Journal of System Simulation

Volume 35 | Issue 4 Article 10

4-12-2023

I-niceMO Enhanced Algorithm Based on Intersection Angel Geometry

Yifan He
1.College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China;, 396981852@qq.com

Yulin He
1.College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China;2.National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518000, China;, yulinhe@gml.ac.cn

Yongda Cai
1.College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China;

Zhexue Huang
1.College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China;2.National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518000, China;

I-niceMO Enhanced Algorithm Based on Intersection Angel Geometry

Abstract

Abstract: To exactly determine the number of cluster centers and correctly identify the candidate cluster centers, an I-niceMO enhanced(I-niceMOEn) algorithm based on intersection angel geometry is proposed. *As many distributions of intersection angles and distances as possible between observation points and data points are utilized to recognize the candidate cluster centers to avoid the neglection of cluster centers. The spectral clustering algorithm is used to automatically merge the candidate cluster centers according to the eigenvalues of Laplacian matrices. The number of final cluster centers is determined by the number of merged candidate cluster centers.* The number of clusters can be automatically determined by I-niceMOEn algorithm and the manual parameter input for clustering is not needed. The experimental results show that I-niceMOEn algorithm is convergent and outperforms the traditional automatic clustering methods and I-niceMO algorithm.

Keywords

automatic clustering, I-nice clustering, spectral clustering, unsupervised learning, observation point mechanism

Recommended Citation

Yifan He, Yulin He, Yongda Cai, Zhexue Huang. I-niceMO Enhanced Algorithm Based on Intersection Angel Geometry[J]. Journal of System Simulation, 2023, 35(4): 797-808.

第 35 卷第 4 期 系统仿真学报© Vol. 35 No. 4
2023 年 4 月 Journal of System Simulation Apr. 2023

基于夹角几何的I-niceMO增强算法

何一帆[1]，何玉林[1,2*]，蔡湧达[1]，黄哲学[1,2]
(1. 深圳大学 计算机与软件学院，广东 深圳 518060；2. 深圳大学 大数据系统计算技术国家工程实验室，广东 深圳 518000)

摘要：针对I-niceMO算法在候选聚类中心合并时中心数目难以确定和中心点识别不准确的问题，提出了基于夹角几何的I-niceMO增强(I-niceMOEn)算法。*利用观测点与数据点之间的距离和角度分布情况找出数据中尽可能多的候选聚类中心，以避免多类别数据聚类中出现的类别丢失的情况；利用谱聚类算法对候选聚类中心进行聚类，根据拉普拉斯矩阵特征值的大小自动地对候选聚类中心进行合并；根据合并后的聚类中心的数量确定最终的数据聚类类别数。*I-niceMOEn算法实现了对数据类别数的自动确定，并且在聚类过程中不需要人为设置参数。实验结果表明：I-niceMOEn算法在收敛的同时能够获得优于传统自动聚类算法和I-niceMO算法的类中心确定表现。

关键词：自动聚类；I-nice聚类；谱聚类；无监督学习；观测点机制
中图分类号：TP391 文献标志码：A 文章编号：1004-731X(2023)04-0797-12
DOI: 10.16182/j.issn1004731x.joss.21-1333
引用格式: 何一帆, 何玉林, 蔡湧达, 等. 基于夹角几何的I-niceMO增强算法[J]. 系统仿真学报, 2023, 35(4): 797-808.
Reference format: He Yifan, He Yulin, Cai Yongda, et al. I-niceMO Enhanced Algorithm Based on Intersection Angel Geometry[J]. Journal of System Simulation, 2023, 35(4): 797-808.

I-niceMO Enhanced Algorithm Based on Intersection Angel Geometry

He Yifan[1], *He Yulin*[1,2*], *Cai Yongda*[1], *Huang Zhexue*[1,2]
(1. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China;
2. National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518000, China)

Abstract: To exactly determine the number of cluster centers and correctly identify the candidate cluster centers, an I-niceMO enhanced(I-niceMOEn) algorithm based on intersection angel geometry is proposed. *As many distributions of intersection angles and distances as possible between observation points and data points are utilized to recognize the candidate cluster centers to avoid the neglection of cluster centers. The spectral clustering algorithm is used to automatically merge the candidate cluster centers according to the eigenvalues of Laplacian matrices. The number of final cluster centers is determined by the number of merged candidate cluster centers.* The number of clusters can be automatically determined by I-niceMOEn algorithm and the manual parameter input for clustering is not needed. The experimental results show that I-niceMOEn algorithm is convergent and outperforms the traditional automatic clustering methods and I-niceMO algorithm.

Keywords: automatic clustering; I-nice clustering; spectral clustering; unsupervised learning; observation point mechanism

收稿日期：2021-12-23 修回日期：2022-02-09
基金项目：国家自然科学基金面上项目(61972261)；深圳市基础研究项目(JCY20210324093609026，JCY120200813091134001)
第一作者：何一帆(1997-)，男，硕士生，研究方向为数据挖掘和机器学习算法及其应用。E-mail：396981852@qq.com
通讯作者：何玉林(1982-)，男，副研究员，博士，研究方向为大数据近似计算技术、多样本统计分析理论与方法。E-mail：yulinhe@gml.ac.cn

0 引言

聚类算法一直是重要的数据分析技术，即根据数据之间一定的相似性将数据分割成2个或多个簇的过程，其中簇内数据尽量相似，而不同簇间的数据差异尽可能显著[1-2]。聚类算法是无监督学习中一个重要的研究方向，多种聚类算法也逐步问世[3-4]。但大多聚类算法需要预先输入簇的个数作为参数来运行算法，然而在数据分析中类的数量是不确定的，需要对数据里簇的数量做出估计[5-6]。不同的估计对聚类算法运行结果有着显著影响，因此，对数据集簇数的估计成为了一个难点。虽然目前很多估计类别个数的算法已被提出，但他们并不能得到满意的结果，或者难以在使用中实现。如Elbow方法[7]、Silhouette方法[8]等算法的效果并不理想。Elbow方法用不同的簇数作为假设对数据集进行K-means聚类[9-10]并评估聚类效果，这一方法的结果需要人为观测斜率并且所得结果也不准确。Silhouette方法对簇结构为凹和簇结构为凸的簇的计算结果不同，导致结果不准确，并且时间复杂度高。

Masud等提出了I-niceSO算法[11-12]，并进一步提出了I-niceMO算法，其效果明显超越了当时的众多聚类算法，但在观测和合并中心点仍有一定缺陷。I-niceMO算法在数据空间中设置多个观测点，计算该观测点到所有数据点的距离，并将这个距离分布拟合为混合高斯分布。其中分量数被认定为该次观测所发现的簇数，并记录每个簇中数据密度最大的数据点作为候选中心点。最后将这些候选中心点根据它们之间的距离选取一定的比例作为合并阈值将它们合并。

