Journal of Computer Applications
计算机应用, 2020, 40(8): 2248 - 2254
ISSN 1001-9081
CODEN JYIIDU
2020-08-10
http://www.joca.cn

文章编号:1001-9081(2020)08-2248-07
DOI:10.11772/j.issn.1001-9081.2020010057

基于数据流的聚类趋势分析算法

樊仲欣*
(大气科学与环境气象国家级实验教学示范中心(南京信息工程大学),江苏 南京 210044)
(*通信作者电子邮箱 fan_zhong_xin@qq.com)

摘　要:聚类趋势分析算法基于抽样原理导致聚类趋势指标不稳定和片面,而且不适应数据流的批量增量特性,因而需要重复进行聚类趋势指数计算。为此,基于全体数据进行整体分析,提出一种基于最小距离连通图(MDCG)的聚类趋势分析算法MDCG-CTI。首先,利用栈的深度优先遍历法更新增量数据的最邻近路径从而降低MDCG的建立复杂度;然后,计算聚类趋势指数并确定可聚类性的判定阈值;最后,将所提算法和批量增量的具有噪声的基于密度的聚类方法(DBSCAN)相结合。在自定义数据集上的实验表明,该算法比现有算法对单簇和含大量噪点的数据的可聚类性判断更为精确;而在大数据集pendigits和avila上,所提算法比基于谱方法的聚类趋势可视化分析(SpecVAT)累计耗时降低了38%和42%,且相较SpecVAT结合批量增量DBSCAN,该算法结合批量增量DBSCAN的聚类平均准确率分别提高了6%和11%,聚类累计耗时则分别降低了7%和8%。实验结果表明该算法可以准确无参地判断聚类趋势,并明显提高增量聚类的有效性和运行效率。

关键词:聚类趋势;最小距离连通图;数据流聚类;批量增量聚类;具有噪声的基于密度的聚类方法
中图分类号:TP312　　**文献标志码**:A

Clustering tendency analysis algorithm based on data stream

FAN Zhongxin*
(*National Experimental Teaching Demonstration Center for Atmospheric Science and Environmental Meteorology (Nanjing University of Information Science and Technology), Nanjing Jiangsu* 210044, *China*)

Abstract: Focusing on the issues that clustering tendency analysis algorithms based on sampling have instability and one-sidedness in clustering tendecy index, and clustering tendency parameters need to be computed repeatedly because the algorithms do not suit the batch incremental property of data stream, an improved Clustering Tendency Index analysis algorithm based on Minimum Distance Connected Graph (MDCG) was proposed, namely MDCG-CTI, which performs overall analysis on all data. First, MDCG was built with complexity optimization by using stack depth-first traversal to update the nearest path of incremental data; then clustering tendency index was computed to determine the judgment threshold of clustering; finally, the proposed algorithm was integrated with batch incremental Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Experimental results on self-built datasets show that the proposed algorithm has higher accuracy of clusterable determination than existing algorithms for single cluster and data with a large number of noises. And on large datasets pendigits and avila, the proposed algorithm has the time consumption reduced by 38% and 42% compared to Spectral Visual Assessment of cluster Tendency (SpecVAT); meanwhile, the proposed algorithm combined with batch incremental DBSCAN has average accuracy of clustering increased by 6% and 11% and time consumption of clustering reduced by 7% and 8% compared to SpecVAT combined with batch incremental DBSCAN. It can be seen that the proposed algorithm not only determines clustering tendency nonparametrically and accurately, but also improves effectiveness and operational efficiency of incremental clustering.

Key words: clustering tendency; Minimum Distance Connected Graph (MDCG); data stream clustering; batch incremental clustering; Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

0　引言

聚类是数据挖掘的一个重要研究领域,它通过对无类别标签数据的属性(如距离、密度、分布等)进行无监督学习,从而将数据划分成多个簇,并使得簇内的数据在属性上具有高相似性,而簇间的数据则在属性上相似性低。聚类趋势分析是聚类的前期预处理步骤,其意义在于确定数据是否具备可聚类性,及其可聚类性的强弱,以便于在此基础上判断是否有继续进行聚类操作的必要性以及聚类结果的可信程度,这对于必然会得到一个聚类结果的各种聚类算法来说不仅能节省计算成本(尤其是大数据集的计算成本),而且也可以作为聚类的一个先验指标。

