基于最小生成树的密度聚类算法研究

王 诚,高兴东
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:传统 DBSCAN 算法对密度分布不均匀的不平衡数据集的聚类效果并不理想,同时传统算法的聚类结果对邻域半径(Eps)以及核心点阈值(MinPts)敏感。针对以上问题,改进了传统算法,提出了一种基于最小生成树的密度聚类算法(MST-DBSCAN)。由于对象之间的距离对聚类结果影响较大,为了更好地表示对象之间的距离特性,首先使用相互可达距离(mutual reachability distance)代替传统算法中的欧氏距离,表示数据集中对象与对象之间的距离,解决因密度分布不均匀导致效果不佳的问题;为了建立对象与对象之间的联系,同时保留对象之间的距离特性,引用 Prim 算法对数据集中的所有对象构建最小生成树;其次根据指定的簇的数目及最小簇对象数数目参数对得到的最小生成树进行剪枝;根据剪枝的结果,将剪枝后的各个部分进行聚类。在公开的 UCI 数据集上的实验结果表明,提出的 MST-DBSCAN 算法与现有 DBSCAN、OPTICS、KANN-DBSCAN 算法相比,在密度分布不均匀的数据集上聚类效果有所提升并且较原有传统算法有较高的聚类准确性。

关键词:DBSCAN;相互可达距离;密度聚类;最小生成树;不平衡数据集
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2022)02-0045-06
doi:10.3969/j.issn.1673-629X.2022.02.007

Research on Density Clustering Algorithm Based on MST

WANG Cheng,GAO Xing-dong
(School of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:The clustering effect of the traditional DBSCAN algorithm on the data set with uneven density distribution is not ideal,and the clustering result of the traditional algorithm is sensitive to the neighborhood radius (Eps) and the core point threshold (MinPts). To solve the above problems,the traditional algorithm is improved,and a density clustering algorithm based on minimum spanning tree (MST-DBSCAN) is proposed. Since the distance between objects has a greater impact on the clustering results,in order to better represent the distance characteristics between objects,first use mutual reachability distance instead of the Euclidean distance in the traditional algorithm to represent the objects in the data set for the problem of poor effect caused by uneven density distribution. In order to establish the connection between the object and the object,while retaining the distance characteristics between the objects,the Prim algorithm is used to construct a minimum generation of all objects in the data set tree. Secondly,pruning the obtained minimum spanning tree according to the specified number of clusters and the minimum number of cluster objects. According to the pruning result,cluster the pruned parts. Experiments on the public UCI data set show that compared with the existing DBSCAN,OPTICS,KANN-DBSCAN algorithm,the proposed MST-DBSCAN algorithm has better clustering performance on data sets with uneven density distribution and higher clustering accuracy than the traditional algorithm.

Key words:DBSCAN;mutual reachable distance;density clustering;minimum spanning tree (MST);unbalanced data set

0 引 言

数据挖掘伴随着互联网数据产生速度的大幅提升,已经在越来越多的领域发挥作用。同时互联网中的数据大多为无标签数据,聚类算法的优势得以体现。聚类分析作为数据挖掘中至关重要的技术之一,目前已经在医疗[1-2]、交通[3]、电力[4]、品质育种[5]等领域得到了广泛应用。基于密度的聚类算法是目前应用较为广泛的聚类分析算法之一。其中典型的算法有 DBSCAN 算法[6]、OPTICS 算法[7]等。文献[6]提出的 DBSCAN 算法对于 Eps 和 MinPts 敏感,同时二者为全

收稿日期:2021-03-28 **修回日期**:2021-07-28
基金项目:江苏省自然科学基金项目(BK20141428)
作者简介:王 诚(1970-),男,副教授,硕导,研究方向为互联网大数据挖掘;高兴东(1997-),男,硕士研究生,研究方向为互联网大数据挖掘。