I-niceMO算法中每个观测点在观测时会将距离相近的多个簇错误地认为是同一个，并且在最后对多个候选中心点进行合并时需要人为地调整合并阈值，不能自动地判断簇的数量。本文提出一种I-niceMO的改进算法解决了这2个问题。每个观测点在观测数据点的距离之后增加了对数据点和原点夹角的观测，使得每个观测点能够发现更多的簇，并且使用了谱聚类[13-16]的方法来合并所有候选中心点，这样就实现了在没有人为调整参数的情况下自动判断数据集的簇数。

1 相关工作

数据集的簇数和每个簇的初始中心点的估计是数据分析中一个非常重要的问题[17]，作为许多聚类算法的参数，簇数和初始中心点设定的好坏对聚类算法的性能有很大影响。Elbow方法是一种经典的自动确定聚类类别数的方法，已经应用于k-means聚类中。该方法设置聚类类别数从小到大依次变化，并计算每一个聚类类别数划分下的类内误差平方和。通常情况下，当数据的划分类别合适的时候，类内误差平方和会取得一个较小的值。通过比较相邻2个聚类类别数的类内误差平方和的变化幅度可以确定聚类类别数。Elbow方法的缺点是存在一些数据的类内误差平方和变化不大的情况，而且，在大量数据的聚类中它的时间复杂度很高。轮廓系数法通过计算数据点与类内数据点的平均距离以及数据点与类间数据点的平均距离来计算系数值。轮廓系数值越大说明聚类效果越好。该方法通过求解不同的聚类类别数设置下的轮廓系数，并选取最大系数对应的值作为最优的结果。这种方法的缺点是对一些复杂结构的数据效果不明显。

为了解决Elbow方法需要人为观测拐点的问题，Hastie等[18]提出了Gap statist确定聚类类别数的方法。该方法通过蒙特卡罗方法产生与原始数据数量相等的均匀分布的随机样本，对测试数据进行不同聚类类别数下的k-均值聚类。然后求随机样本损失与实际样本损失之差，取最大差值所对应的K值就是最佳簇数。然而，Gap statist在多类别数据所得的结果并不理想。由于基于图的方法对任意形状的数据都具有良好的聚类效果，卢鹏丽等[19]提出了基于谱聚类的自动确定聚类数目的方法。该方法通过计算拉普拉斯矩阵特征值并求出降序排列的相邻特征值之间的差值来判断聚类类别数。当差值最大时，前面的小特征值的数量即为聚类数目。然

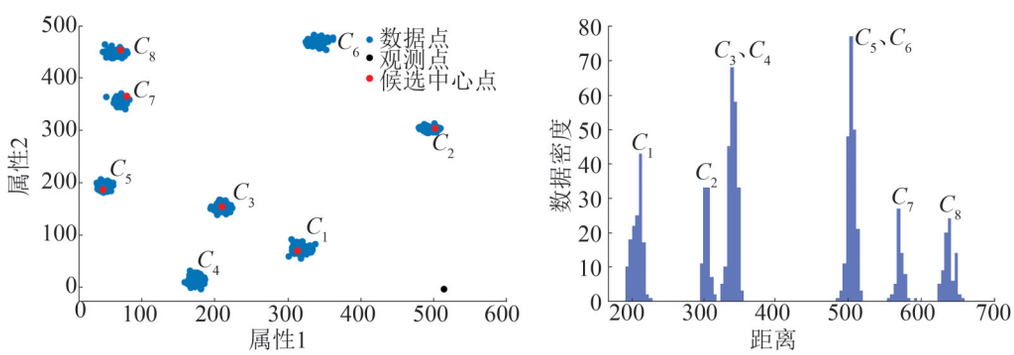
DOI: 10.16182/j.issn1004731x.joss.21-1333

而，这种方法受初始构图的影响很大。在多个类别数据之间的距离靠近的时候相邻特征值之间的差值很小，很难找到一个准确的类别数。

基于密度的聚类方法受到广泛关注[20]。Campello等[21]提出了使用数据点附近其他数据点个数作为密度估计的方法评估该数据点成为一个簇的中心点的可能性大小，但这类方法需要预先设定“近邻”的距离范围和区别中心点与边缘点的近邻个数作为阈值，所以在实际应用中并不理想。Rodriguez等[22]提出了一种基于密度的确定聚类类别数目的方法。该方法通过计算每个数据点的密度和数据点与其最近的高密度点的距离生成一个决策图。一个合适的聚类中心通常具有高密度的特征并且与其他高密度点的距离较远。依据这个准则，可以通过决策图来确定聚类类别数。然而，在混合属性数据中这种方法的效果比较差。在此基础上，陈晋音等[23]提出了基于密度的聚类中心自动确定的混合属性数据的聚类算法。该方法通过粒子群算法计算与数据密度有关的参数以获得更好的密度测量效果。然而，基于密度的方法需要人为设置参数去确定决策图中聚类类别数。Masud等[24]也使用密度估计的方法以改进K-means算法需要初始中心点的缺陷，提出了IK-means算法，但难以应用。Masud等提出了I-niceSO算法，即在数据空间中设置观测点，计算每个数据到观测点的距离，根据不同距离的不同数据密度，将他们拟合成混合高斯模型，将模型中的构件数量当作簇数量的估计。I-niceMO算法设计了多个观测点分别观测，最终将观测结果汇总得到更加精准的结果。

2 I-niceMO算法分析

I-niceMO算法类似于一个观察者观测周围的山峰。以一个人工合成的二维数据集为例，将800个二维数据用蓝色描绘在对应位置，并用黑色描绘了观测点的位置。计算所有数据点到观测点的距离，并计算不同距离的数据密度，拟合成混合高斯模型。该模型根据“山峰”的数量估计簇的数量，再从每个簇中选取密度最大的数据点作为该簇的中心点。这样单个观测点得到的结果并不精准，就如我们在图1中看到的有2个簇C_4和C_6不能被识别，因为该簇与另一个簇到观测点的距离十分接近，以至于被观测点认为是同一个簇。如图2所示，根据观测点的距离分割数据空间后，C_3和C_4处于同一空间，C_5和C_6处于同一空间，从而导致了误判。在数据量极大的数据集里，这种情况十分常见。因此，I-niceMO算法在数据空间的边缘位置布置了多个不同位置的观测点以消除每个观测点的盲区所带来的误判。另外，为了解决这一问题，Azhar等[25]采用分层结构观测，其原理为根据距离把数据分割之后对每个区域布置另一位置的观测点继续观测和分割直至确认该区域只有一个簇为止。因为各子区域的数据类别相比原始数据大大减少，所以其观测效果得到提升。