目前聚类趋势分析研究主要分三个方面[1]:1)基于统计检验的方法,以Hopkins[2]、Cow-Lewis[3]、T-平方(T-squared)[4-5]和Elberhardt[6]统计量为主要代表;2)基于图论的方法,以IC聚类趋势指数(Clustering Tendency Index)[7]为代表;3)基于可

收稿日期:2020-01-19;**修回日期**:2020-03-17;**录用日期**:2020-03-24。　**基金项目**:国家重点研发计划项目(2018YFC1505804)。
作者简介:樊仲欣(1981—),男,江苏南京人,工程师,硕士,主要研究方向:数据挖掘、神经网络。

视化分析的方法,以 VAT(Visual Assessment of cluster Tendency)[8]、sVAT(scalable VAT)[9]、bigVAT(VAT for large datasets)[10]、SpecVAT(Spectral VAT)[11]、cSpecVAT(cosine metric SpecVAT)和 GMMMVS-VAT(Gaussian Mixture Model and Multi-Viewpoint based Similarity VAT)[12]为代表。这些方法中统计检验依赖于假设检验,应用空间抽样原理基于不同的最近邻模式建立 Hopkins 等各种统计量。由于该方法实现简单且时间空间复杂度低,所以实践应用较多,其中 Hopkins 统计量应用在电力负荷预测[13]、UCI 机器学习库的混合型数据聚类[14],T-平方统计量应用在电铁 W 牵引负载模型建模[15]等方面都已有成功先例。图论方法将多维空间的数据看作是一个无向图,如果图中不同簇的顶点都是相邻的,那么称其为 K 部完全图,算法利用地图着色原理提出了 IC 指数,该指数表示建立 K 部完全图需要增加的边数。可视化分析方法通过将数据的相异度信息用灰度图像的形式进行展示以判断是否具有潜在聚类结构。最初的 VAT 方法因为高时间复杂度(数据量的平方)而不适用于大样本或关联数据集,因此改进的 sVAT 和 bigVAT 算法用采样技术以及二维剖面图代替相异度图的办法在一定程度上解决了上述问题,而后来 SpecVAT、cSpecVAT、GMMMVS-VAT 方法的提出又进一步提高了可视化方法的聚类效果判断精准度,并将其适用的数据集类型扩展到了图像和视频上。

然而,上述各种聚类趋势分析算法均存在两个问题有待解决:

1)抽样导致的聚类趋势指标不稳定以及片面性。Hopkins 统计量需要随机均匀抽样,因此指标稳定性差,尤其是在高维度稀疏数据集的情况下;Cow-Lewis 虽然在高维空间检测有优势,但有时的结果片面不稳定;T-平方在实践中效果相对较好,但是抽样窗口在高维度空间中的设置难度很大,容易导致指标不稳定;而 Elberhardt 也和 T-平方存在同样的问题[1]。

2)现有的聚类趋势算法是针对离线数据设计的,因此不适用于数据流这种在线动态增量数据的聚类。近年来,基于实时数据的数据挖掘越来越受重视,因此涌现出了不少基于数据流的聚类方法,如在具有噪声的基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)[16]的基础上出现了增量 DBSCAN[17]和批量增量 DBSCAN[18]的算法,以及其改进算法 R-DBSCAN(Rough-DBSCAN)[19]、DG2CEP(Density-Grid Clustering using Complex Event Processing)[20]等用于流数据识别[19,21],但是这些算法都忽视了聚类趋势分析的重要性,所以只能实现定量数据(单个数据或固定量多个数据)的增量聚类。而现有的聚类趋势分析算法中,抽样技术应用于递增数据序列时需要每增一项数据就重复进行一次抽样,这样无法充分利用数据的动态特性。图论方法构造 K 部完全图需要数据完备,因此也不适合应用在动态增量数据序列上。可视化分析方法早期的 VAT、sVAT 和 bigVAT 效率低、耗时长,且只适用于相异度矩阵是方阵的情况,后来提出的 SpecVAT、cSpecVAT、GMMMVS-VAT 方法虽然提高了适用范围和准确性,但是由于使用到了谱方法(需要数据一次性到齐)并且需要确定特征向量的选取个数,因此算法复杂且耗时更久,也不适用于增量性的数据流。