局参数,虽然可以发现任何形状的簇,但在运算过程中参数无法更改,所以无法正确处理密度分布不均匀的样本。文献[7]提出了 OPTICS 算法,该算法并不直接得出结果簇,而是得出一个增广的簇排序,即得到基于任何参数 Eps 和 MinPts 组合的 DBSCAN 算法的聚类结果,在结果中选择聚类效果好的参数组合,该算法并没有在本质上解决如何选择参数的问题,且过程产生了不必要的运算量;同时这种排序导致低密度点与高密度点的相邻关系在映射时被分离[8]。另外,文献[9]提出首先使用 K-Means 算法对数据集进行分析以确认输入参数再进行数据集的聚类;文献[10]提出可以利用绘制 k-dist 图的方式确认 Eps 的值进而确认 MinPts 的值;文献[11]通过分析数据集平均数等统计特性,运用多组数据之间的比较得以确定 Eps 参数;文献[12]提出 KANN-DBSCAN 算法利用 k -近邻确定 Eps 候选集合,通过 k 的数学期望确定 MinPts 参数。文献[13]提出了一种基于局部特征的层次聚类算法,对不同密度样本分别处理,解决样本密度分布不均的聚类的问题。

受上述研究启发,可以使用对于算法结果影响较小的参数降低原算法中对于参数的敏感程度。该文提出一种基于最小生成树的密度聚类算法。首先通过计算样本点之间的相互可达距离构建无向有权图并求解其最小生成树;再通过预设的最小簇对象数 n 对最小生成树进行剪枝,将最小生成树中的节点中,子孙节点数不大于最小簇对象数的节点及其子孙节点标记为噪声;最后,根据预设的簇的个数 k_c 分割最小生成树并对分割结果进行聚类,解决了传统算法对于参数敏感的问题(将该算法命名为 MST-DBSCAN)。

1 DBSCAN 算法

DBSCAN 算法是一种经典的空间密度聚类算法,该算法可以聚类任何形状的簇并能够自动确定簇的数量,不需要人为设定具体的簇数。与其他聚类算法不同,该算法通过从随机的数据对象开始向外扩散的方式在数据集中进行对象的聚类。该算法在众多聚类算法中应用较为广泛且效果较好,其中主要定义如下:

定义 1(Eps 邻域):在数据集样本中随机选择 1 个数据对象 p ,其邻域 $N_{\text{Eps}}(p)$ 定义为以 p 为核心,Eps 的值为半径的多维区域,即:

$$N_{\text{Eps}}(q)=\{q \in D \mid \text{dist}(p,q) \leqslant \text{Eps}\} \tag{1}$$

其中, D 为数据集中所有对象的集合, q 为数据集中的对象, $\text{dist}(p,q)$ 表示 p 、q 两对象之间的距离。

定义 2(核心点):对于数据集中的对象 p ,如果在对象 p 的邻域内的对象数大于等于核心点阈值 MinPts,则将对象 p 称为核心点。

定义 3(直接密度可达):数据对象 p 、q 关于 Eps 和 MinPts 直接密度可达当且仅当满足公式 2、公式 3。

$$p \in N_{\text{Eps}}(q) \tag{2}$$

$$|N_{\text{Eps}}(q)| \geqslant \text{MinPts} \tag{3}$$

其中, q 为核心点。

定义 4(密度可达):数据对象 p 、q ,如果存在对象链路 $p,p_1,\cdots,p_n,q$,对于 p_{i+1} 和 p_i 关于 Eps 和 MinPts 直接密度可达,则称 p 、q 关于 Eps 和 MinPts 密度可达。

定义 5(密度相连):类比定义 4,当对象链路上的所有对象关于 Eps 和 MinPts 密度可达,则称 p 、q 关于 Eps 和 MinPts 密度相连。

定义 6(噪声、簇):从一个核心点出发,所有与当前核心点密度可达的点的集合构成一个簇;同理,所有核心点都密度不可达的对象称为噪声。

基于以上定义,DBSCAN 算法核心思想为,将数据集中所有的对象按照密度是否可达划分为不同的簇,对于密度不可达的对象标记为噪声对象。其过程为:

(1)随机选取数据集中的一个对象,判断该对象是否为核心点。如果为核心点,则将该对象标记为核心点并作为簇的起点;如果不是核心点标记为非核心点;

(2)从(1)中选取的核心点出发,寻找所有和该对象密度可达的所有对象,加入当前核心点的簇中,并以核心点的簇中的对象为起始点继续迭代寻找更多的核心点放入核心点簇中;直至当前核心点簇中的所有核心点无法找到更多满足条件的密度可达的对象为止;