(a) 观测点与样本点的距离　　(b) 距离分布

图1　I-niceMO

Fig. 1　I-niceMO algorithm

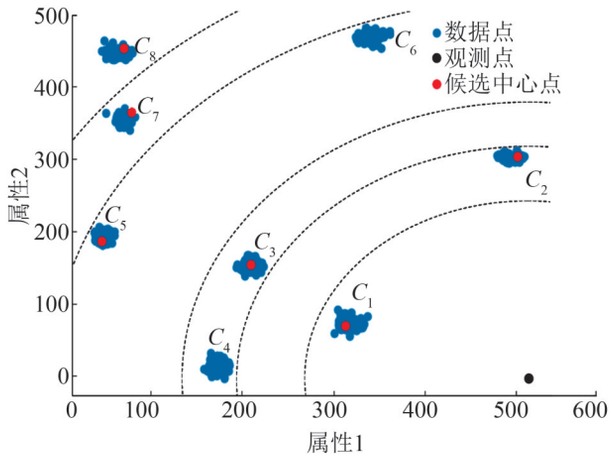


图2 I-niceMO算法的聚类结果
Fig. 2 Clustering result of I-niceMO algorithm

由于I-nice聚类是通过分析观测点与所有数据的距离分布情况进行聚类的，所以观测点位置的选取对聚类结果有非常重要的影响。然而，目前还没有成熟的观测点选取方法，I-niceMO聚类在观测点的选取上还存在2个问题：① I-niceMO聚类算法对观测点的位置比较敏感；②对于多类别的高维数据，I-niceMO在高度重合的距离分布中无法检测出所有类别。I-niceMO算法伪代码如下。

算法1：I-niceMO算法
输入：高维度数据集
输出：初始中心点，聚类结果
初始化：
在数据空间边缘生成 T 个观测点 $\{P_1, P_2, \cdots, P_T\}$；
为观测点设置最大分量数 $M_{\max}$；
观测阶段：
for t: = 1 to T do
计算观测点 P_t 到所有数据点的距离集合 X_t；
for M: = 1 to $M_{\max}$ do
将 X_t 拟合为 $GMM(t, M)$ 并对其使用EM算法计算拟合程度；
计算 $AICc$ 值；
end for
将 $AICc$ 值最小的混合高斯模型作为这次观测的结果，其中有 C 个高斯分量，数据也被分割至 C 个子集中；
for c: = 1 to C do
计算对应子集中每个数据点的数据密度；
选择数据密度最大的数据点作为一个候选中心点；
end for
end for
合并候选中心点阶段：
计算观测到的 K 个候选中心点两两之间的距离，生成相似矩阵 $\boldsymbol{d}(K)$；
从相似矩阵中选取一定比例的距离，取平均数作为阈值 $threshold$；
设置遍历数组 $\boldsymbol{visited}[K]=[0, 0, \cdots, 0]$；
设置队列 q；
for i: = 1 to K do
if $\boldsymbol{visited}[i]$ == 0 then
$\boldsymbol{visited}[i]$ = 1；
i 进入 q；
while q 不为空 do
q 的队头元素 j 出队；
for k: = 1 to K do
if distance(k, j) < $threshold$ and k! = j and $\boldsymbol{visited}[k]$ == 0 then
k 进入 q；
$\boldsymbol{visited}[k]$ = 1；
end if
end for
end while
将 i 设定为一个初始中心点；
end if
end for
得到 n 个初始中心点后，将数据点依据距离分割至 n 个子集中；

伪代码中合并候选中心点阶段采用了广度优先算法，使每个候选中心点去搜索遍历附近距离小于阈值的其他候选中心点，直至遍历结束，使所有候

DOI: 10.16182/j.issn1004731x.joss.21-1333

选中心点成为若干个连通分量，如图3所示。其中连通分量的数量作为对数据集中簇的个数估计。

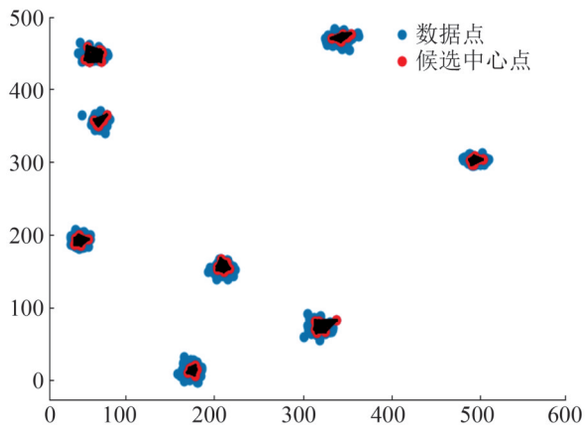


图3　广度优先算法遍历附近候选中心点
Fig. 3　Traversing nearby candidate centers by BFS

3　基于几何夹角的I-niceMO增强算法

为了解决上述问题，本文提出了融合距离、角度的I-niceMOEn聚类方法。通过距离密度的分布情况，只能发现与观测点距离不等的簇。当聚类数据比较复杂的时候存在多个簇与观测点距离相同或接近的情况，可使用角度进一步区分。

然而，通过聚类和角度的方法找出大量的候选聚类中心的同时也带来了聚类中心合并上的困难。传统的I-niceMO算法通过人为设置近邻阈值方法对聚类中心进行合并，在候选聚类中心很多的时候很容易出现类别的溢出。本文提出利用谱聚类方法对候选中心进行合并，根据拉普拉斯矩阵特征值的大小自动地确定聚类类别数。

3.1　观测点选取距离与角度观测

I-nice聚类算法是一种k-means类型的聚类方法。与k-means不同的是，它能够自动地确定聚类类别数和聚类中心。I-nice算法确定聚类类别数的原理为通过计算观测点与所有聚类数据的欧式距离来获取距离的分布数据；构建高斯混合模型来拟合分布数据并从中选择最优的混合模型；根据最优混合模型确定聚类类别数和聚类中心。本文提出了一种欧式距离和角度双重观测的I-nice聚类算法，改进了传统方法因只分析距离分布情况而不能很好地区分多类别数据的缺点。