因此,本文提出了一种可以适应数据流的基于全体数据最小距离连通图(Minimum Distance Connected Graph, MDCG)的聚类趋势算法——MDCG-CTI(Clustering Tendency Index analysis algorithm based on MDCG)来解决上述问题。

1　聚类趋势算法原理

本文算法的实现原理如下:

1)针对全体数据生成 MDCG,同时使图的生成可以适应数据流的递增数据序列进行动态调整,从而大幅度降低时间复杂度;

2)以 MDCG 为基础,用肘阈值分割出簇间和簇内的边,再综合变异系数、均值等统计量,计算增量数据序列的聚类趋势指数;

3)根据聚类趋势指数动态决定批量递增的数据量,确保后续聚类的有效性及运行效率。

1.1　建立递增数据序列的 MDCG

定义 1　最小距离连通图(MDCG)。由 n 个顶点的唯一标识集合 ID 和 $n-1$ 条连接顶点的边集 $\boldsymbol{E}$ 组成的图中,如果图无向无环,且这 $n-1$ 条边是 n 个顶点按照最邻近距离连接而成的,其中任意两个顶点间都有一条唯一的最邻近路径,则称该图为最小距离连通图 MDCG$[ID,\boldsymbol{E}]$,如图 1。

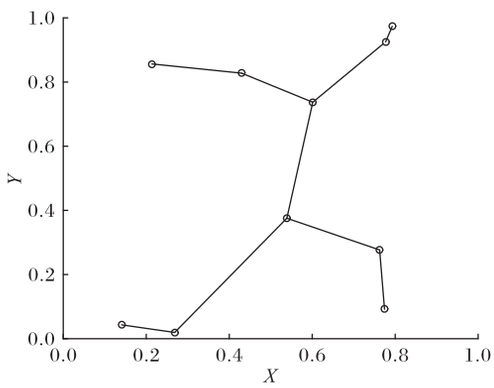


图 1　最小距离连通图示意图

Fig. 1　Schematic diagram of MDCG

基于增量数据动态生成最小距离连通图 MDCG$[ID,\boldsymbol{E}]$:

步骤 1　初始顶点数据 $\boldsymbol{X}$ 为空,MDCG 的唯一标识集合 ID 和最邻近边集 $\boldsymbol{E}$ 也为空,新增数据 $x_1=\{x_{11},x_{12},\cdots,x_{1m},id_1\}$ 后,$\boldsymbol{X}=[x_{11},x_{12},\cdots,x_{1m},id_1]$,$ID=\{id_1\}$,$\boldsymbol{E}=\varnothing$;新增 $x_2=\{x_{21},x_{22},\cdots,x_{2m},id_2\}$ 后,$\boldsymbol{X}=\begin{bmatrix}x_{11},x_{12},\cdots,x_{1m},id_1\\x_{21},x_{22},\cdots,x_{2m},id_2\end{bmatrix}$,$ID=\{id_1,id_2\}$,$\boldsymbol{E}=[id_2,id_1,d_{21}]$。其中:$m$ 表示顶点数据的维度;d_{21} 表示第 1 个和第 2 个顶点间的欧氏距离(简称顶点距离);id_1 和 id_2 表示第 1 个和第 2 个顶点数据的唯一标识(简称顶点)。按 $new=\{3,4,\cdots,n\}$ 的顺序依次执行步骤 2~6。

步骤 2　计算新增数据 $x_{new}=\{x_{new1},x_{new2},\cdots,x_{newm},id_{new}\}$($new=\{3,4,\cdots,n\}$)和 $\boldsymbol{X}$ 的所有欧氏距离集合 $D=\{d_{new1},d_{new2},\cdots,d_{new(new-1)}\}$,并同时记录下 D 的首个最小顶点距离 d_{newi} 以及其对应的顶点 id_i 作为新增边 $[id_{new},id_i,d_{newi}]$ 加入最邻近边集 $\boldsymbol{E}$ 成为新增行。

