DOS-IN聚类算法优化

摘要

DOS（Delta Open Set）聚类算法能够识别复杂形状的簇，但是其对输入参数的依赖性大，且由于其通过特定函数生成规则邻域用于识别开集，导致DOS在处理重叠的集群和高斯集群时表现较差。本文复现的DOS-IN（Irregular Neighborhoods）聚类算法基于对象之间的相似度生成不规则邻域，能够自适应对象的分布，不仅可以准确地区分重叠的簇，而且具有更少的输入参数。此外，DOS-IN引入了小簇合并机制，解决DOS在识别高斯簇方面的不足。但DOS–IN算法在数据预处理与小簇合并机制设计上仍存在不足，为此，本文在程序中加入了特征工程的数据预处理方法并优化了小簇合并机制，有效提高该算法的准确率。

关键词： 开集聚类；数据预处理；小簇合并机制

1 引言

聚类是数据挖掘中的关键技术之一，它的目的是揭示数据的内在结构，核心是识别数据中的簇数目以及簇中心。聚类的过程是将样本点分为一组簇的过程，基于相似度度量，同一簇中的样本点彼此靠近，而不同簇中的样本点彼此远离。目前，聚类已广泛应用于模式识别、数据压缩、图像分割、空间数据分析等等领域。

DOS（Delta Open Set）聚类算法[1]是第一个引入集合概念的聚类算法，它擅长识别复杂形状的簇，但它还存在以下三大不足：1.对于重叠的数据集，可能会错误地将相邻的簇识别为一个类别。2.对于高斯数据集，可能会错误地将高斯簇的边界对象识别为噪声。3.对输入参数的依赖性大，不同的输入参数可能会产生显著不同的聚类结果。可见，DOS算法还存在很大的优化空间。本文复现的DOS-IN（Irregular Neighborhoods）聚类算法[2]从DOS算法入手，优化创新，通过优化邻域设计方法和引入小簇合并机制，解决了DOS算法的三大缺陷。

本文提到的DOS、DOS-IN聚类算法都是基于距离和相似度展开聚类的，各变量间欧氏距离的取值与这些聚类算法密切相关，而距离对各特征的取值范围十分敏感，若某个特征的尺度比其他特征的尺度要大很多，那么它对距离的影响将会远大于其他特征，导致其掩盖了其他特征对总距离的影响，这样聚类算法便不能很好的识别所有特征，导致聚类有效性降低。同时，为了得到更好的聚类结果，在聚类时往往使用较多的指标进行综合分析，导致变量冗杂、计算成本高的问题。为了解决上述两种问题，本文尝试通过五种特征工程方法对DOS-IN聚类算法的实验数据进行预处理，其中包括：4种特征缩放方法（Min-Max缩放，方差缩放，L2标准化，鲁棒标准化）和主成分分析法。从实验结果中发现，特征缩放的4种方法对DOS-IN聚类算法的聚类有效性有显著的优化提升作用，主成分分析法能有效降低其聚类计算负担。

此外，DOS-IN算法中所用的小簇合并机制是找出距离该对象最近的大簇对象，合并到该大簇对象所在簇。但由于边界对象分布的散乱且有可能与其他簇的边界出现重叠，这样的合并机制显然缺乏合理性。因此，本文在复现时对这一机制进行了优化，先计算每个初始大簇

的簇中心，通过寻找最近的大簇簇中心合并小簇。通过实验发现，这一优化对小簇的分类是合理有效的。

2 相关工作

聚类模型大致可以分为4种：基于中心的模型、基于密度的模型、基于层次结构的模型和基于网络的模型。2018年，Shuliang Wang等学者首次在聚类算法中引入了集合概念，称为DOS(Delta Open Set)[1]聚类算法，它不同于上述的4种聚类模型。

2.1 DOS聚类算法

本节将简要描述DOS聚类算法定义的用于识别开集的DOS-δ-邻域。

设o_i是数据集O中的第i个对象，o_{ig}是o_i的第g个最近邻，$\mathrm{dis}(o_i, o_j)$ 是o_i和o_j之间的欧氏距离。接下来，将描述理解DOS算法与DOS-IN算法的重要定义：

定义1.（δ-邻域）如果$\mathrm{N}_\delta(o_i)$是对象o_i的δ-邻域，对于$\forall o_j \in \mathrm{O}$，如果$\mathrm{dis}(o_i, o_j) < \delta$，那么$o_j \in \mathrm{N}_\delta(o_i)$。即$\mathrm{N}_\delta(o_i) = \{o_j | \mathrm{dis}(o_i, o_j) < \delta\}$。

定义2.（K-邻域）如果$\mathrm{N}_\mathrm{K}(o_i)$是对象$o_i$的K-邻域，对于$\forall o_j \in \mathrm{O}$，如果$\mathrm{dis}(o_i, o_j) \le \mathrm{dis}(o_i, o_{iK})$，则$o_j \in \mathrm{N}_\mathrm{K}(o_i)$.也就是说，$\mathrm{N}_\mathrm{K}(o_i) = \{o_{ig} | \mathrm{g} \le K\}$。

定义3.（开集）如果A是开集，则对于$\forall o_i \in A$，$\exists\delta$，使得如果有$o_j \in \mathrm{N}_\delta(o_i)$，则有$o_j \in A$。

开集聚类的核心思想是将对象的δ-邻域重叠的对象确定为一个开集，一个开集对应一个类。而如何定义这个用于识别开集的邻域半径就是开集聚类算法的重难点。

DOS算法通过在δ-邻域$\mathrm{N}_\delta(o_i)$中增设了一个邻域半径系数，记为$\mathrm{Rad}(o_i)$，从而构造DOS-δ-邻域，记为$\mathrm{N}_\delta^\dagger(o_i)$，$\mathrm{N}_\delta^\dagger(o_i) = \{o_j | \mathrm{dis}(o_i, o_j) < \mathrm{Rad}(o_i) \cdot \delta\}$，将$\mathrm{N}_\delta^\dagger(o_i)$重合的对象集确定为开集，即聚为一类。其中$\mathrm{Rad}(o_i)$由特定函数式(1)定义。

$$\mathrm{Rad}(o_i) = \begin{cases} k\frac{1-e^{\frac{\rho(o_i)-\mu}{\sigma}}}{1+e^{\frac{\rho(o_i)-\mu}{\sigma}}}, & if\rho(o_i) < \mu - \sigma, \\ 1, & if\rho(o_i) \ge \mu - \sigma, \end{cases} \tag{1}$$

其中，$\rho(o_i)$为对象o_i的密度，μ和σ为数据集中所有对象的平均值和标准差。k通常是一个较大的正整数。通过简单推导可以发现，稀疏对象的邻域半径大于密集对象的邻域半径。因此，一旦簇的边界重叠，边界区域中的稀疏对象将被由于大的邻域半径而被错误地视为同一开集，如图1(A)所示。同时，DOS在半径系数中引入新的输入参数K，使得DOS算法需要两个输入参数k和δ，不同的输入参数可能产生显著不同的聚类结果，如图1(C)所示。