(3)重复过程(1)(2),直到所有的密度可达的对象都聚类成簇。

DBSCAN 算法使用欧氏距离表示对象和对象之间的距离,同时使用从核心点向外拓展寻找簇的方式,可以发现任意形状的簇。

2 MST-DBSCAN

传统 DBSCAN 算法存在几点弊端:第一,传统算法无法使用同一组参数将密度分布不均匀的数据集中的两种或多种密度的数据进行准确的聚类。第二,需要人为指定参数且参数对于算法结果的影响很大。在 MinPts 确定的情况下,选择的 Eps 参数越小,簇中对象和对象之间的距离越小,簇的密度越高。但如果选择了过小的 Eps,使得对象的选择过于苛刻,会导致大多数的对象被错误标记为噪声,增加了噪声的数量导致算法准确性降低,同时原本为一个簇的对象也会被拆分为多个簇,将相似的对象簇进行了过度拆分;如果选择了过大的 Eps 会导致对象的选择条件降低,很多噪

声被错误地归入簇,原本独立的各个簇也会被归为一类,降低了聚类的有效性。在 Eps 确定的情况下,选择的 MinPts 越大,同一簇中的将包含更多的对象簇的密度越高。选择了过小的 MinPts 会使得大量的对象被标记为核心点,同一个簇中从核心点出发会包含更多的噪声;过大的 MinPts 会导致核心点数量减少,使得一些边缘对象被错误舍弃。

针对以上弊端对算法进行改进,改进后的 MST-DBSCAN 算法定义如下参数:

定义 7(簇的个数 k_c):对数据集聚类后期望得到簇的个数。

定义 8(最小簇对象数 n):数据集聚类期望得到簇中对象数的最小值。

改进算法的步骤如下:

Step1:距离计算。

传统算法中常使用欧氏距离表示对象与对象之间的距离,但考虑到数据集中对象的密度分布不均匀,而欧氏距离反映的是空间内的真实距离,对于分布不均的数据集的对象之间的距离无法归一化,密度大的区域的对象之间的距离和密度小的区域的对象之间的距离没有可比性,所以使用相互可达距离,如公式 4,代替样本点之间的真实距离。经过距离替换后大密度区域的对象之间距离不受影响,小密度区域的对象之间的距离被放大,起到了归一化的作用,使得大密度区域和小密度区域的对象之间的距离可以进行比较,增加了聚类算法对散点的鲁棒性;同时使用相互可达距离代替欧氏距离可以使单链路聚类算法更加贴合地去拟合数据集中的密度分布[14]。

$$d_{\text{mreach-k}}(p,q)=\max\{\text{core}_k(p),\text{core}_k(q),d(p,q)\} \tag{4}$$

其中, k 表示距离当前对象第 k 近的对象,该参数对聚类结果影响不大,通常设置为 $N*0.02$(N 为数据集对象总数)[15-16]; $d_{\text{mreach-k}}(p,q)$ 表示对象 p 、q 之间的相互可达距离; $d(p,q)$ 表示二者之间的欧氏距离; $\text{core}_k(p)$ 表示对象 p 的核心距离,即与对象 p 与第 k 近的对象的距离,具体计算方法如下;

$$\text{core}_k(x)=d(x,N^k(x)) \tag{5}$$

其中, $N^k(x)$ 表示数据集中以 x 为中心第 k 近的对象。

Step2:构建最小生成树。

利用 Step1 中求解的相互可达距离构建距离加权无向图。构建该无向图的邻接矩阵:

$$\boldsymbol{A}_s=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \tag{6}$$

其中, d_{qp} 表示数据集中对象 p 、q 之间的相互可达距离。

使用 Prim 算法计算当前邻接矩阵 $\boldsymbol{A}_s$ 的最小生成树,建立数据集中所有对象之间的联系。该算法是图论中一种常用的求解最小生成树的算法,以权值最小的边为主导,再加权有向图中搜寻最小生成树,使得生成树的权最小,即:

$$W(T)=\sum_{pq\in A_s} d_{pq} \tag{7}$$

其中, $W(T)$ 表示树 T 的权, A_s 表示树 T 的边的集合, d_{qp} 表示有向图中点 (p,q) 之间的权。

算法步骤如下:

(1)初始化:定义图中的所有顶点集合为 V ,全部边的集合为 E ,选中的顶点的集合为 V_{new} ,选中的边的集合为 E_{new} ,将 V_{new} 及 E_{new} 初始化为 $V_{\text{new}}=\{\}$, $E_{\text{new}}=\{\}$ 。随机在 V_{new} 中放入集合 V 中的任意节点 p_0 作为算法的起始节点,即 $V_{\text{new}}=\{p_0\}$;

(2)构建树:在所有 $p\in V_{\text{new}}$, $q\in V-V_{\text{new}}$ 的边 $(q,p)\in E$ 中找到权值最小的边 $e_0=(p_0,q_0)$ 放入集合 E_{new} 中,同时将 q_0 放入 V_{new} 中;

(3)递归:重复步骤(2)中构建树的过程,直到所有的顶点都放入到 V_{new} 中为止,即 $V_{\text{new}}=V$ 。

此时 E_{new} 中存在 $N-1$ 条边及 N 个点。将 E_{new} 中的所有边,按照起点-终点-权重的方式封装,并将封装后的结果放入列表 PTList 中。

Step3:簇的聚类。

通过 Step2 后生成的最小生成树,将数据集中的所有对象连接起来。通过切断节点和节点之间的边,将最小生成树切断成为多个独立的子树。具体步骤如下:

(1)将生成的 PTList 按照距离降序排序;

(2)依次选取步骤(1)排序后距离最大的边的两个节点 i 、j ,将这两个节点之间的边断开,并计算节点 i 、j 的所有子孙节点的个数,记为 N_i 、N_j ,不妨设 $N_i\leqslant N_j$ 。如果某一个节点的子孙节点个数不小于最小簇对象数,即 $\min(N_i,N_j)\geqslant n$,则将节点 N_i 及其所有子孙节点归为聚类的一个簇,放入队列 C_uList 中;否则将上述子孙节点全部标记为噪声节点;

(3)重复步骤(2),直到有 k_c-1 个簇被放入了 C_uList 中;

(4)将没有被标记的节点组成标记为一个簇放入 C_uList 中。

3　实验及结果分析

为了对改进后算法的聚类有效性进行评估,进行以下对比实验。将该算法与 DBSCAN 算法、文献[12]

提出的 KANN-DBSCAN 算法以及 OPTICS 算法在公开数据集上进行对比。

实验环境:Intel(R) Core(TM) i7-8565U CPU @ 1.80 GHz 1.99 GHz,软件环境:JDK1.8、IDEA。

该文使用的聚类算法的评价指标选择调整兰德指数(ARI)[17]、归一化互信息(NMI)[18]、完整性指标及同质性指标[19]。

调整兰德指数(ARI)是一种衡量两个数据分布的吻合程度的指标。兰德指数(RI)对两个随机的划分上存在缺陷,使得 RI 结果不是一个接近 0 的常数。ARI 解决了该问题,其指标的取值范围为[-1,1],值越大表示其聚类结果越接近真实结果,且对于随机聚类的 ARI 都非常接近 0,其计算方法如下:

$$\mathrm{ARI}=\frac{\mathrm{RI}-E[\mathrm{RI}]}{\max(\mathrm{RI})-E[\mathrm{RI}]} \tag{8}$$

其中,RI 为兰德指数,计算方法如下:

$$\mathrm{RI}=\frac{a+b}{C_2^{n_{\mathrm{samples}}}} \tag{9}$$

其中,n_{samples} 表示数据集中的样本个数,$C_2^{n_{\mathrm{samples}}}$ 表示任意两个样本为一类共有多少种情况,是数据集中的对象可以组成的总对象对的个数,a、b 分别表示聚类正确和错误的元素个数。

归一化互信息(NMI)指标是对互信息的归一化,是信息论中的一种信息度量的方式,可以理解为一个随机事件中包含另外一个随机事件的信息量。可以通过随机时间发生的概率计算得到。

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x),p(y)}\right) \tag{10}$$

其中,$I(X;Y)$ 表示两随机事件的互信息,$p(x,y)$ 表示二者的联合分布概率,$p(x)$ 及 $p(y)$ 分别表示两事件单独发生的概率。

NMI 将互信息的取值归一化在[0,1]之间,其取值越接近 0 表示聚类结果越差,越接近 1 表示聚类结果越接近真实结果,计算公式如下:

$$\mathrm{NMI}(C,T)=\frac{\sum_{i=1}^{k(C)}\sum_{j=1}^{k(T)}n_{i,j}\log\left(\frac{n\cdot n_{i,j}}{n_i\cdot n_j}\right)}{\sqrt{\left(\sum_{i=1}^{k(C)}n_i\log\frac{n_i}{n}\right)\left(\sum_{j=1}^{k(T)}n_j\log\frac{n_j}{n}\right)}} \tag{11}$$

