减少了样本在边界区域的过度分配，在保留模糊信息的同时，也能够避免必要的确定性信息的丢失。

通过表3数据可知，SS3W-DBSCAN算法在识别核心点和边界点时更加准确，使得聚类结果更加精确，特别是在处理类间差异明显的数据集时优势尤为明显，有效地减少了样本点在边界区域的过度分配，降低了不确定性的影响，显著提升了在三支聚类任务中的表现。

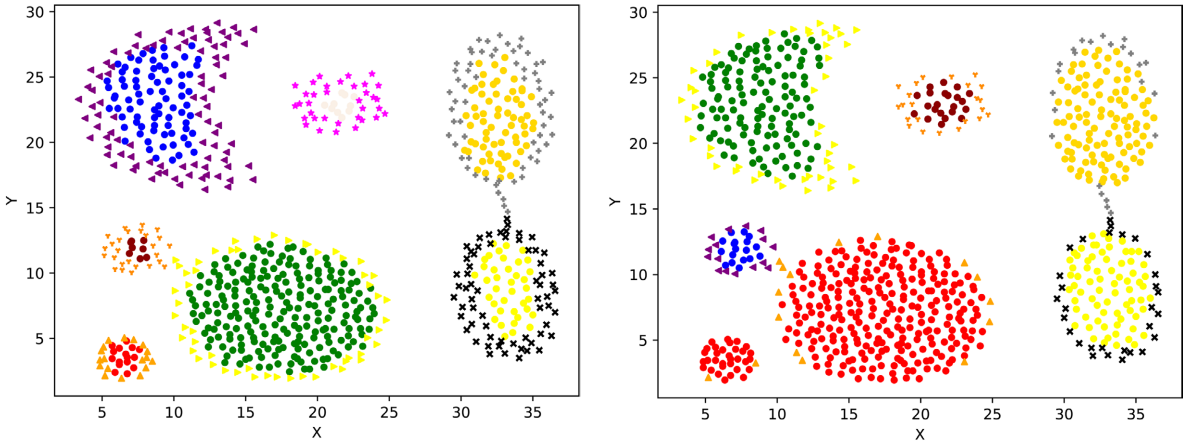
Table 2. Parameter settings for synthetic datasets
表2. 人工数据集参数设置

数据集	Flame	Aggregation	Square	R15	Cure
Eps	1.550	2.737	3.420	0.772	0.267
MinPts	13	31	126	32	155