此外，如果集群中的对象数量小于数据集中对象总数的1%-2%，DOS将把集群中的对象视为噪声。在高斯数据集中，边界对象的密度远小于核心对对象的密度。为了避免相邻的集群被识别为一个类别，DOS只能设置一个小的k值来限制边界对象的邻域半径，导致许多边界对象被分配到小簇中，从而被错误地识别为噪声，如图1(B)所示。

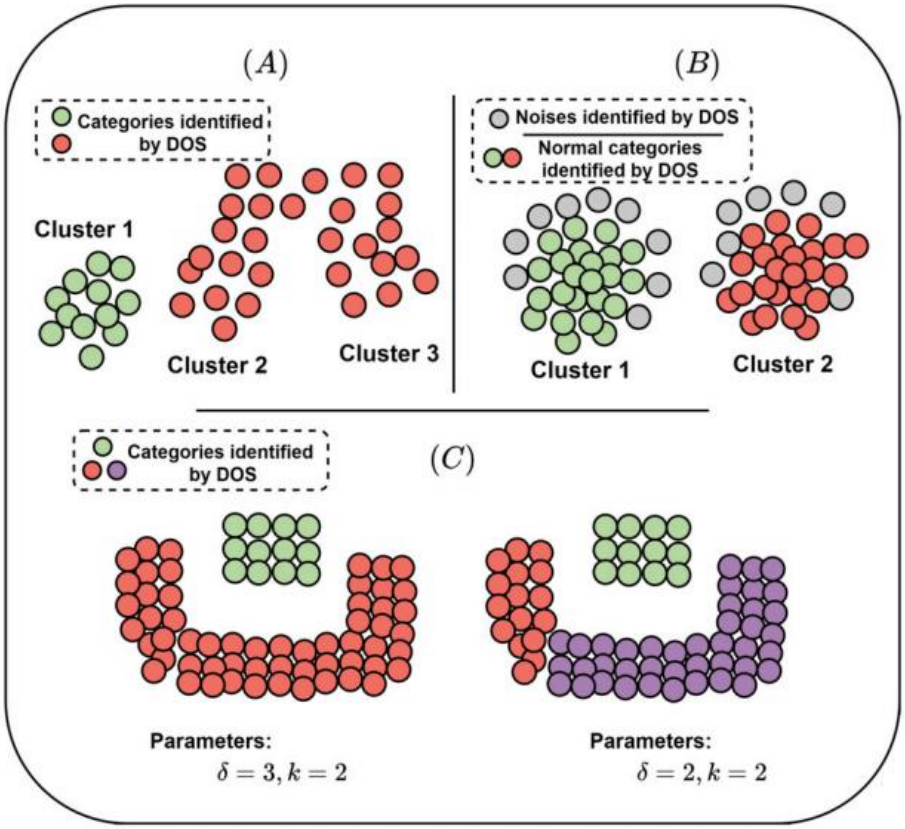

图 1. DOS的缺陷： (A) DOS错误地将重叠数据集中的相邻聚类识别为一个类别；(B) DOS错误地将高斯簇的边界对象识别为噪声；(C)不同的DOS输入参数产生显著不同的聚类结果。

3 DOS-IN聚类算法

3.1 DOS-IN聚类算法概述

为了解决DOS的缺陷，Qi L等学者提出了一种新的DOS-IN（Irregular Neighborhoods）聚类算法[2]。DOS-IN仍然通过识别开集来识别集群。与DOS不同的是，DOS-IN放弃了使用特定的模型来确定邻域半径，并引入了小簇合并机制。具体来说，DOS-IN包括以下3个步骤：

步骤1.计算相似度：DOS-IN根据对象 δ -邻域中的对象数量确定K-邻域的K值，然后通过K-邻域之间的Jaccard系数计算对象之间的相似度，计算公式如式(2)。

$$\mathrm{Sim}(o_i, o_j) = \frac{|\mathrm{N_K}(o_i) \cap \mathrm{N_K}(o_j)|}{|\mathrm{N_K}(o_i) \cup \mathrm{N_K}(o_j)|} \tag{2}$$

其中，$|x|$表示集合x中的对象的数量,$\mathrm{K} = \frac{|\mathrm{N}_\delta(o_i) + \mathrm{N}_\delta(o_j)|}{2}$。显然，$\mathrm{Sim}(o_i, o_j)$在[0,1]内，其值越接近1，$\mathrm{N_K}(o_i)$和$\mathrm{N_K}(o_j)$的重叠越多，$o_i$和$o_j$的相似度越大。

步骤2.确定IN-δ-邻域：基于相似度，DOS-IN对每个对象相对于不同的对象设置不同的邻域半径，使每个对象获得一个不规则的IN-δ-邻域，记为$\mathrm{N}_\delta^*(o_i)$，以应对不同的对象分布。$\mathrm{N}_\delta^*(o_i) = \{o_j | \mathrm{dis}(o_i, o_j) < \mathrm{Rad}^*(o_i|o_j) \cdot \delta\}$，其中$\mathrm{Rad}^*(o_i|o_j) = \mathrm{Sim}(o_i, o_j)$。

步骤3.识别集群： DOS-IN识别一组对象，其IN-δ-邻域重合，作为一个开放集，即一个初始集群。接下来，DOS-IN将初始小簇中的对象分配给最近的初始大簇，以得到最终的聚类结果。

3.2 DOS-IN算法的优势

DOS-IN算法所设计的不规则邻域$N_\delta^*(o_i)$，可以正确地区分重叠的集群。虽然重叠簇之间的边界对象彼此接近，但大多数边界对象都更靠近它们自己的簇中的核心对象（以下简称为自身核心对象）。因此，大多数边界对象与自身核心对象的K-邻域更一致，因此它们之间的相似度更大。因为一个对象的邻域半径关于不同对象之间的相似度成正比，所以大多数边界对象的IN-δ-邻域将偏向于自身核心对象而不是偏向在相邻的集群中的对象，这确保DOS-IN可以正确区分重叠的集群。通过图2所示的复现实验结果，可以证明DOS-IN算法的这一优势。