步骤 3　令栈数据结构 ST 为空,将边 $[id_{new},id_i,d_{newi}]$ 作为一项数据入栈,再设置检索排除顶点数据唯一标识集合为 $EXI=\{id_{new}\}$。

步骤 4　栈 ST 数据出栈,出栈的数据项表示为 $[id_j,id_k,d_{jk}]$,$EXI=EXI\cup\{id_k\}$,然后在最邻近边集 $\boldsymbol{E}$ 中检索出 1 条一个顶点为 id_k 且另一顶点 id_{ka} 不在集合 EXI 中的边 e_k。如果 e_k 存在,则将刚出栈的数据 $[id_j,id_k,d_{jk}]$ 入栈,以及将 e_k 中的顶点顺序调整为 $\{id_k,id_{ka}\}$ 后再将 $[id_k,id_{ka},d_{kka}]$ 入栈,调用

步骤5并返回。如果e_k不存在且栈ST也为空(此时出栈的数据项$[id_j, id_k, d_{jk}]=[id_{new}, id_i, d_{newi}]$且$id_k$无法拓展出$EXI$以外的顶点),则在最邻近边集$\boldsymbol{E}$中检索出1条一个顶点为$id_j$且另一顶点$id_{ja}$不在集合$EXI$中的边$e_j$。如果$e_j$存在,则将$e_j$中的顶点顺序调整为$\{id_j, id_{ja}\}$后再将$[id_j, id_{ja}, d_{jja}]$入栈,调用步骤5并返回。重复本步骤直到$EXI = ID \cup \{id_{new}\}$,或者距离集合$D$中小于最邻近边集$\boldsymbol{E}$的顶点距离最大值的所有距离$D2 = \{d_{newp}, d_{newq}, \cdots, d_{newq}\}$($D2 \subseteq D$)对应的$ID2 = \{id_p, id_q, \cdots, id_q\}$($ID2 \subseteq ID$)全部属于$EXI$,则进入步骤6。

步骤5 遍历栈ST,获取栈的各条边数据中顶点距离的首个最大值d_{m1m2}所在边$\boldsymbol{e}_{m1m2}=[id_{m1}, id_{m2}, d_{m1m2}]$,如果$d_{m1m2}$大于距离集合$D$中$id_{new}$到$id_{m2}$($id_{m2} \in ID$)的距离$d_{newm2}$($d_{newm2} \in D$),则从最邻近边集$\boldsymbol{E}$中删除边$\boldsymbol{e}_{m1m2}$并新增边$\boldsymbol{e}_{newm2}=[id_{new}, id_{m2}, d_{newm2}]$,同时清空栈$ST$和$EXI$,并将新增边$\boldsymbol{e}_{newm2}$入栈再设置$EXI = \{id_{new}\}$。

步骤6 将x_{new}加入到$\boldsymbol{X}$中,将id_{new}加入到ID中,生成当前的顶点数据$\boldsymbol{X}$以及最小距离连通图MDCG$[ID, \boldsymbol{E}]$。

步骤1~6计算新增顶点的MDCG流程在于用栈数据结构记录从新增顶点深度优先遍历到当前任意一个已有顶点id_{exist}的最邻近路径(该路径从最邻近边集$\boldsymbol{E}$中选取,因此栈ST的每条边都包含在最邻近边集$\boldsymbol{E}$的各边中,但栈中各条边顶点的先后顺序会根据路径拓展顺序而有所调整),然后如果该最邻近路径的距离最大值大于了新增顶点到id_{exist}的直接距离,则表示有更短的最邻近路径可以生成,因此需要更新$\boldsymbol{E}$,并清空栈ST和EXI后重新开始遍历。由于使用栈的遍历过程中新增顶点到id_{exist}的直接距离在总体趋势上是不断增大的,因此当其大于$\boldsymbol{E}$的最大顶点距离时便不会有更短的最邻近路径可以生成,此时$\boldsymbol{E}$的更新已完成可以提前结束遍历。

1.2 基于MDCG计算聚类趋势指数

定义2 肘阈值。在最小距离连通图MDCG$[ID, \boldsymbol{E}]$的最邻近边集$\boldsymbol{E}$中,计算得到顶点距离差分最大值的连续相邻两个距离值的均值即为肘阈值。