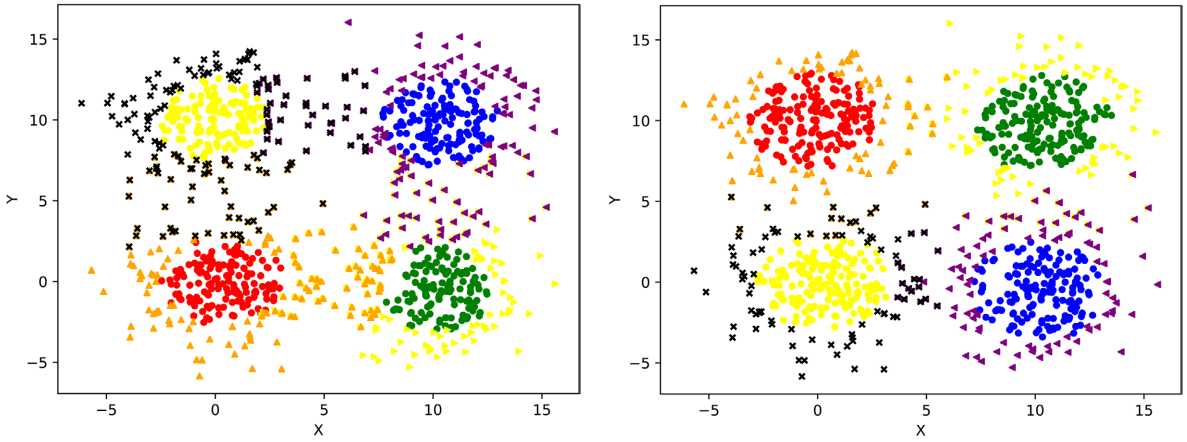
(a) 3W-DBSCAN　　(b) SS3W-DBSCAN

Figure 3. Clustering results of flame
图3. Flame聚类结果



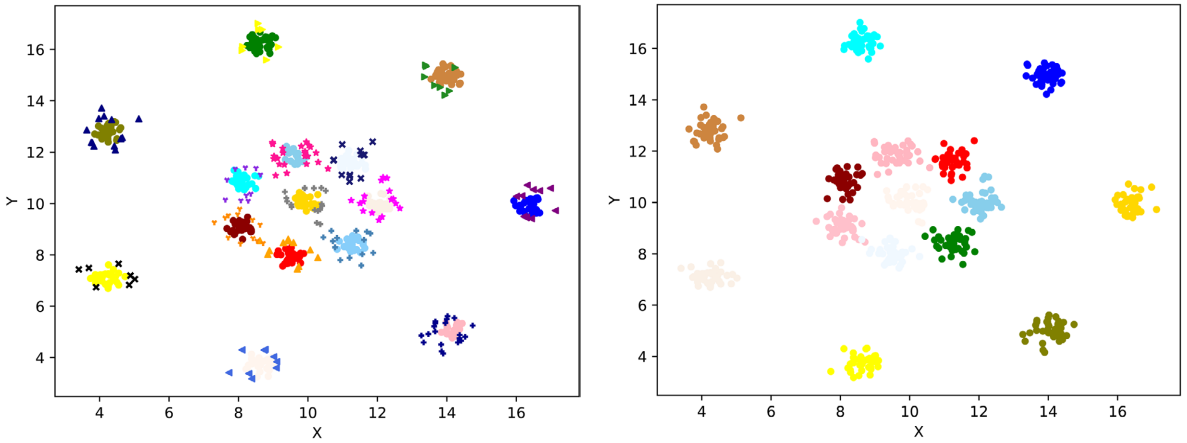
(a) 3W-DBSCAN (b) SS3W-DBSCAN

Figure 4. Clustering results of aggregation
图 4. Aggregation 聚类结果



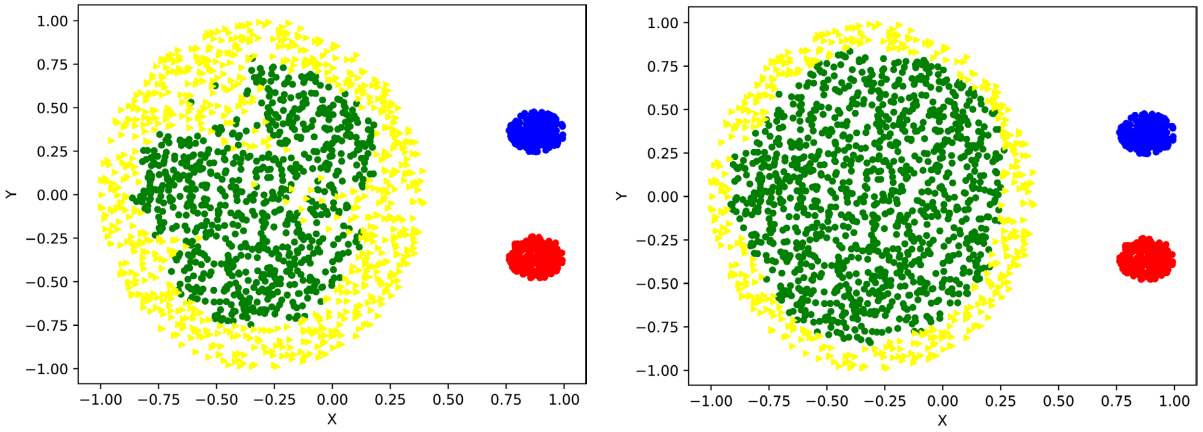
(a) 3W-DBSCAN (b) SS3W-DBSCAN

Figure 5. Clustering results of square
图 5. Square 聚类结果



(a) 3W-DBSCAN (b) SS3W-DBSCAN

Figure 6. Clustering results of R15
图 6. R15 聚类结果



(a) 3W-DBSCAN　　(b) SS3W-DBSCAN

Figure 7. Clustering results of cure
图 7. Cure 聚类结果

Table 3. Three-way clustering evaluation results
表 3. 三支聚类评价结果

数据集	γ		α		α^*	
	3W-DBSCAN	SS3W-DBSCAN	3W-DBSCAN	SS3W-DBSCAN	3W-DBSCAN	SS3W-DBSCAN
Flame	0.550	0.829	0.522	0.794	0.522	0.809
Aggregation	0.473	0.722	0.331	0.693	0.473	0.722
Square	0.530	0.796	0.526	0.764	0.526	0.764
R15	0.672	1.000	0.672	1.000	0.672	1.000
Cure	0.598	0.728	0.832	0.886	0.598	0.728
Iris	0.540	0.860	0.526	0.854	0.523	0.854
Wine	0.500	0.854	0.351	0.795	0.489	0.792
Seeds	0.519	0.857	0.494	0.799	0.495	0.800
Ecoli	0.476	0.878	0.388	0.883	0.476	0.878
Heart	0.535	0.772	0.504	0.695	0.513	0.696