假设数据集为$X=\{x_1,x_2,\cdots,x_N\}$，且每一条数据记录$x_i\in\mathbb{R}^D$，D为维度。随机选取一个观测点$p\in\mathbb{R}^D$，则每一个数据$x_i=(x_{i1},x_{i2},\cdots,x_{iD})$与观测点$p=(p_1,p_2,\cdots,p_D)$之间的欧式距离为

$$d(x_i,p)=\sqrt[2]{\sum_{j=1}^{D}(x_{ij}-p_j)^2} \tag{1}$$

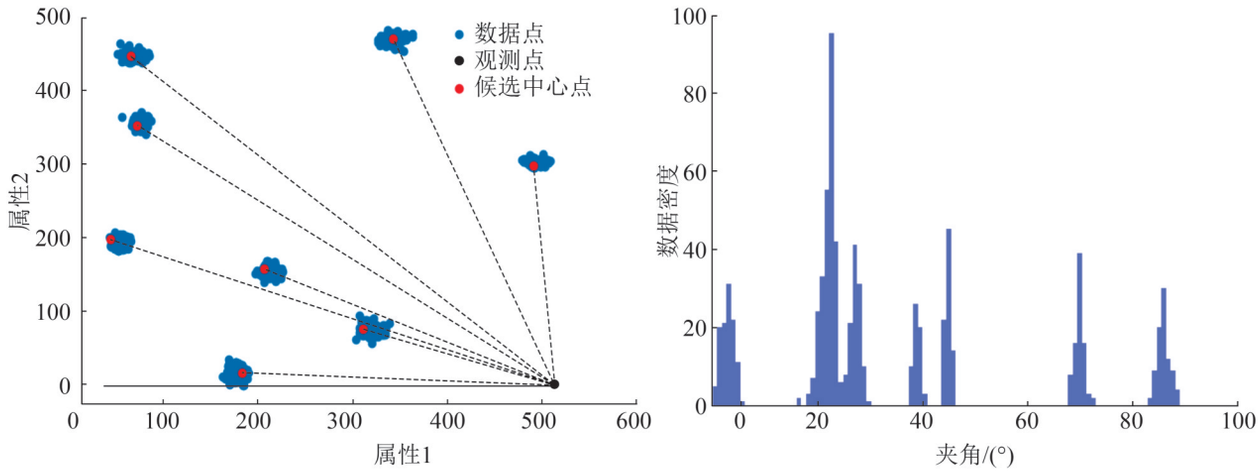
引言中提到，如果2个簇到观测点的距离非常接近，观测点会错误地将2个簇认为是1个簇。为了得到更好的观测效果，将观测点布置在数据空间的边缘以尽可能地避免这种情况。观测点对每一个数据点进行计算便可得到所有数据的距离向量$\boldsymbol{X}_d$。因为不同类别的数据通常与观测点的距离不同，所以可以通过分析距离向量的分布情况推算出聚类数据的类别数量。然而，距离只能反映观测点与数据之间的远近关系无法判断出方向。本文提出基于距离和角度的双重观测模型。首先，利用式(1)计算距离向量。然后以数据空间中最小值作为整个数据空间的原点O，则

$$O=(O_1,O_2,\cdots,O_D) \tag{2}$$

其中，$O_i=\min\{x_{1i},x_{2i},\cdots,x_{Ni}\}$，作观测点$p$到原点$O$的向量$\boldsymbol{a}=O-p$，观测点$p$到数据点$x$的向量$\boldsymbol{\beta}=x_i-p$，计算2个向量的夹角作为该观测点观测这个数据点的角度：

$$\theta(x_i,p)=\arccos\frac{a\cdot\beta}{|a|\cdot|\beta|} \tag{3}$$

如图4所示，以数据空间每一维度的最小值作为原点O，作观测点到原点和观测点到数据点的2个向量，并计算夹角。这样每一个数据点就有对应的夹角大小，类似地得到一个角度向量。可以根据夹角再次拟合混合高斯模型，以得到另一种观测结果从而发现原始数据中更多的簇。再将角度向量拟合为混合高斯分布，以发现在距离观测中未发现的簇。这样每个观测点综合2种不同的观测使得整个观测对于观测点位置不再那么敏感。



(a) 观测点与样本点的几何夹角　　(b) 几何夹角分布

图4　夹角几何I-niceMO

Fig. 4　Intersection angel geometry-based I-niceMO algorithm

3.2 拟合混合高斯模型

所有数据点对于一个观测点的距离或者角度向量的混合高斯模型定义为

$$P(d\,|\delta)=\sum_{k=1}^{M}\pi_k g(d\,|\delta_k),\quad d\geqslant 0 \tag{4}$$

式中：$P(d\,|\delta)$为在混合高斯模型为δ的假设下数值为d的数据点个数；δ为混合高斯模型中多个构件δ_k的集合；M为其分量数量；π_k为第k个构件的权重大小；$g(d\,|\delta_k)$为在分量δ_k的密度函数即高斯函数里数值为d的结果，δ_k代表该分量的均值和方差。任何模型应当满足$\sum_{k=1}^{M}\pi_k=1$。

可以设置最大的高斯分量参数$M_{\max}$，对一个向量进行多次不同构件数量的拟合，M从1变化到$M_{\max}$并求解每一个M所得到的混合高斯模型。

为了得到最适合的混合高斯模型，通过Akaike信息准则二阶变体来确定最优模型。该准则定义为

$$AICc=-2L(\delta^*)+2q\left(\frac{N}{N-q-1}\right) \tag{5}$$

式中：$L(\delta^*)$为高斯混合模型中最大似然函数值；$q=3M$即分量数量的3倍；N为聚类数据总量。具有使$AICc$值最小的高斯混合模型为最优模型。由此每个观测点p_t得到2个混合高斯模型：

$$GMM_t^{(1)}=\sum_{i=1}^{I}\pi_i^{(1)}N(\mu_i^{(1)},\sigma_i^{(1)}) \tag{6}$$

$$GMM_t^{(2)}=\sum_{j=1}^{J}\pi_j^{(2)}N(\mu_j^{(2)},\sigma_j^{(2)}) \tag{7}$$

式中：$GMM_t^{(1)}$为根据距离得到的混合高斯模型；$GMM_t^{(2)}$为根据角度得到的混合高斯模型；I和J为高斯分量个数；$N(\mu,\sigma)$为正态分布。不论是距离观测还是角度观测，最佳拟合结果中的分量数量就是该次观测所发现的聚类类别个数，同时记录结果中每个分量的峰值在整个向量中的位置，设定为候选中心点。