肘阈值反映的是数据序列变化的最大差异阈值,应用在MDCG上即可以分割出簇间距离(大于肘阈值的距离)和簇内距离(小于肘阈值的距离),对于一个具有明显聚类趋势的MDCG来说,应该具有以下三个特性:

1)簇内距离数量远大于簇间距离数量。

2)簇内距离的均值远小于簇间距离均值。

3)簇内距离的变异系数远小于整体距离的变异系数。

上述三个特性中:第1点反映的是簇内可聚类数据的数据量,越大越好;第2点反映的是簇内和簇间距离的差异性,越大越好;第3点反映的是簇内和数据总体在离散程度上的差异性,越大越好。因此得出聚类趋势指数R_{MDCGCTI}如下:

$$num\_i = \frac{num\_s - num\_b}{num\_s + num\_b} \tag{1}$$

$$mean\_i = 2 \times \left(\frac{mean\_b - mean\_s}{mean\_b}\right) - 1 \tag{2}$$

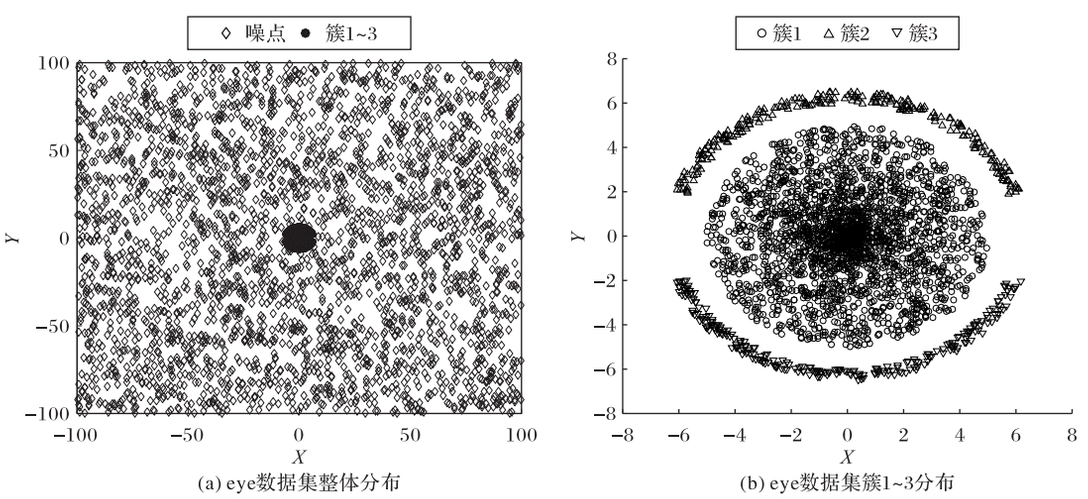
$$cv\_i = \frac{cv\_a - cv\_s}{cv\_a + cv\_s} \tag{3}$$

$$R_{\mathrm{MDCGCTI}} = \begin{cases} -1, & num\_s = num\_b = 0 \\ \dfrac{num\_i + mean\_i + 1}{3}, & mean\_s = 0 \\ \dfrac{num\_i + mean\_i + cv\_i}{3}, & \text{其他} \end{cases} \tag{4}$$

其中:num_s表示簇内距离数量,num_b表示簇间距离数量;$mean_s$表示簇内距离均值,$mean_b$表示簇间距离均值;cv_s表示簇内距离变异系数,cv_a表示全部距离变异系数。式(4)对式(1)~(3)求均值,R_{MDCGCTI}的数据范围是$[-1, 1)$。

为确定R_{MDCGCTI}的聚类趋势阈值,使用eye数据集进行逐项递增数据的聚类趋势指数分析,该数据集有6 000个二维数据,其整体分布如图2(a),其中簇1至簇3(位于图2(a)的中心部位)的分布如图2(b),具有多簇和大量噪点,且各簇的体积不均,形状也不同。