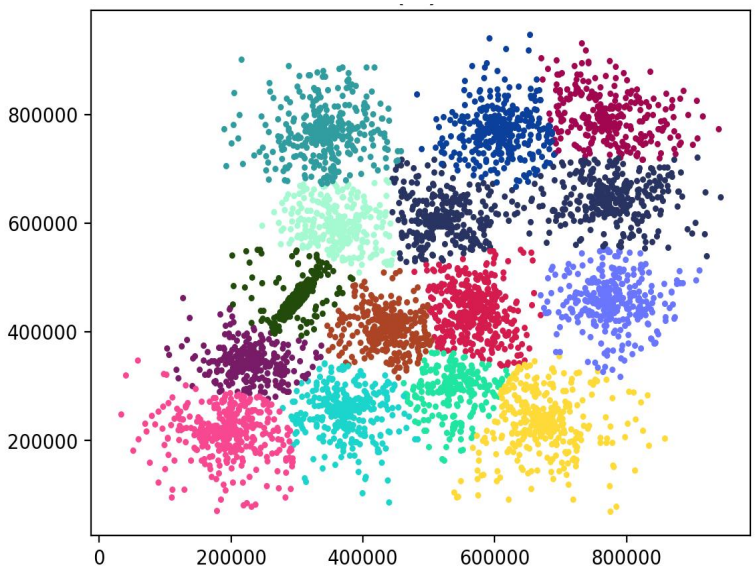

图 2. DOS-IN能够在重叠数据集中很好得识别不同的集群（不同颜色标记不同簇）

另外，DOS-IN算法的输入参数更少，IN-δ-邻域只涉及一个输入参数δ，因此比DOS-δ-邻域更容易获得最优结果，对输入参数的依赖性较小。

同时，DOS-IN算法通过引入小簇合并机制，将初始小集群中的对象分配给最近的初始大集群。在处理高斯数据集时，则可以将被错误识别为噪声的边界点，尽可能得分到他们所属的集群中。通过图3所示的复现实验结果，可以证明DOS-IN算法能够很好地处理高斯数据的边界点，不会将其识别为噪声。

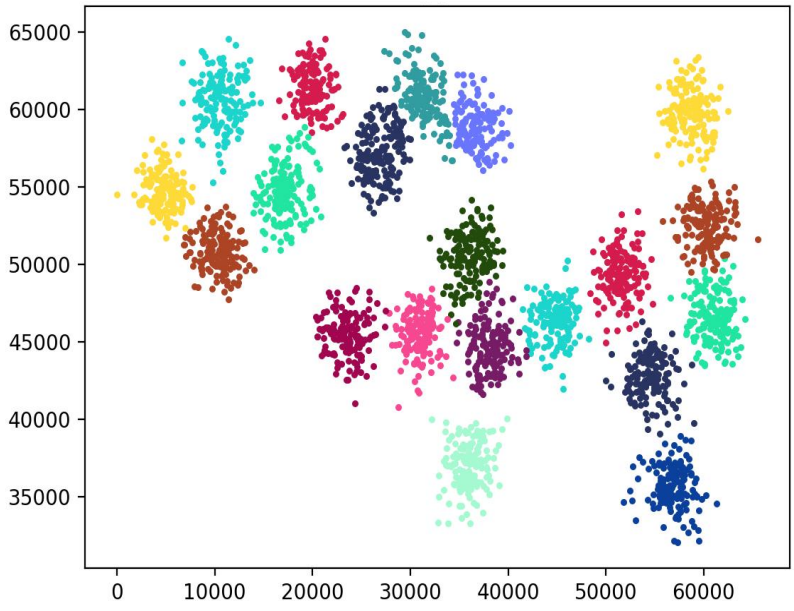

图 3. DOS-IN能够很好得识别高斯数据集及其边界点（不同颜色标记不同簇）

在优化了DOS算法的三大缺陷的情况下，DOS-IN算法依旧保留了DOS算法能够很好得识别复杂形状的簇的优点，通过图4所示的复现实验结果，可以证明DOS-IN算法能够正确识别不同形状的簇，对复杂和多样的分布具有鲁棒性。

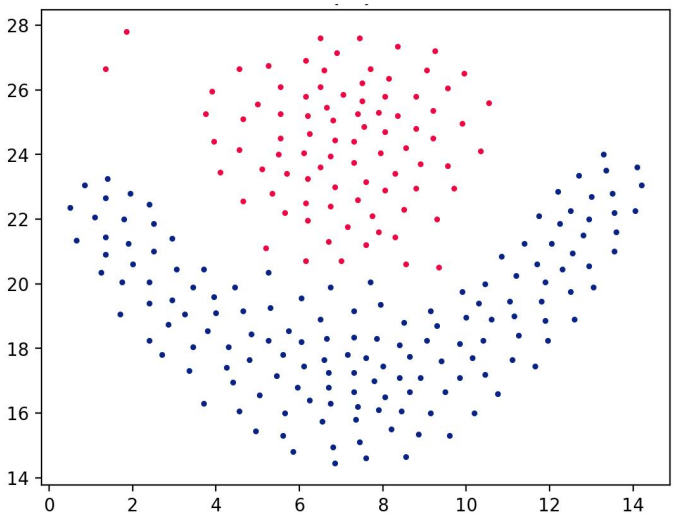

图 4. DOS-IN能够正确识别不同形状的簇（不同颜色标记不同簇）

4 复现细节

4.1 与已有开源代码对比

本实验参考DOS-IN算法的开源代码（链接：https://github.com/Youth-49/2023-DOS-IN）对其进行复现并优化。

4.1.1 使用特征工程方法对数据进行预处理

根据DOS-IN算法的开源代码与数据可以发现，Li Qi等学者所发表的有关DOS-IN论文[2]实验中所用的部分数据存在量纲差异大、样本维度高的问题，如图5所示为其中BreastTissue——乳腺组织数据集的9个变量的核密度估计曲线。

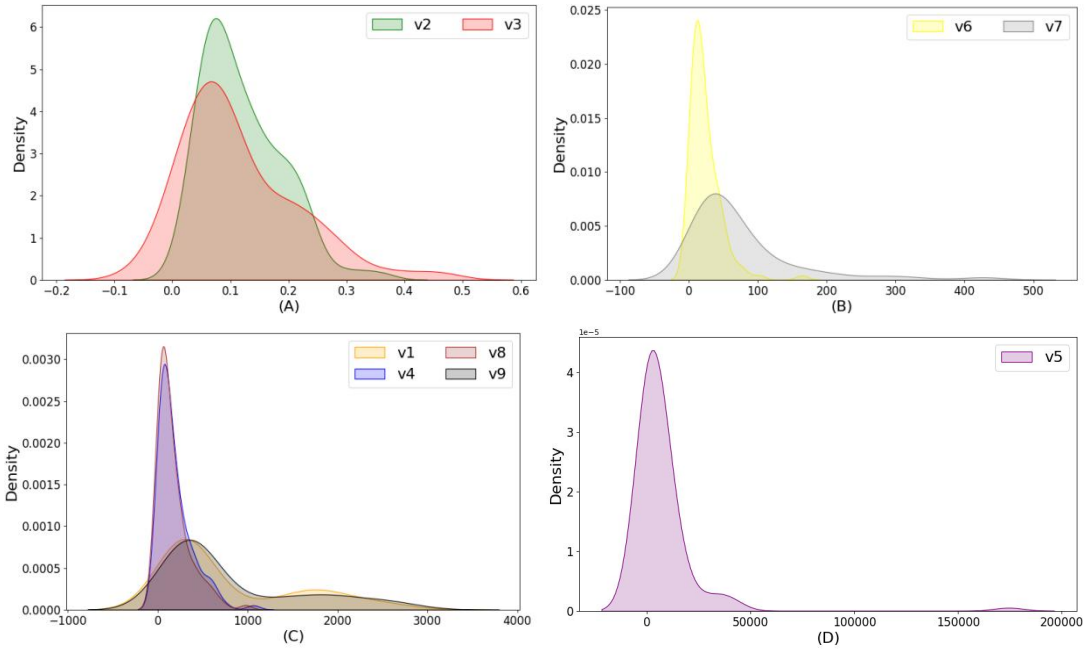

图5：DOS-IN原文实验所用数据集中BreastTissue的9个变量的密度曲线

由图5可见，BreastTissue数据集中有两个变量的核密度函数自变量取值范围在[-0.2,0.6]之间，有两个变量在[-100,500]之间，有四个变量在[-1000,4000]之间，还有一个变量在[-10000,200000]之间，不同变量的量纲差异非常大。但其论文的实验中并没有对数据进行特征工程等预处理，而是直接进行DOS-IN聚类。因此，本次复现的第一个优化操作是——对实验数据进行特征工程的数据预处理。