3.3 合并类中心

最优高斯混合模型中高斯分量的个数就是该次观测所发现的聚类数，每个簇当中的峰值记录为该簇的候选中心点，这样就避免了不同簇的数据之间的交集所带来的识别错误。如图5所示，所有观测点观测结束后，将所有候选中心点汇总$C=\{C_1,C_2,\cdots,C_Q\}$，不同观测点对同一个簇会得出不同的簇中心，用谱聚类的方法合并这些簇中心。

根据Kang等[26]提出的鲁棒图构造(RGC)的构图方法，先将C去噪后得到Z，得到相似矩阵

$$\boldsymbol{S}=(s_{ij})_{Q\times Q} \tag{8}$$

DOI: 10.16182/j.issn1004731x.joss.21-1333

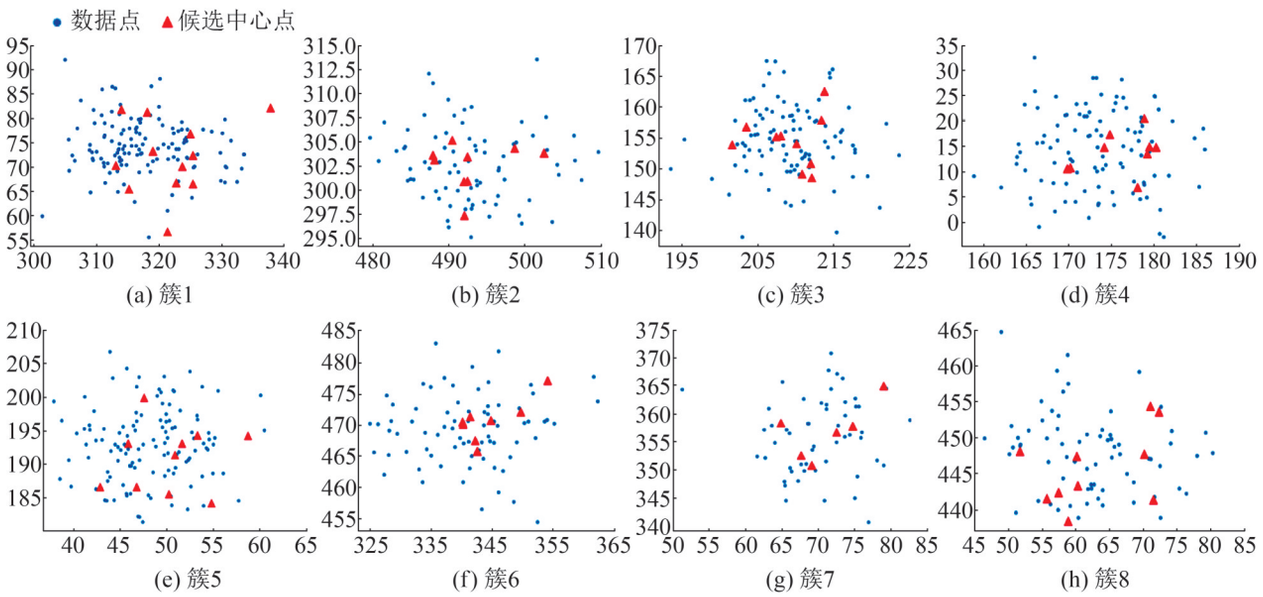


图5 不同簇中候选中心点

Fig. 5 Candidate centers in different clusters

$$s_{ij}=\begin{cases}\dfrac{f_{i,k+1}-f_{ij}}{kf_{i,k+1}-\sum_{r=1}^{k}f_{i,r}}, & f_{ij}<f_{i,k+1}\\ 0, \text{ otherwise}\end{cases} \tag{9}$$

式中：$f_{ij}=d^2(Z_i,Z_j)$；$f_{i,j}$为$f_{i1},f_{i2},\cdots,f_{in}$从小到大排序后的第$j$个分量；$k$为近邻数，一般设置为5。

一个用于分析出类别数的可行方法是分析拉普拉斯矩阵的特征值。设$\boldsymbol{S}$为相似矩阵，$\boldsymbol{D}$为度矩阵，$\boldsymbol{L}$为拉普拉斯矩阵，其中$\boldsymbol{D}$的对角线上元素为$\boldsymbol{S}$对应的行中所有元素的和，对角线外所有元素均为0，即

$$\boldsymbol{D}_{ii}=\sum_{j=1}^{Q}\boldsymbol{s}_{ij} \tag{10}$$

则拉普拉斯矩阵为

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S} \tag{11}$$

将其规范化后得到

$$\hat{\boldsymbol{L}}=\boldsymbol{D}\boldsymbol{D}^{-1}-\boldsymbol{S}\boldsymbol{D}^{-1}=\boldsymbol{I}-\boldsymbol{S}\boldsymbol{D}^{-1} \tag{12}$$

式中：$\boldsymbol{I}$为单位矩阵。

Zelnik-Manor等[27]发现拉普拉斯矩阵的特征值0的个数与图中连通分量的个数相同，即簇的数量，也等价于$\boldsymbol{S}\boldsymbol{D}^{-1}$这个矩阵的特征值1的个数，意味着可以通过特征值1的数量来估计数据的类别数。根据测试发现，若数据不同类之间有明显的区分，则这个方法是明显可行且正确的。但如果数据之间存在交错融合的情况，或者其中有干扰数据，特征值会偏离1，这便会给判断带来困难。

得到簇的个数之后，便可以用K-means的方法聚类所有的候选点。I-niceMOEn算法伪代码如下。其中计算混合高斯模型与算法1中操作相同，此处省略详细过程。

算法2：I-niceMOEn算法

输入：高维度数据集

输出：初始中心点，聚类结果

初始化：

在数据空间边缘生成T个观测点$\{P_1,P_2,\cdots,P_T\}$；

为观测点设置最大分量数$M_{\max}$；

观测阶段：

for $t:=1$ to T do

计算观测点P_t到所有数据点的距离集合X_{dis}；

计算最佳混合高斯模型，并求得I个中心点；

计算原点到观测点和数据点到观测点的夹角集合X_θ；

计算最佳混合高斯模型，并求得J个中心点；

for end

合并候选中心点阶段：

将所有的中心点存放于一个集合中，得到Q个中心点；

计算观测到的Q个候选中心点两两之间的距离，生成拉普拉斯矩阵$\boldsymbol{L}$；

计算该矩阵的特征值，得到其中1的个数U；

将U作为参数，对所有中心点作K-means聚类得到合并后的中心点集合；

根据到这些中心点的距离将所有数据点分类至子集作为聚类结果；

4 实验结果与分析

为了测试算法的有效性、合理性，以及可行性，本文设计了多个人工合成的数据集用于测试，也将该算法用于真实数据集上验证效果。由于这些数据集的真实聚类数量和各个数据的标签都是已知的，可以用纯度来评估聚类效果。

$$P=\frac{1}{N}\sum_{i=1}^{K}\max_j|C_i\cap Y_j| \tag{13}$$

式中：N为数据数量；K为算法得出的聚类数量；C_i为聚类i中的所有数据；Y_j为真实分类j的所有数据。纯度的含义为将每个聚类与真实分类中最吻合的分类计算交集大小，纯度的范围在0~1之间。