为方便分析,图2的eye数据集数据的递增顺序为簇1(2 500个数据)、簇2(250个数据)、簇3(250个数据)、噪点(3 000个数据)。其num_i、$mean\_i$、cv_i和R_{MDCGCTI}随数据量递增的变化情况如图3所示。



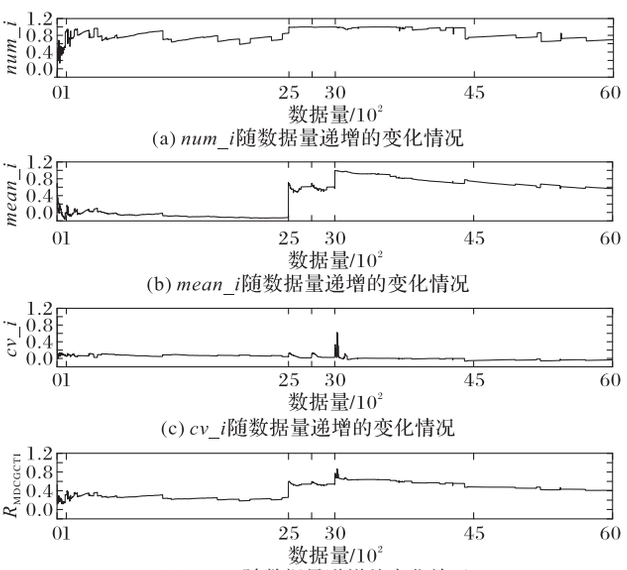
(a) eye数据集整体分布 (b) eye数据集簇1~3分布

图2 eye数据集的分布

Fig. 2 Distribution of eye dataset

通过分析图3可以发现,当数据在簇1中随机递增时,一开始的统计量波动都很大,这是因为在数据量极少的情况下,数据聚类趋势很不明显,因此各统计量都会偏向于极值。但当数据量超过100以后各统计量便基本稳定下来,其中num_i在0.7至0.9的范围内波动,$mean_i$和cv_i在0左右浮动,因此R_{MDCGCTI}取值在0.2至0.4,此时由于只有单簇簇1,聚类趋势是不显著的,所以R_{MDCGCTI}统计量较小。但当数据量超过2 500,递增到簇2和簇3时,由于多簇具有显著的聚类趋势,

因此*mean_i*统计量上升明显,*num_i*略有上升并达到最大值,*cv_i*则有小幅度波动,导致了R_{MDCGCTI}统计量大幅度上升至0.6左右,此时具有了显著的聚类趋势。此外值得注意的是在2 500个和2 750个数据(递增到了新簇簇2和簇3)的位置时,*mean_i*和*cv_i*均有一个抬升回落的峰值现象,从而使得R_{MDCGCTI}也出现同样的情况,这是因为在新簇数据量极少时会被判定为噪点,而噪点具有数据非常分散的特点,因此导致了肘阈值分割后簇间距离(大于肘阈值的距离)数量的大幅度增加,而后面随着新簇数据量增加簇间距离数量会快速减少并达到稳定,因此*mean_i*和*cv_i*统计量才出现了峰值。最后在数据量超过3 000刚刚递增到噪点时,*mean_i*和*cv_i*又出现了峰值,这和数据量刚刚递增到簇2时情况类似,此时的R_{MDCGCTI}值亦有所增加,在0.7至0.8(由于多簇含少量噪点依然是具有显著聚类趋势的,因此R_{MDCGCTI}值的变化是合理的);但是此后随着噪点不断增多,*num_i*和*mean_i*便逐步下降(*mean_i*稳步下降,*num_i*则在噪点量达到约1 500时会突然骤降),*cv_i*则基本保持不变,因此R_{MDCGCTI}也逐步下降并在接近4 500个数据前降至了0.5以下,考虑到此时噪点的数量已近1 500个,约为簇数据量的50%,而大量的噪点(多噪点)导致了簇的聚类趋势不再明显,所以R_{MDCGCTI}值低于0.5正体现了这种非显著性的聚类趋势。