4.1.2 优化小簇合并机制

本次复现的第二个优化操作，是对DOS-IN算法引入的小簇合并机制进行优化。论文及其开源代码中所用的合并机制是找出距离该对象最近的初始大簇对象，合并到该大簇对象所在簇。但由于边界对象分布的散乱且有可能与其他簇的边界出现重叠，这样的合并机制显然缺乏合理性。因此本次复现时对这一机制进行了优化，先计算每个初始大簇的簇中心，通过寻找小簇中每个对象对应的最近大簇簇中心来合并小簇中的对象。

4.2 实验设置

4.2.1 实验数据

本次复现从DOS-IN原文[2]的实验数据中选择了9个真实数据集(BreastTissue, lenses, zoo, fertility_Diagnosis, led7digit, skewed, ecoli, Spriral, divorce)进行了实验。

4.2.2 聚类有效性评价指标

本文将沿用原文[2]中所使用的聚类有效性评价指标：NMI（Normalized Mutual Information，归一化互信息）和RI（Rand Index，兰德系数），其计算公式分别由式(3)和式(4)给出。

$$\mathrm{NMI}(\mathrm{U},\mathrm{V}) = \frac{\sum_{i=1}^{|V|}\sum_{j=1}^{|U|}\mathrm{P}(i,j)\log\frac{\mathrm{P}(i,j)}{\mathrm{P'}(j)\log\mathrm{P'}(j)}}{\sqrt{\sum_{i=1}^{|U|}\mathrm{P}(i)\log\mathrm{P}(i)\times\sum_{j=1}^{|V|}\mathrm{P'}(j)\log\mathrm{P'}(j)}} \tag{3}$$

其中，U为真实分类标签向量，$\mathrm{U_i}$为第i个样本的真实标签，V为聚类结果向量，$\mathrm{V_i}$为第i个样本的聚类结果，N为样本数量，$\mathrm{P}(i) = \frac{|\mathrm{U_i}|}{\mathrm{N}}, \mathrm{P'}(i) = \frac{|\mathrm{V_i}|}{\mathrm{N}}, \mathrm{P}(i,j) = \frac{|\mathrm{U_i}\cap\mathrm{V_j}|}{\mathrm{N}}$. NMI的取值范围在[0,1]之间，其值越接近于1，聚类结果越准确。

$$RI(U,V) = \frac{a+b}{\binom{N}{2}} \tag{4}$$

其中，样本中取任意两个样本点组成样本点对，共有$\binom{N}{2}$个样本点对，a是样本点对中的两个样本点真实分类标签为同类且在聚类结果中也为同一类的样本点对数量，b是样本点对中的两个样本点真实分类标签为不同类且在聚类结果中也为不同类的样本点对数量。RI的取值范围在[0,1]之间，其值越接近于1，聚类结果越准确。

5 实验结果分析

5.1 对BreastTissue数据集的聚类实验结果展示

本节将通过本次复现实验数据集中的BreastTissue数据集的实验结果，展示5种特征工程数据预处理操作以及小簇合并机制优化操作对DOS-IN聚类算法的优化作用。

5.1.1 原论文中对BreastTissue数据集的聚类实验结果

原论文中DOS-IN算法在对数据集BreastTissue聚类时的聚类有效性指标如图6所示，NMI=0.39，RI=0.64，聚类所用时间为0.0314秒。

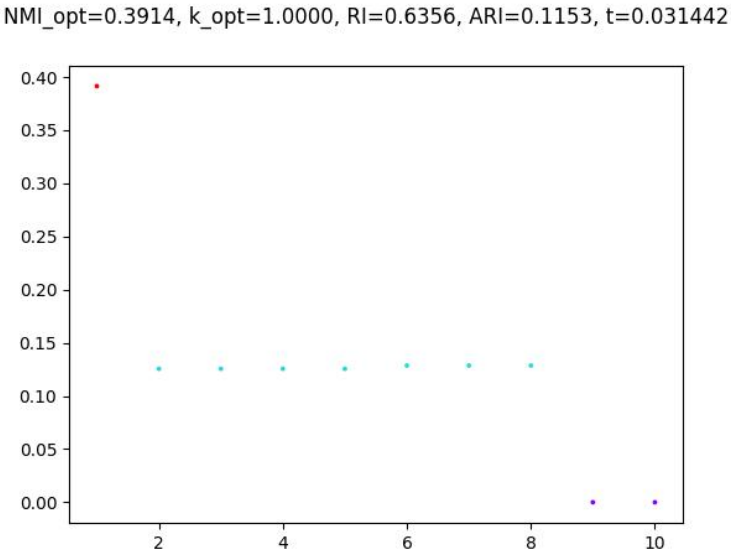

图6：原文中DOS-IN对数据集BreastTissue聚类的有效性指标值

5.1.2 通过Min-Max缩放进行预处理后的聚类实验结果

Min-Max缩放是一种常见的数据归一化方法，它将所有的数据压缩（或拉伸）到[0,1]的范围内，其缩放公式如式(5)：

$$\tilde{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

这种缩放方式只和样本中的最大值$max(x)$和最小值$min(x)$有关，可以消除量纲对最终结果的影响，使不同变量具有可比性[3]。

Min-Max缩放预处理后的BreastTissue数据集各变量的核密度曲线如图7所示(注意各变量值均在[0,1]之间，但其核密度曲线的横坐标取值不一定为[0,1])。