其中,$\mathrm{NMI}(C,T)$ 表示聚类结果 C 与真实结果 T 的归一化互信息,n 为样本总数,$k(C)$ 和 $k(T)$ 分别为聚类结果和真实结果的簇数,$\sum_{i=1}^{k(C)}\sum_{j=1}^{k(T)}n_{i,j}\log\left(\frac{n\cdot n_{i,j}}{n_i\cdot n_j}\right)$ 表示 C、T 之间的互信息。

完整性及同质性是一种基于条件熵的聚类评估方法。二者往往存在一定的负相关关系。其中,完整性(completeness)表示在真实结果中某一簇的所有成员,经过聚类后被分配到单一簇的程度;同质性(homogeneity)表示聚类结果中每一个簇包含单一类别的成员的程度。两个指标的取值范围为[0,1],越接近 0 表示聚类后的结果分布松散,对数据集的错误聚类情况越多,越接近 1 表示聚类结果越准确。计算方法如公式 12 和公式 13:

$$h=1-\frac{H(C\mid K)}{H(C)} \tag{12}$$

$$c=1-\frac{H(K\mid C)}{H(K)} \tag{13}$$

其中,h 表示同质性,c 表示完整性,h、c 的取值范围为[0,1];$H(C\mid K)$ 及 $H(K\mid C)$ 是给定簇 C 和 K 的条件熵,$H(C)$ 及 $H(K)$ 是簇的熵。

选用以上 4 个参数作为聚类结果的评价指标,可以从聚类结果与真实结果的吻合度、聚类结果的准确性及聚类程度等多方面衡量聚类算法的聚类效果。

实验分别选择 UCI 公开数据集中的 5 个常用数据集,数据集包含了不同维度、不同分类数的数据集,各个数据集中的具体参数如表 1 所示。

表 1 UCI 数据集

数据集	对象数	特征数	分类数
Iris	150	4	3
Thyroid	215	5	3
Glass	214	9	6
Wine	178	13	3
Breast Cancer	569	32	2