4.4. 文本聚类实验

本文选用了 THUCNews 数据集和 SogouCS 数据集，THUCNews 由清华大学自然语言处理实验室收集和发布包含 74 万篇新闻文本，涵盖 14 个类别，选取体育、科技、教育、家居、财经 5 个类别，每个类 240 篇，共 1200 篇文档。SogouCS 数据集由搜狐公司推出包含 270 多万篇新闻文本涵盖社会、经济、科技等多个类别，选取 IT、健康、教育、旅游、体育 5 个类别，1210 篇文档，其中，IT、健康每个类 245 篇，其余每个类 240 篇。关于文本数据集的详细信息如表 4 所示。

Table 4. Text dataset
表 4. 文本数据集

数据集	类别数	具体类别	数据总量
THUCNews	5	体育、科技、教育、家居、财经	1200
SogouCS	5	IT、健康、教育、旅游、体育	1210

4.4.1. 文本关于类别数量变化实验

本次实验目的是测试聚类效果关于文档类别数量的变化。对比实验分成两类：DBSCAN 算法、本文提出的 SS3W-DBSCAN 算法，使用 THUCNews 数据集，分成五组试验：第一组实验的数据集包含 5 类，每一类 100 篇共 500 篇文档；第二组实验的数据集包含 6 类，每一类 100 篇共 600 篇文档；以此类推，第五组实验的数据集是 9 类共 900 篇文档。每组实验均采用 BERT 模型处理文档向量化，使用去噪自编码器进行降维，每组实验过程完全一致。ACC 指标随类别数量变化如图 8 所示。

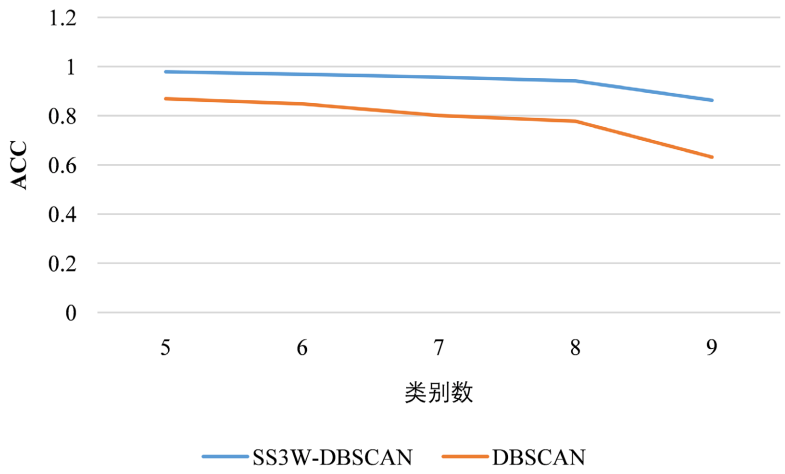


Figure 8. ACC metrics vary with the number of categories
图 8. ACC 指标随类别数量变化图

从实验结果可以知道文本聚类的 ACC 值会随着文本类别的增加而降低，当类别数较小时，ACC 值很高，当类别数增加时，ACC 值逐渐降低。使用 SS3W-DBSCAN 算法后，聚类效果更好且稳定，在聚类类别小于 10 时，文档聚类的 ACC 值一直超 80 个百分点。

4.4.2. 文本综合对比实验

本次实验目的是验证 SS3W-DBSCAN 算法的文本聚类效果。对比实验分成五类：DBSCAN 算法、FDBSCAN 算法[29]、KANN-DBSCAN 算法[18]、3W-DBSCAN 算法[10]和本文提出的 SS3W-DBSCAN 算法。实验均采用 BERT 模型进行文本向量化，使用去噪自编码器进行降维，实验过程完全一致，仅聚类算法不同。

实验结果如表 5~7 所示，SS3W-DBSCAN 算法在 THUCNews 和 SogouCS 数据集上的各项指标均优于对比算法，其 ACC 指标分别达到 0.898 和 0.901，NMI 和 ARI 指标也显著提升。相比经典 DBSCAN 算法，ACC 指标分别提升了 16.9%和 14.1%，相比 3W-DBSCAN 算法，也有显著提升，体现了更高的聚类精度。无论是在较小规模的 THUCNews 数据集还是在较大规模的 SogouCS 数据集上，SS3W-DBSCAN 均表现出了较好的性能。