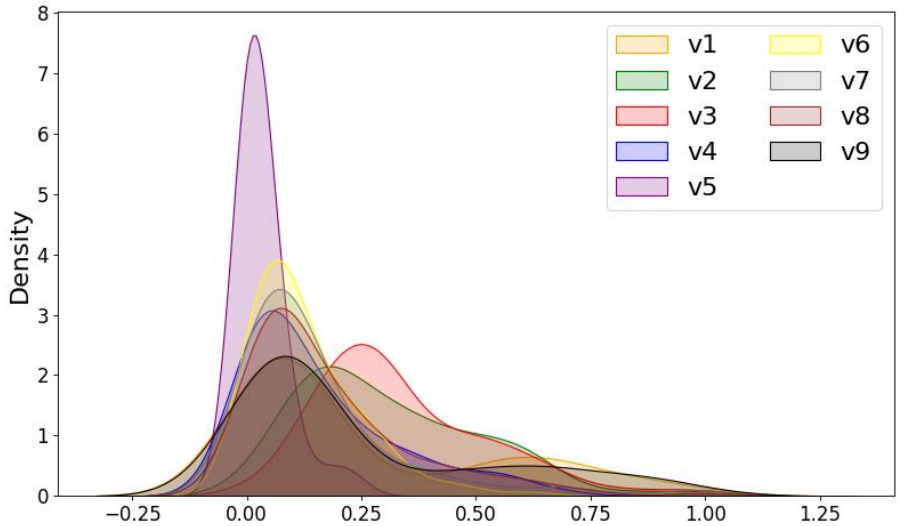

图7：Min-Max缩放后BreastTissue数据集各变量的核密度曲线

由图7可见，Min-Max缩放有效消除了各变量间量纲的差异。

经过Min-Max缩放的预处理后，DOS-IN算法对数据集BreastTissue聚类时的聚类有效性指标如图8所示，NMI=0.57，RI=0.86，聚类用时为0.0072秒。

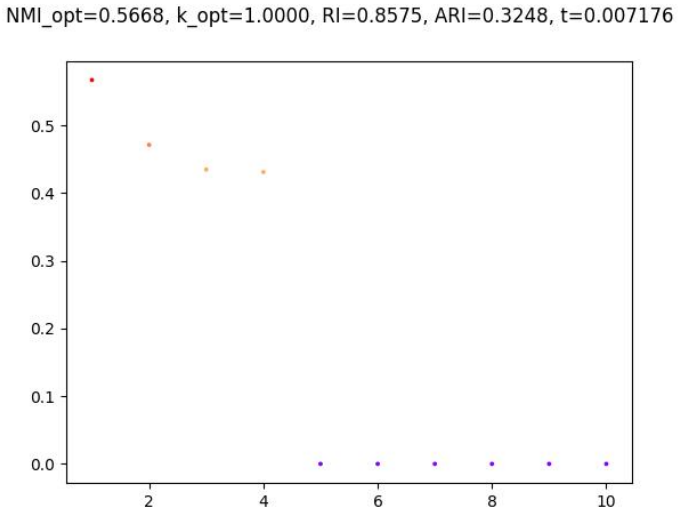

图8：经过Min-Max缩放后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中直接进行DOS-IN聚类的聚类有效性，NMI提高了45%，RI提高了35%.

5.1.3 通过方差缩放进行预处理后的聚类实验结果

方差缩放是根据原始数据的均值和标准差进行标准化，将数据转化为均值为0，标准差为1的新数据，其缩放公式如式(6)：

$$\tilde{x} = \frac{x - mean(x)}{SD(x)} \tag{6}$$

这种缩放方式和样本的每个值都有关，通过样本总体的均值$mean(x)$和标准差$SD(x)$体现，其产生的新数据有相同的方差，因此高权重不会分配给具有较高方差的变量。其输出范围与数据的标准差相关，它表示的是原始值与均值之间差多少个标准差，是一个相对值，有去除量纲的作用。

方差缩放预处理后的BreastTissue数据集各变量的核密度曲线如图9所示。

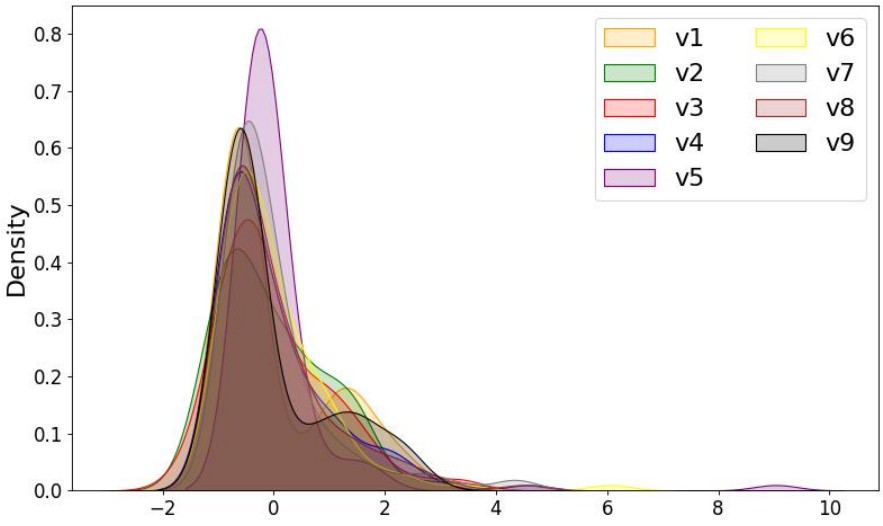

图9：方差缩放后BreastTissue数据集各变量的核密度曲线

由图9可见，方差缩放也能有效消除变量间量纲的差异，同时，方差缩放产生的新数据有相同的方差，这可能导致变量方差中所包含的部分信息损失。

经过方差缩放的预处理后，DOS-IN算法对数据集BreastTissue聚类时的聚类有效性指标如图10所示，NMI=0.51，RI=0.84，聚类用时为0.0089秒。

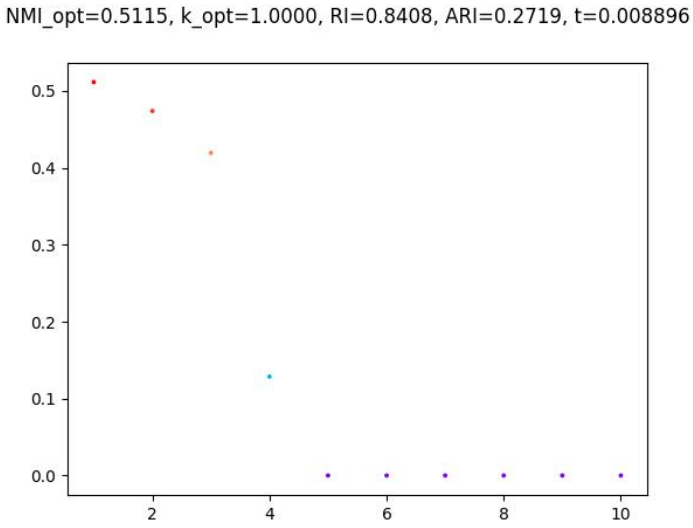

图10：经过方差缩放后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中直接进行DOS-IN聚类的聚类有效性，NMI提高了31%，RI提高了32%.