4 种算法在 UCI 数据集上的评价指标对比如表 2 所示。

将 5 个数据集在 DBSACAN、OPTICS、KANN-DBSCAN、MST-DBSCAN 算法中的各个性能指标绘制成折线图,如图 1 所示。

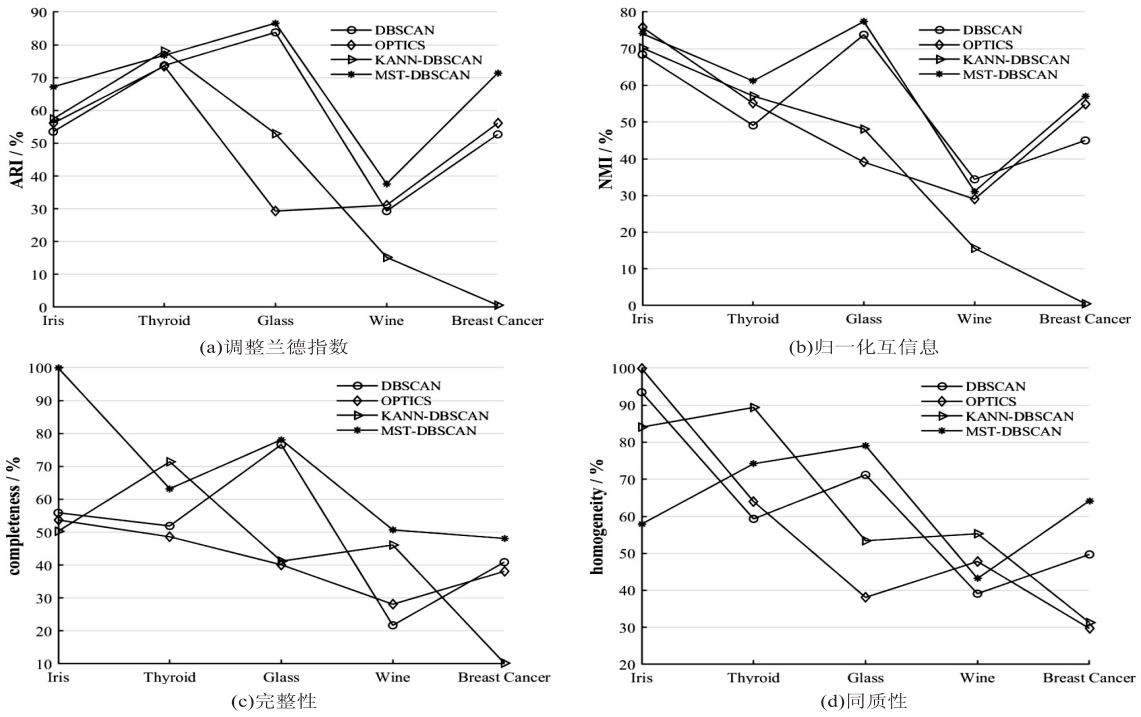
通过以上对比结果可以看出,MST-DBSCAN 算法在各个数据集中的评价指标相比于传统算法都有着不同程度的改进,相比其他改进算法,该算法在整体效果上优于其他算法。虽然在 Iris 数据集中 OPTICS 算法的 NMI 及完整性指标高于 MST-DBSCAN;在 Thyroid 数据集中 KANN-DBSCAN 算法的 ARI 及完整性指标优于 MST-DBSCAN 算法等,但从整体效果对比来看,随着数据集的特征数增加,MST-DBSCAN 算法的 ARI、同质性等聚类评价指标明显高于其他聚类算法。

在密度分布不均匀的数据集及高维度数据集中,改进的算法体现了优越性。在处理密度分布不均匀的数据集时,如 Glass 及 Wine 数据集,其中 Glass 数据集的不平衡问题最为突出,通过对比图 1 可以发现,MST-DBSCAN 算法在 4 种评价指标上的表现都优于其他

算法,其中 NMI 指标最为明显。在处理高维度数据集时,如 Breast Cancer 数据集,该数据集的数据维度为 32,从评价指标中可以看出,MST-DBSCAN 算法的表现同样优于其他算法。尤其对于 KANN-DBSCAN 算法,在 ARI 及 NMI 指标上有明显优势。

表 2　4 种算法在 UCI 数据集上的评价指标对比

数据集	算法	ARI	NMI	homogeneity	completeness
Iris	DBSCAN	0.535	0.684	0.559	0.935
	OPTICS	0.561	0.758	0.537	0.999
	KANN-DBSCAN	0.574	0.702	0.503	0.841
	MST-DBSCAN	0.672	0.741	0.999	0.579
Thyroid	DBSCAN	0.737	0.491	0.519	0.593
	OPTICS	0.735	0.551	0.486	0.640
	KANN-DBSCAN	0.781	0.571	0.714	0.894
	MST-DBSCAN	0.768	0.612	0.632	0.742
Glass	DBSCAN	0.838	0.738	0.766	0.712
	OPTICS	0.293	0.391	0.401	0.381
	KANN-DBSCAN	0.529	0.481	0.412	0.534
	MST-DBSCAN	0.866	0.774	0.781	0.791
Wine	DBSCAN	0.293	0.344	0.217	0.391
	OPTICS	0.311	0.290	0.281	0.478
	KANN-DBSCAN	0.152	0.156	0.461	0.553
	MST-DBSCAN	0.376	0.311	0.507	0.432
Breast Cancer	DBSCAN	0.527	0.450	0.409	0.497
	OPTICS	0.561	0.548	0.381	0.297
	KANN-DBSCAN	0.006	0.005	0.102	0.313
	MST-DBSCAN	0.714	0.571	0.481	0.642



(a)调整兰德指数　(b)归一化互信息　(c)完整性　(d)同质性

图 1　改进前后各指标对比结果

通过上述比较发现,该文提出的 MST-DBSCAN 算法的聚类效果最好,在密度分布不均及高维度数据集上的表现较好,DBSCAN 和 OPTICS 算法的效果一般,KANN-DBSCAN 算法在数据维度高的数据集的聚类效果最不理想。