本文所提算法称为I-niceMOEn算法，由Python实现，运行IDE为Pycharm 2021.2.1专业版。谱聚类构图由Matlab实现，版本为R2018b。运行计算机CPU为Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz，运行内存为8 GB，操作系统为Windows 10专业版。

I-niceMOEn算法与Elbow算法、Silhouette算法、I-niceSO算法、I-niceMO算法进行比较，其中I-niceSO使用1个观测点，I-niceMO使用5个观测点，I-niceMOEn算法每次运行默认使用5个观测点。在I-niceMO中合并阈值选取所有候选中心点相互距离值的最小10%的平均值。

表1和表2是人工合成数据集和真实数据集的描述。合成数据集使用多维正态分布生成，数据范围在10~500之间。

表1 仿真数据集
Table 1 Description of synthetic datasets

数据集	数据量	类别数	维度
S1	1 000	10	10
S2	2 000	20	10
S3	3 000	30	10
S4	4 000	40	10
S5	1 000	10	20
S6	2 000	20	20
S7	3 000	30	20
S8	4 000	40	20

表2 真实数据集
Table 2 Description of real-world datasets

数据集	数据量	类别数	维度
Wine	178	3	13
Iris	150	3	4
Breast	569	2	30

4.1 验证可行性

使用I-niceMO算法和I-niceMOEn算法在2个数据集上运行，观察随着观测点数量逐渐增大时，算法对簇的数量的估计结果。

首先针对数据集S4和S8比较I-niceMO和I-niceMOEn不同观测点数量对运行的影响。S4和S8都有40个簇，每个簇有100个数据点，共计4 000个数据点。其中S4的特征数为10，S8的特征数为20，图6为2个数据集的平面图。

图7是2个不同算法在2个数据集上运行的比较，可以看到在S4中I-niceMOEn算法至少使用5个观测点可以得到正确的簇数；而I-niceMO无法得到正确的簇数；在S8中I-niceMOEn算法至少使用4个观测点可以得到正确的簇数，而I-niceMO需要13个。说明I-niceMOEn算法具有更强的可行性。当观测点数量过少时，候选中心点不足会导致I-niceMOEn算法中鲁棒图构造的构图方法不能良好运行，所以通常选取5个观测点对数据观测。

DOI: 10.16182/j.issn1004731x.joss.21-1333

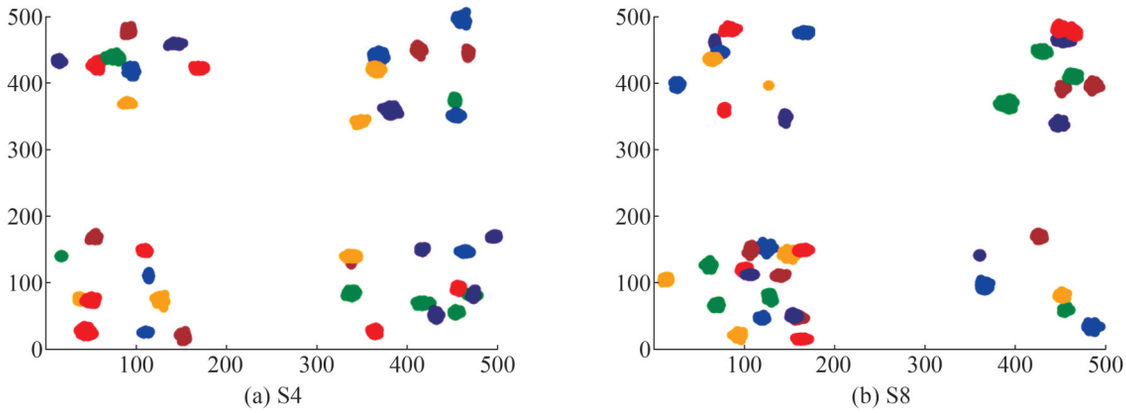


图6　仿真数据集S4和S8的分布图
Fig. 6　Description of synthetic datasets of S4 and S8

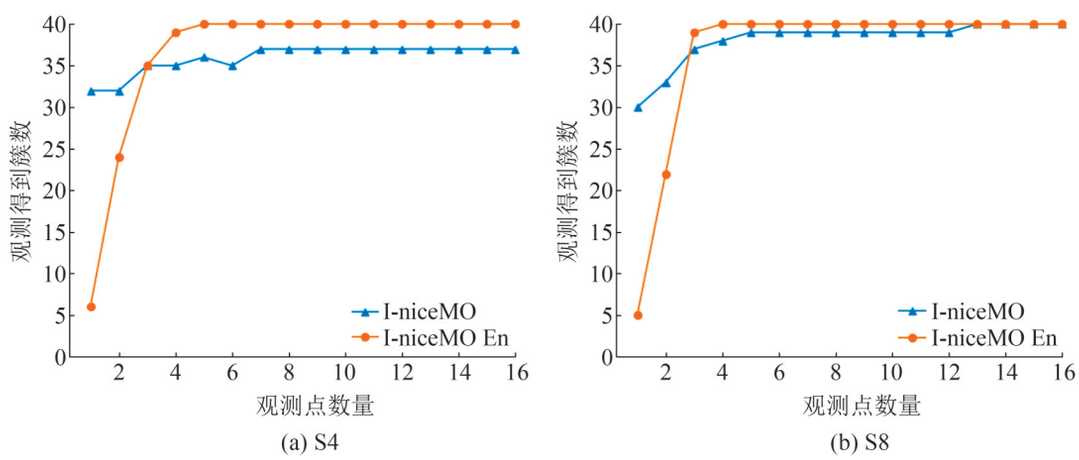


图7　不同观测点数量对两种算法的影响
Fig. 7　Impacts of different numbers of observation points on performances of two algorithms

4.2 验证合理性

使用I-niceMO算法和I-niceMOEn算法在含有混合数据的数据集上进行实验，检验通过夹角观测是否能达到满意的效果。

设计1个人工合成的二维数据集M1，该数据中有16个簇，每个簇有100个数据点，总共1 600个数据点，如图8所示。与之前的数据集不同的是，该数据集中不同的簇之间存在混合的数据，程序很可能将2个簇观测为同1个簇。