5.1.4 通过L2缩放进行预处理后的聚类实验结果

L2范数（也称为欧几里得范数）度量向量在坐标空间中的长度，而L2标准化则是通过L2范数标准化原始特征值，其变换公式如式(7)：

$$\tilde{x} = \frac{x}{\|x\|_2} \tag{7}$$

其中$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$，$n$为样本数量。L2标准化与样本的每个值都有关，通过L2范数体现，有去除变量量纲的作用。同时，这一技术没有对样本数据进行平移，即没有在原始特征值中减去一个量，因此能够使用在稀疏数据上。

L2标准化预处理后的BreastTissue数据集各变量的核密度曲线如图11所示。

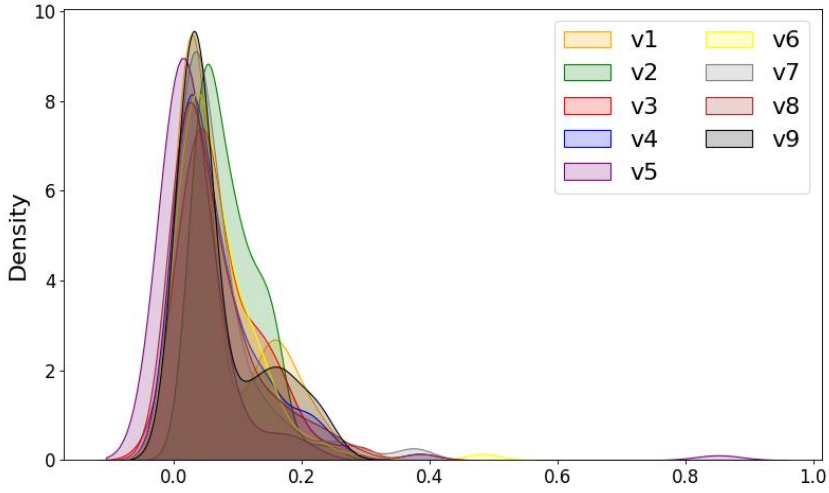

图11：L2标准化后BreastTissue数据集各变量的核密度曲线

通过图11可以发现，L2标准化有效消除了各变量间量纲的差异，同时削弱了各变量间方差的差异。

经过L2标准化的预处理后，DOS-IN算法对数据集BreastTissue聚类时的聚类有效性指标如图12所示，NMI=0.52，RI=0.84，聚类用时为0.0072秒。

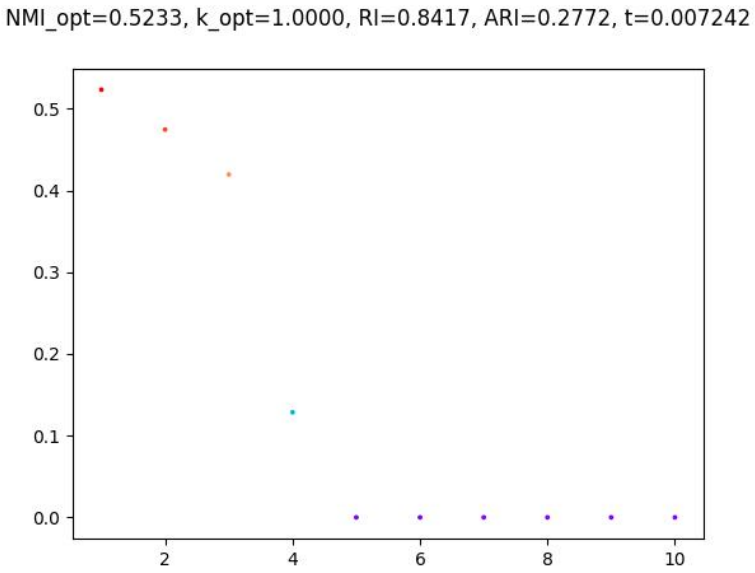

图12：经过L2标准化后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中直接进行DOS-IN聚类的聚类有效性，NMI提高了34%，RI提高了32%.

5.1.5 通过RobustScaler进行预处理后的聚类实验结果

RobustScaler，又称鲁棒标准化，它根据中位数和四分位数范围对数据进行缩放，确保每个特征的统计属性都位于同一范围。其变换公式如式(8)：

$$\tilde{x} = \frac{x - x_{50}}{x_{75} - x_{25}} \tag{8}$$

其中，x_{50}为中位数，x_{75}为75百分位数，x_{25}为25百分位数。RobustScaler在处理存在异常值的数据标准化时更具鲁棒性，能够获得更健壮的特征缩放结果。

RobustScaler预处理后的BreastTissue数据集各变量的核密度曲线如图13所示。

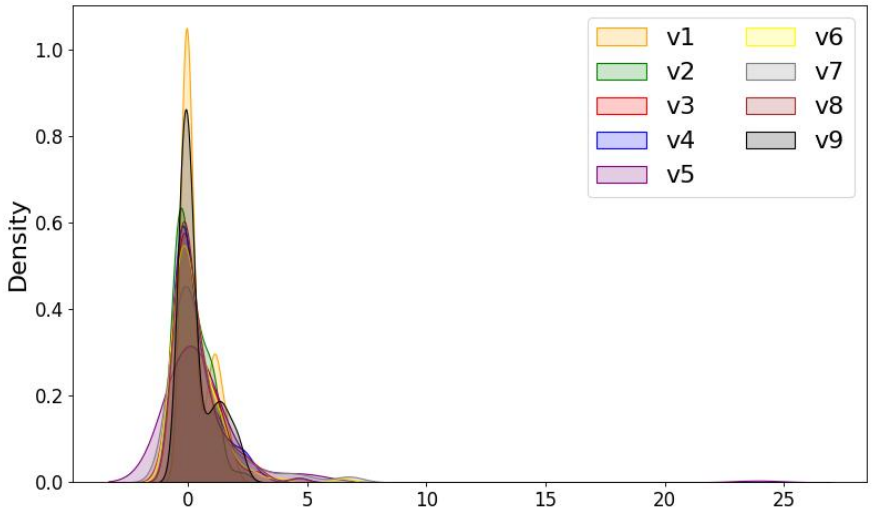

图13：RobustScaler后BreastTissue数据集各变量的核密度曲线

通过图13可以发现，RobustScaler可以有效消除各变量间量纲的差异。

经过RobustScaler的预处理后，DOS-IN算法对数据集BreastTissue聚类时的聚类有效性指标如图14所示，NMI=0.52，RI=0.73，聚类用时为0.0115秒。

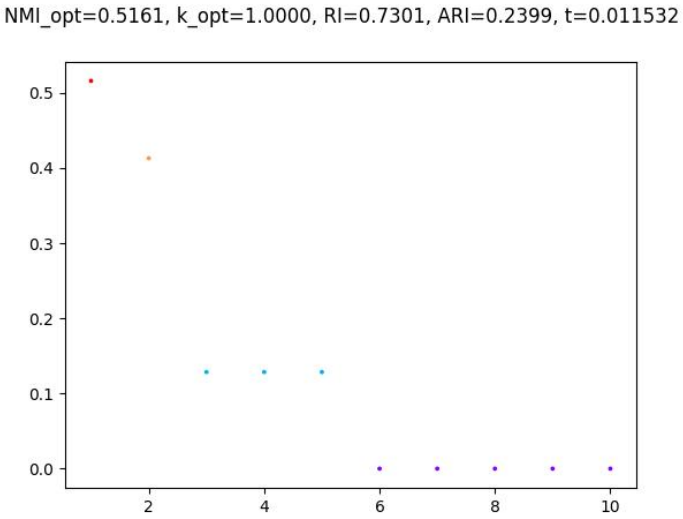

图14：经过RobustScaler后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中直接进行DOS-IN聚类的聚类有效性，NMI提高了32%，RI提高了15%.