4　结束语

该文运用最小生成树思想,对数据集中的对象构

建最小生成树,通过指定簇的个数及最小簇对象数进行剪枝,因为指定了簇的数目,使聚类结果更加接近真实结果。进一步优化了 DBSCAN 对 Eps 及 MinPts 敏感的问题,选择使用更加容易确定的参数代替较难确定的参数,省去了原算法中对于 Eps 及 MinPts 选择的过程。同时通过在对象之间使用相互可达距离代替 DBSCAN 中使用的欧氏距离在一定程度上避免了维度灾难问题,解决了 DBSCAN 算法对于密度分布不均与数据集聚类效果不理想的问题。实验结果表明,MST-DBSCAN 算法可以完成对于密度分布不均匀的数据集的聚类;同时,对高维数据集的聚类效果优于原 DBSCAN 算法。但 MST-DBSCAN 算法在低维数据集聚类效果不稳定,同时算法增加了聚类的复杂度,对于规模较大的数据集消耗时间较长。接下来的工作将继续降低算法的时间复杂度及低维数据集的聚类稳定性,将算法的聚类效果最优化。

参考文献:

[1] YIN L,LIU J,LIN X,et al. Nutritional features-based clustering analysis as a feasible approach for early identification of malnutrition in patients with cancer[J]. European Journal of Clinical Nutrition,2021,75:1291-1301.

[2] BONNEFOND A,FROGUEL P. Clustering for a better prediction of type 2 diabetes mellitus[J]. Nature Reviews Endocrinology,2021,17:193-194.

[3] 宫同伟,运迎霞. 基于因子分析和聚类分析的城市轨交站区功能识别方法[J]. 统计与决策,2020,36(5):177-180.

[4] 申建建,张楠男,程春田,等. 基于聚类分析和决策树的"一库多级"水电站日调度方法[J]. 中国电机工程学报,2019,39(3):652-663.

[5] 王 颖,田应金,蒋 伟,等. 基于热图和聚类分析的马铃薯矿质元素含量评价[J]. 分子植物育种,2019,17(19):6483-6488.

[6] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd international conference on knowledge discovery and data mining. Portland,Oregon:AAAI Press,1996:226-231.

[7] ANKERST M,BREUNIG M M,KRIEGEL H P,et al. OPTICS:ordering points to identify the clustering structure[J]. ACM SIGMOD Record,1999,28(2):49-60.

[8] 曾依灵,许洪波,白 硕. 改进的 OPTICS 算法及其在文本聚类中的应用[J]. 中文信息学报,2008,22(1):51-55.

[9] 雷小锋,谢昆青,林 帆,等. 一种基于 K-Means 局部最优性的高效聚类算法[J]. 软件学报,2008,19(7):1683-1692.

[10] 周 董,刘 鹏. VDBSCAN:变密度聚类算法[J]. 计算机工程与应用,2009,45(11):137-141.

[11] 夏鲁宁,荆继武. SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院研究生院学报,2009,26(4):530-538.

[12] 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.

[13] 王鹏宇,王国宇,贾 贞,等. 一种基于局部特征的层次聚类算法[J]. 中国海洋大学学报:自然科学版,2019,49(z1):176-184.

[14] ELDRIDGE J,BELKIN M,WANG Y. Beyond hartigan consistency:merge distortion metric for hierarchical clustering[C]//Proceedings of the 28th conference on learning theory. Cambridge MA:JMLR,2015:588-606.

[15] DU M,DING S,JIA H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J]. Knowledge-Based Systems,2016,99:135-145.

[16] CHEN M,LI L,WANG B,et al. Effectively clustering by finding density backbone based-on kNN[J]. Pattern Recognition,2016,60:486-498.

[17] VINH N X,EPPS J,BAILEY J. Information theoretic measures for clusterings comparison[C]//Proc of the international conference on machine learning. Montreal,Quebec,Canada:ACM,2010:2837-2854.

[18] NGUYEN T P Q,KUO R J. Partition-and-merge based fuzzy genetic clustering algorithm for categorical data[J]. Applied Soft Computing,2019,75:254-264.

[19] ZHANG H, GUO H, WANG X, et al. Clothescounter:a framework for star-oriented clothes mining from videos[J]. Neurocomputing,2020,377:38-48.