图9是2个算法的运行结果，I-niceMO算法仅观测了14个簇，未被观测的2个簇在于混合的数据中，而I-niceMOEn算法正确地观测了16个簇。可以明显看出I-niceMOEn算法可以观测出混杂的簇，证明I-niceMOEn算法具有合理性。

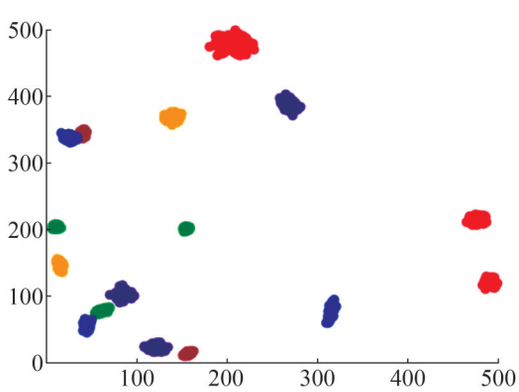
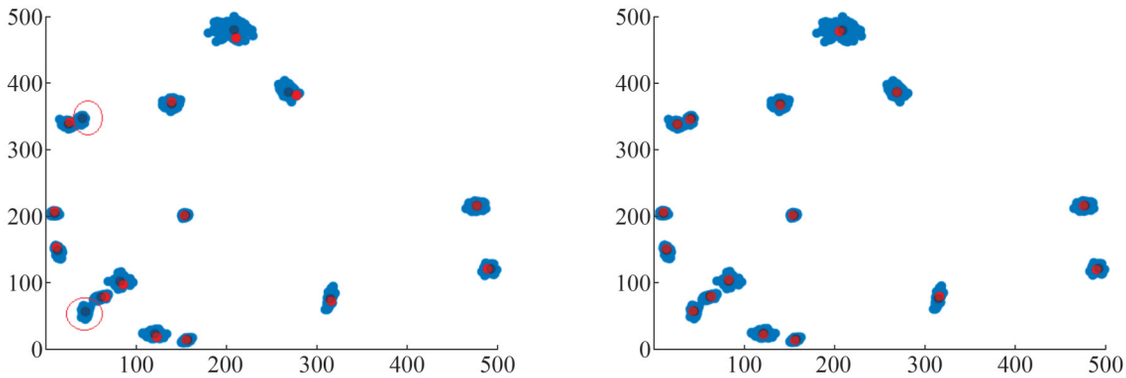


图8　仿真数据集M1
Fig. 8　Data set M1



(a) I-niceMO算法 (b) I-niceMOEn算法

图9 2种算法在仿真数据集M1上的运行结果

Fig. 9 Results of 2 algorithms on dataset M1

4.3 验证有效性

表3~4是各算法在合成数据集和真实数据集上估计的簇的数量。表5~6是不同算法所得到的纯度对比。表3中，I-niceMO算法在8个人工合成数据集中正确判断了其中3个聚类数量，而I-niceMOEn算法正确判断了所有的合成数据集，其他算法的效果均不理想。表4中，I-niceMOEn算法正确判断了3个真实数据集的聚类数量。通过结果可以发现，在人工合成数据集中，I-niceMO和I-niceMOEn算法有相接近的结果。但在真实数据集中，本文提出的I-niceMOEn算法相比I-niceMO算法有明显的超越。表5中，I-niceMOEn算法在人工合成数据集上得到的纯度(见式(13))普遍达到了100%，在2个数据集上得到的纯度低于I-niceMO算法所得到的纯度，是因为I-niceMOEn得到的中心点是由K-means算法计算簇内数据平均值得到的，而I-niceMO算法是在数据集中选择数据密度峰值点得到的，因此I-niceMOEn在选择中心点时会存在微小误差。另外，在部分数据集上I-niceMO算法得到的纯度达到100%是因为估计簇数过多使得一个簇被认为是多个簇，从而提高了纯度。在表6中，只有Iris数据集上I-niceMO算法的纯度超过了I-niceMOEn算法得到的纯度，也是因为估计簇数过多导致纯度提升。另外，I-niceMO仍需人为地猜测并调整候选点合并阈值的比例，而I-niceMOEn是完全无需这一操作的。I-niceMO这一缺陷在真实数据集中表现得更加突出，因为这种猜测在真实数据集中更加困难了。

表3 仿真数据集上簇数估计的对比

Table 3 Comparison of cluster number on synthetic datasets

数据集	类别数	Elbow	Silhouette	I-niceSO	I-niceMO	I-niceMOEn
S1	10	13	**10**	7	9	**10**
S2	20	28	19	15	**20**	**20**
S3	30	36	33	24	25	**30**
S4	40	56	46	34	37	**40**
S5	10	12	9	9	**10**	**10**
S6	20	25	18	17	18	**20**
S7	30	44	27	25	**30**	**30**
S8	40	56	41	24	39	**40**

DOI: 10.16182/j.issn1004731x.joss.21-1333

表4 真实数据集上簇数估计的对比
Table 4 Comparison of cluster number on real-world datasets

数据集	类别数	Elbow	Silhouette	I-niceSO	I-niceMO	I-niceMOEn
Wine	3	7	2	2	5	**3**
Iris	3	6	2	2	4	**3**
Breast	2	6	**2**	4	6	**2**

表5 仿真数据集上纯度的对比
Table 5 Comparison of purity on synthetic datasets

数据集	类别数	I-niceSO/%	I-niceMO/%	I-niceMOEn/%
S1	10	72.9	87.1	100
S2	20	79.1	100	98.6
S3	30	82.6	100	99.3
S4	40	82.4	92.8	98.5
S5	10	91.0	100	100
S6	20	83.8	73.5	100
S7	30	76.3	100	100
S8	40	49.6	97.6	100

表6 真实数据集上纯度的对比
Table 6 Comparison of purity on real-world datasets

数据集	类别数	I-niceSO/%	I-niceMO/%	I-niceMOEn/%
Wine	3	68.5	85.0	96.0
Iris	3	66.6	95.0	88.0
Breast	2	87.7	85.9	89.5

5 结论

为了解决经典I-niceMO算法在实际应用中的不足，本文基于夹角几何策略对I-niceMO算法进行了改进，提出了I-niceMOEn聚类算法用以更加准确地确定聚类中心和类中心个数。通过一系列实验对I-niceMOEn算法的可行性、合理性和有效性进行了验证，并基于簇数和纯度2个评价指标将提出的I-niceMOEn算法与Elbow算法、Silhouette算法、I-niceSO算法、I-niceMO算法进行了对比。实验结果表明本文所提出的I-niceMOEn算法具有收敛性，同时能够获得比对比算法更好的聚类效果。

未来我们将在以下3个方面展开研究：①进一步探索不同构图方法对聚类中心合并的影响；②在分布式环境下实现I-niceMOEn算法；③实现用于处理高维数据聚类问题的子空间I-niceMOEn算法。

参考文献：

[1] Havens T C, Bezdek J C, Palaniswami M. Scalable Single Linkage Hierarchical Clustering for Big Data[C]// 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing. Melbourne, VIC, Australia: IEEE, 2013: 396-401.