Table 5. ACC comparison on text datasets
表 5. 文本数据集上的 ACC 对比

Model	数据集	
	THUCNews	SogouCS
	ACC	ACC
DBSCAN	0.729	0.675
FDBSCAN	0.809	0.777
KANN-DBSCAN	0.822	0.821
3W-DBSCAN	0.871	0.881
SS3W-DBSCAN	0.898	0.901

Table 6. NMI comparison on text datasets
表 6. 文本数据集上的 NMI 对比

Model	数据集	
	THUCNews	SogouCS
	NMI	NMI
DBSCAN	0.719	0.707
FDBSCAN	0.800	0.778
KANN-DBSCAN	0.809	0.785
3W-DBSCAN	0.816	0.828
SS3W-DBSCAN	0.842	0.848

Table 7. ARI comparison on text datasets
表 7. 文本数据集上的 ARI 对比

Model	数据集	
	THUCNews	SogouCS
	ARI	ARI
DBSCAN	0.655	0.652
FDBSCAN	0.728	0.703
KANN-DBSCAN	0.747	0.727
3W-DBSCAN	0.781	0.798
SS3W-DBSCAN	0.839	0.803

5. 结束语

本文提出的基于阴影集的共享最邻近三支 DBSCAN 聚类算法(SS3W-DBSCAN)，有效解决了传统

DBSCAN 在处理文本数据时存在的模糊边界和不确定性问题。通过引入阴影集理论，避免了将不确定的边界点强制归类的风险；而与共享最邻近算法的结合，进一步细化了边界点的划分，提升了文本聚类的准确性和鲁棒性。实验结果表明，SS3W-DBSCAN 在不同文本数据集上的表现优于传统 DBSCAN 算法。接下来研究可以考虑进一步优化算法的计算效率，降低复杂度，以便更好地适应更多应用场景。

基金项目

哈尔滨师范大学双一流-提高人才培养质量项目(1504120015)；哈尔滨师范大学计算机科学与信息工程学院教育教学改革项目(JKYJGY202205)。

参考文献

[1] Wang, P., Yang, X., Ding, W., Zhan, J. and Yao, Y. (2024) Three-Way Clustering: Foundations, Survey and Challenges. *Applied Soft Computing*, **151**, Article ID: 111131. https://doi.org/10.1016/j.asoc.2023.111131

[2] Leuski, A. (2001) Evaluating Document Clustering for Interactive Information Retrieval. *Proceedings of the Tenth International Conference on Information and Knowledge Management*, Atlanta, 5-10 October 2001, 33-40. https://doi.org/10.1145/502585.502592

[3] Mei, Q. and Zhai, C. (2005) Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, 21-24 August 2005, 198-207. https://doi.org/10.1145/1081870.1081895

[4] Nandwani, P. and Verma, R. (2021) A Review on Sentiment Analysis and Emotion Detection from Text. *Social Network Analysis and Mining*, **11**, Article No. 81. https://doi.org/10.1007/s13278-021-00776-6

[5] Ester, M., Kriegel, H.P., Sander, J., *et al*. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, 2-4 August 1996, 226-231.

[6] Rehman, S.U., Asghar, S., Fong, S. and Sarasvady, S. (2014) DBSCAN: Past, Present and Future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies* (*ICADIWT* 2014), Bangalore, 17-19 February 2014, 232-238. https://doi.org/10.1109/icadiwt.2014.6814687

[7] Deng, D. (2020) DBSCAN Clustering Algorithm Based on Density. 2020 7*th International Forum on Electrical Engineering and Automation* (*IFEEA*), Hefei, 25-27 September 2020, 949-953. https://doi.org/10.1109/ifeea51475.2020.00199

[8] Ienco, D. and Bordogna, G. (2016) Fuzzy Extensions of the DBScan Clustering Algorithm. *Soft Computing*, **22**, 1719-1730. https://doi.org/10.1007/s00500-016-2435-0

[9] Ertöz, L., Steinbach, M. and Kumar, V. (2003) Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. *Proceedings of the* 2003 *SIAM International Conference on Data Mining*, San Francisco, 1-3 May 2003, 47-58. https://doi.org/10.1137/1.9781611972733.5

[10] Yu, H., Chen, L., Yao, J. and Wang, X. (2019) A Three-Way Clustering Method Based on an Improved DBSCAN Algorithm. *Physica A: Statistical Mechanics and Its Applications*, **535**, Article 122289. https://doi.org/10.1016/j.physa.2019.122289