5.1.6 通过主成分分析法进行预处理后的聚类实验结果

主成分分析法（Principal Components Analysis, PCA），又称主分量分析技术，是一种常用的降维技术，旨在通过线性变换将高维数据转换为低维数据，同时保留原始数据中的最大方差信息。其目标是找到数据中的主成分，这些主成分是原始数据中变化最大的方向，即方差最大的方向[4]。

本节通过主成分分析法对BreastTissue数据集进行降维，该数据集各变量间相关系数矩阵热力图如图15所示。

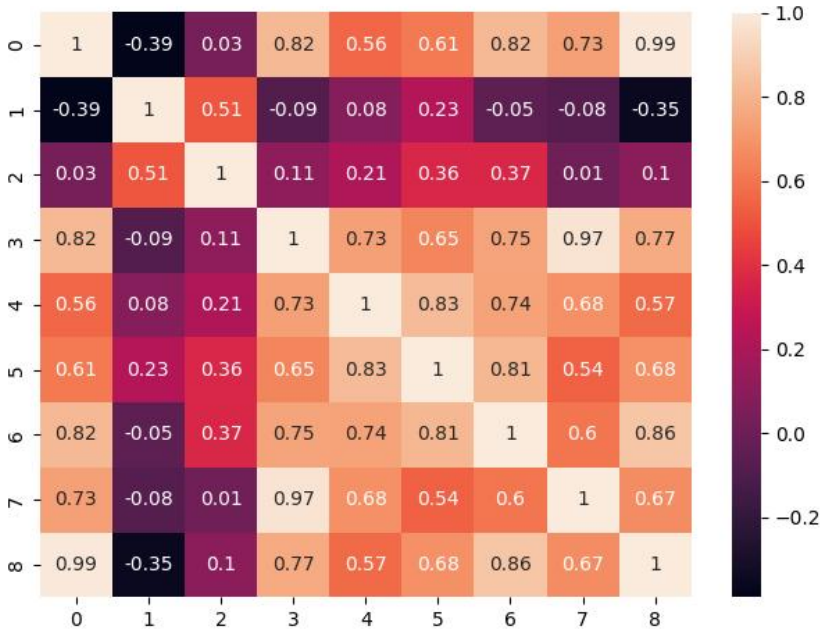

图15：BreastTissue数据集各变量间相关系数矩阵热力图（其中0-8表示第1-9个变量）

对BreastTissue数据集标准化后进行主成分分析并计算每个主成分能解释的方差的百分比，通过图16所示的方差解释累计百分比图，判断所需的主成分数量。

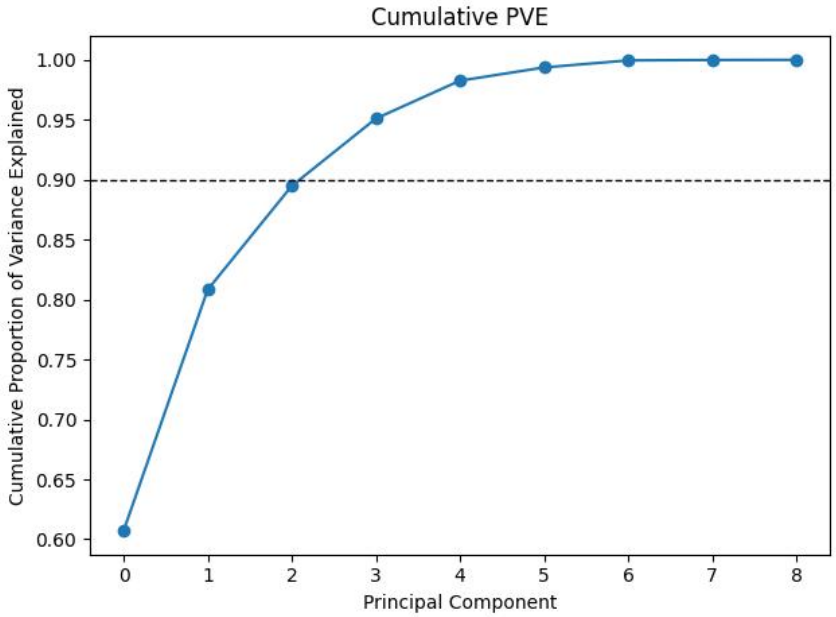

图16：BreastTissue数据集主成分的方差解释累计百分比（其中0-8表示第1-9个主成分）

由图16发现，只需要前3个主成分就可以解释数据中90%的方差，因此将原始数据中的9个变量减少为按照主成分核载矩阵对原始变量进行线性变换所产生的前3个主成分，通过这3个主成分进行DOS-IN聚类，其聚类结果的有效性指标如图17所示，NMI=0.52，RI=0.83，聚类用时为0.0071秒。

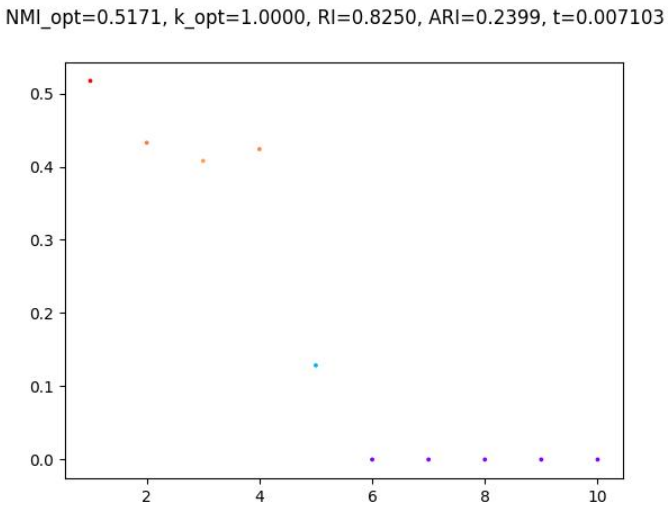

图17：经过主成分分析法降维后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中直接进行DOS-IN聚类的聚类有效性，NMI提高了32%，RI提高了30%．由于主成分分析时首先进行了数据标准化（方差缩放），通过主成分分析法预处理后的结果与仅对数据进行方差缩放的结果对比可以发现，主成分分析法预处理后的聚类有效性相较方差缩放预处理后的聚类有效性差异不大，可见主成分分析法在降维的同时有效保留了变量的更多信息，而聚类有效性提升的来源主要源于主成分分析前的方差缩放操作，而非主成分分析的核心操作。但在聚类所用的时间上主成分分析法预处理后相较直接进行DOS-IN聚类或只进行方差缩放再聚类都有所改进，是本文所有实验中用时最少的。对比直接进行DOS-IN