[2] 金建国. 聚类方法综述[J]. 计算机科学, 2014, 41(增2): 288-293.
Jin Jianguo. Review of Clustering Method[J]. Computer Science, 2014, 41(S2): 288-293.

[3] Ma Z, Lai Y, Xie J, et al. Dirichlet Process Mixture of Generalized Inverted Dirichlet Distributions for Positive Vector Data with Extended Variational Inference[J]. IEEE Transactions on Neural Networks and Learning Systems(S1045-9227), 2021, 99: 1-14.

[4] Masud M A, Huang J Z, Zhong M, et al. Generate Pairwise Constraints from Unlabeled Data for Semi-Supervised Clustering[J]. Data & Knowledge Engineering(S0169-023X), 2019, 123: 101715.

[5] Meyer C, Race S, Valakuzhy K. Determining the Number of Clusters Via Iterative Consensus Clustering[C]//2013

SIAM International Conference on Data Mining. Texas, USA: Society for Industrial and Applied Mathematics, 2013: 94-102.

[6] De Amorim R C, Hennig C. Recovering the Number of Clusters in Data Sets with Noise Features Using Feature Rescaling Factors[J]. Information Sciences(S0020-0255), 2015, 324: 126-145.

[7] Thorndike R L. Who Belongs in the Family? [J]. Psychometrika(S0033-3123), 1953, 18(4): 267-276.

[8] Rousseeuw P J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. Journal of Computational and Applied Mathematics (S0377-0427), 1987, 20: 53-65.

[9] Sinaga K P, Yang M S. Unsupervised K-means Clustering Algorithm[J]. IEEE Access(S2169-3536), 2020, 8: 80716-80727.

[10] 贾瑞玉, 李玉功. 类簇数目和初始中心点自确定的K-means算法[J]. 计算机工程与应用, 2018, 54(7): 152-158.
Jia Ruiyu, Li Yugong. K-means Algorithm of Clustering Number and Centers Self-Determination[J]. Computer Engineering and Applications, 2018, 54(7): 152-158.

[11] Masud M A, Huang J Z, Wei C, et al. I-nice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres[J]. Information Sciences (S0020-0255), 2018, 466: 129-151.

[12] He Y, Wu Y, Qin H, et al. Improved I-nice Clustering Algorithm Based on Density Peaks Mechanism[J]. Information Sciences(S0020-0255), 2021, 548: 177-190.

[13] Ng A Y, Jordan M I, Weiss Y. On Spectral Clustering: Analysis and an Algorithm[C]//14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA, United States: MIT Press, 2001: 849-856.

[14] Zhu X, Zhu Y, Zheng W. Spectral Rotation for Deep One-Step Clustering[J]. Pattern Recognition(S0031-3203), 2020, 105: 107175

[15] Wen G. Robust Self-Tuning Spectral Clustering[J]. Neurocomputing(S0925-2312), 2020, 391: 243-248.

[16] Druskin Vladimir, Mamonov Alexander V, Zaslavsky Mikhail. Distance Preserving Model Order Reduction of Graph-Laplacians and Cluster Analysis[J]. Journal of Scientific Computing(S0885-7474), 2022, 90(1): 1-30.

[17] Jang J, Hitchcock D B. Model-Based Cluster Analysis of Democracies[J]. Journal of Data Science(S1683-8602), 2021, 10(2): 297-319.

[18] Hastie T, Tibshirani R, Walther G. Estimating the Number of Data Clusters via the Gap Statistic[J]. Journal of the Royal Statistical Society Series B-Statistical Methodology(S1369-7412), 2001, 63: 411-423.

[19] 卢鹏丽, 才彦姣. 一种自动确定特征向量与类别数目的谱聚类算法[J]. 兰州理工大学学报, 2018, 44(2): 90-94.
Lu Pengli, Cai Yanjiao. A Spectral Clustering Algorithm for Automatic Determination of Eigenvectors and Category Quantity[J]. Journal of Lanzhou University of Technology, 2018, 44(2): 90-94.

[20] Wang Shuliang, Li Qi, Zhao Chuanfeng, et al. Extreme Clustering-A Clustering Method via Density Extreme Points[J]. Information Sciences(S0020-0255), 2020, 542: 24-39.

[21] Campello R J G B, Kröger P, Sander J, et al. Density-Based Clustering[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery(S1942-4787), 2020, 10(2): e1343.

[22] Rodriguez A, Laio A. Clustering by Fast Search and Find of Density Peaks[J]. Science(S0036-8075), 2014, 344 (6191): 1492-1496.

[23] 陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究[J]. 自动化学报, 2015, 41(10): 1798-1813.
Chen Jinyin, He Huihao. Research on Density-Based Clustering Algorithm for Mixed Data with Determine Cluster Centers Automatically[J]. Journal of Automatica Sinica, 2015, 41(10): 1798-1813.

[24] Masud M A, Rahman M M, Bhadra S, et al. Improved K-Means Algorithm Using Density Estimation[C]//2019 International Conference on Sustainable Technologies for Industry 4.0(STI). Dhaka, Bangladesh: IEEE, 2019: 1-6.

[25] Azhar M, Huang J Z, Masud M A, et al. A Hierarchical Gamma Mixture Model-Based Method for Estimating the Number of Clusters in Complex Data[J]. Applied Soft Computing(S1568-4946), 2020, 87: 105891.

[26] Kang Z, Pan H, Hoi S C H, et al. Robust Graph Learning from Noisy Data[J]. IEEE Transactions on Cybernetics (S2168-2267), 2019, 50(5): 1833-1843.

[27] Zelnik-Manor L, Perona P. Self-tuning Spectral Clustering[C]//17th International Conference on Neural Information Processing Systems. Cambridge, MA, United States: MIT Press, 2004: 1601-1608.

DOI: 10.16182/j.issn1004731x.joss.21-1333