[11] Pedrycz, W. (1998) Shadowed Sets: Representing and Processing Fuzzy Sets. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, **28**, 103-109. https://doi.org/10.1109/3477.658584

[12] Pedrycz, W. and Vukovich, G. (2002) Granular Computing with Shadowed Sets. *International Journal of Intelligent Systems*, **17**, 173-197. https://doi.org/10.1002/int.10015

[13] Pedrycz, W. (2005) Interpretation of Clusters in the Framework of Shadowed Sets. *Pattern Recognition Letters*, **26**, 2439-2449. https://doi.org/10.1016/j.patrec.2005.05.001

[14] Pedrycz, W. (2009) From Fuzzy Sets to Shadowed Sets: Interpretation and Computing. *International Journal of Intelligent Systems*, **24**, 48-61. https://doi.org/10.1002/int.20323

[15] Jiang, C., Li, Z. and Yao, J. (2022) A Shadowed Set-Based Three-Way Clustering Ensemble Approach. *International Journal of Machine Learning and Cybernetics*, **13**, 2545-2558. https://doi.org/10.1007/s13042-022-01543-5

[16] Zhang, Y., Zhang, T., Peng, C., Ma, F. and Pedrycz, W. (2024) Rough Fuzzy K-Means Clustering Based on Parametric Decision-Theoretic Shadowed Set with Three-Way Approximation. *International Journal of Fuzzy Systems*, **26**, 1698-1715. https://doi.org/10.1007/s40815-024-01700-8

[17] Zhang, X. and Zhou, S. (2023) WOA-DBSCAN: Application of Whale Optimization Algorithm in DBSCAN Parameter

Adaption. *IEEE Access*, **11**, 91861-91878. https://doi.org/10.1109/access.2023.3307412

[18] 李文杰, 闫世强, 蒋莹, 等. 自适应确定 DBSCAN 算法参数的算法研究[J]. 计算机工程与应用, 2019, 55(5): 1-7, 148.

[19] Kim, J., Choi, J., Yoo, K. and Nasridinov, A. (2018) AA-DBSCAN: An Approximate Adaptive DBSCAN for Finding Clusters with Varying Densities. *The Journal of Supercomputing*, **75**, 142-169. https://doi.org/10.1007/s11227-018-2380-z

[20] Smiti, A. and Eloudi, Z. (2013) Soft DBSCAN: Improving DBSCAN Clustering Method Using Fuzzy Set Theory. 2013 6*th International Conference on Human System Interactions* (*HSI*), Sopot, 6-8 June 2013, 380-385. https://doi.org/10.1109/hsi.2013.6577851

[21] 申秋萍, 张清华, 高满, 等. 基于局部半径的三支 DBSCAN 算法[J]. 计算机科学, 2023, 50(6): 100-108.

[22] Yao, Y. (2010) Three-Way Decisions with Probabilistic Rough Sets. *Information Sciences*, **180**, 341-353. https://doi.org/10.1016/j.ins.2009.09.021

[23] Yao, Y. (2011) The Superiority of Three-Way Decisions in Probabilistic Rough Set Models. *Information Sciences*, **181**, 1080-1096. https://doi.org/10.1016/j.ins.2010.11.019

[24] Yu, H., Zhang, C. and Wang, G. (2016) A Tree-Based Incremental Overlapping Clustering Method Using the Three-Way Decision Theory. *Knowledge-Based Systems*, **91**, 189-203. https://doi.org/10.1016/j.knosys.2015.05.028

[25] Yu, H. (2017) A Framework of Three-Way Cluster Analysis. *Rough Sets*: *International Joint Conference*, *IJCRS* 2017, Olsztyn, 3-7 July 2017, 300-312. https://doi.org/10.1007/978-3-319-60840-2_22

[26] 鞠哲, 曹隽喆, 顾宏. 用于不平衡数据分类的模糊支持向量机算法[J]. 大连理工大学学报, 2016, 56(5): 525-531.

[27] Maji, P. and Pal, S.K. (2007) RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets. *Fundamenta Informaticae*, **80**, 475-496. https://doi.org/10.3233/fun-2007-80408

[28] Yang, F., Xie, H. and Li, H. (2019) RETRACTED: Video Associated Cross-Modal Recommendation Algorithm Based on Deep Learning. *Applied Soft Computing*, **82**, Article 105597. https://doi.org/10.1016/j.asoc.2019.105597

[29] 周水庚, 周傲英, 曹晶, 等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11): 1287-1292.