聚类所用聚类时间优化了77%，对比仅进行方差缩放后再DOS-IN聚类所用聚类时间优化了20%，可见主成分分析法能够有效降低DOS-IN聚类分析的计算负担。

5.1.7 优化小簇合并机制后的聚类实验结果

优化小簇合并机制后，DOS-IN算法对数据集BreastTissue聚类时的聚类有效性指标如图18所示，NMI=0.42，RI=0.65，聚类用时为0.0115秒。

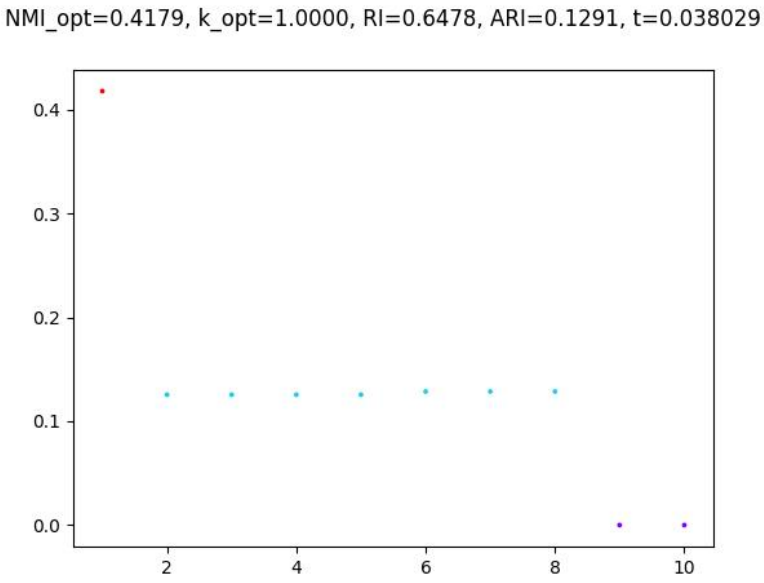

图18：优化小簇合并机制后DOS-IN对数据集BreastTissue聚类的有效性指标值

对比原论文中DOS-IN聚类的聚类有效性，NMI提高了7%，RI提高了2% .由于该机制只用于处理初始聚类结果中的小簇，而小簇在总体中所占比例本就较少，所以虽然对总体的优化百分比相对较小一点，但也说明这一优化对小簇的分类是合理有效的。

5.2 在9个真实数据集上的聚类优化实验结果

综合各类特征工程方法在9个真实数据集上的数据预处理稳定性来看，最终选择了Min-Max缩放对数据进行预处理。通过该预处理后，有4个数据集的聚类有效性有所提升，4个数据集的有效性不变，只有1个数据集的有效性略微有所降低。可见对数据进行Min-Max缩放处理是有效的，提升效果如表1所示。

表1：Min-Max缩放预处理后对比直接DOS-IN聚类的优化百分比

数据集	NMI优化百分比	RI优化百分比
breastTissue	45%	35%
lenses	13%	10%
zoo	5%	6%
fertility_Diagnosis	5%	0%
skewed	0%	0%
led7digit	0%	0%
Spriral	0%	0%
divorce	0%	0%
ecoli	-5%	-1%

优化小簇合并机制后，有2个数据集的聚类有效性有所提升，5个数据集的有效性不变，在2个数据集上的聚类有效性略微有所降低，提升效果如表2所示。

表2：优化小簇合并机制后对比原文DOS-IN聚类的优化百分比

数据集	NMI优化百分比	RI优化百分比
breastTissue	7%	2%
ecoli	2%	1%
lenses	0%	0%
zoo	0%	0%
fertility_Diagnosis	0%	0%
Spriral	0%	0%
divorce	0%	0%
skewed	-2%	0%
led7digit	-4%	-1%

结合本节上述两种优化操作后，提升效果如表3所示。

表3：结合两种优化操作后对比原文DOS-IN聚类的优化百分比

数据集	NMI优化百分比	RI优化百分比
breastTissue	44%	35%
lenses	13%	10%
zoo	6%	6%
fertility_Diagnosis	5%	0%
ecoli	4%	1%
Spriral	0%	0%
divorce	0%	0%
skewed	-2%	0%
led7digit	-4%	-1%

由表3可见，通过本次复现的两步优化操作，大多数数据集聚类有效性存在显著提升，有的数据集聚类有效性不变，只有少数数据集的聚类有效性出现小幅下降，总体而言，本次复现优化是有效的。

6 总结与展望

通过特征工程的方法对数据进行预处理后，能有效提高基于距离的聚类算法的聚类有效性、降低聚类算法的计算负担。在本文进行的特征工程预处理后使用DOS-IN聚类算法对BreastTissue数据集进行聚类的结果显示，Min-Max缩放对该数据的预处理效果最好，相较直接进行DOS-IN聚类，NMI提高了45%，RI提高了35%；L2缩放效果次之，NMI提高了34%，RI提高了32%；RobustScaler效果相对其他特征工程较差，NMI提高了32%，RI提高了15%；方差缩放和主成分分析法效果相近，分别为NMI提高了31%，RI提高了32%和NMI提高了32%，RI提高了30%. 同时，主成分分析法预处理后有效降低了DOS-IN聚类的计算负担，对比直接进行DOS-IN聚类所用聚类时间优化了77%. 可见Min-Max缩放、方差缩放、L2标准化、主成分分析法这四种特征工程方法都是对DOS-IN聚类算法有效的优化方法。

通过优化小簇合并机制，对比原论文中DOS-IN聚类的聚类有效性，NMI提高了7%，RI提高了2% .由于该机制只用于处理初始聚类结果中的小簇，所以虽然对总体的优化百分比相对较小一点，但也说明这一优化对小簇的分类是合理有效的。

结合特征工程数据预处理和优化小簇合并机制两种优化操作后，实验的9个真实数据集中，超过一半的数据集聚类有效性都存在显著提升，只有极少数数据集的聚类有效性出现小幅下降，总体而言，本次复现优化是有效的。

本次复现仍存在少数数据集的优化实验结果不太理想，后续可以尝试通过更多类型的数据实验，总结本次复现提出的两种优化方法，针对哪些数据集有较好的优化效果，对哪些数据集的优化效果较差。

参考文献

[1] Shuliang Wang,Qi Li,Hanning Yuan,等.δ-Open set clustering—A new topological clustering method[J].Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4):e1262.

[2] Qi L ,Guochen Y ,Shuliang W , et al.A novel open-set clustering algorithm[J].Information Sciences,2023,648

[3] 罗向阳,王道顺,汪萍等.基于图像多域特征缩放与BP网络的信息隐藏盲检测[J].东南大学学报(自然科学版),2007,(S1):87-91.

[4] 庞博文,李治军.基于PCA-GA-XGBoost模型的吉林省水资源承载力评价[J/OL].人民珠江,1-14[2024-01-04]http://kns.cnki.net/kcms/detail/44.1037.TV.20231228.1401.012.html.