(a) *num_i*随数据量递增的变化情况

(b) *mean_i*随数据量递增的变化情况

(c) *cv_i*随数据量递增的变化情况

(d) R_{MDCGCTI}随数据量递增的变化情况

图3　eye数据集随数据量递增的统计量变化情况

Fig. 3　Statistics of eye dataset varying with incremental data

综上所述,R_{MDCGCTI}统计量在单簇和多簇的情况下主要受*mean_i*和*cv_i*的影响而呈现单簇取值低于0.5,多簇取值高于0.5的现象,其中*mean_i*决定了R_{MDCGCTI}的总体变化趋势,*cv_i*则提升了R_{MDCGCTI}对新簇出现的敏感性。同时R_{MDCGCTI}在噪点增多的情况下主要受*num_i*和*mean_i*的影响而呈现取值递减的现象,并最终在大量噪点使得聚类趋势不显著时减至0.5以下,其中*mean_i*依然决定了R_{MDCGCTI}的总体变化趋势,*num_i*则进一步提升了R_{MDCGCTI}对噪点减弱聚类趋势的敏感性。因此可以认为:$R_{\mathrm{MDCGCTI}} \leqslant 0$时没有聚类趋势(不可聚类);$R_{\mathrm{MDCGCTI}} \in (0, 0.5]$时具有不显著的聚类趋势(非显著可聚类);$R_{\mathrm{MDCGCTI}} > 0.5$时则具有显著聚类趋势(显著可聚类)。

结合聚类趋势分析的批量增量数据聚类趋势分析的目的是为聚类算法服务的,因此为了说明本文的MDCG-CTI算法如何与具体的DBSCAN聚类算法相结合,给出算法流程如图4所示。图4的步骤一为MDCG-CTI算法,预处理部分要读取增量DBSCAN的区域半径*Eps*和邻域对象数量*MinPts*阈值,以及不断地在线读取新增数据(*m*维数据值及其唯一标识符)和删除数据(唯一标识符),然后经MDCG-CTI算法后判断R_{MDCGCTI},如$R_{\mathrm{MDCGCTI}} \leqslant 0.5$则将新增的数据放入$\Delta C$数据集合。步骤二,如果$R_{\mathrm{MDCGCTI}} > 0.5$,则要对$\Delta C$的新增数据进行批量增量DBSCAN[18]。由此可见,由于MDCG-CTI算法能检验出数据是否具备可聚类性,因此在批量增量DBSCAN前先使用MDCG-CTI算法作为前期步骤,可以实现先验性的非固定量增量数据聚类,这不仅解决了批量增量的数据量不易确定的问题,而且还可以避免无效聚类,提高聚类结果的有效性。

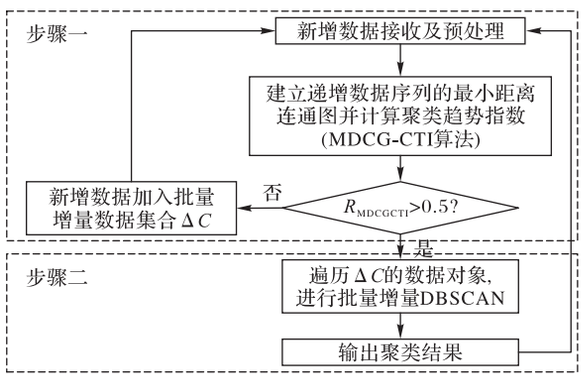


图4　结合MDCG-CTI算法的批量增量DBSCAN流程

Fig. 4　Flowchart of batch incremental DBSCAN combined with MDCG-CTI algorithm

2　时间空间复杂度

首先,MDCG新增和删除单个顶点数据的空间复杂度主要为$O(n(m+2)+n(n-1)/2+3(n-1))$。新增顶点的时间复杂度主要为$O\left(nm+\sum_{i=1}^{t}s_i\right)$,其中:*n*表示已有数据数量;*m*表示数据维度;s_i表示栈*ST*的深度;*t*表示从新增顶点数据出发,在已有数据的最邻近边集$\boldsymbol{E}$上深度优先地搜索到所有顶点的最邻近路径,直到*D*中小于$\boldsymbol{E}$的最大顶点距离的距离集合$D2(D2 \subseteq D)$对应的顶点集合$ID2(ID2 \subseteq ID)$都完成遍历后的顶点遍历个数,当数据量足够大时,$O(t)$是远小于$O(n)$的。

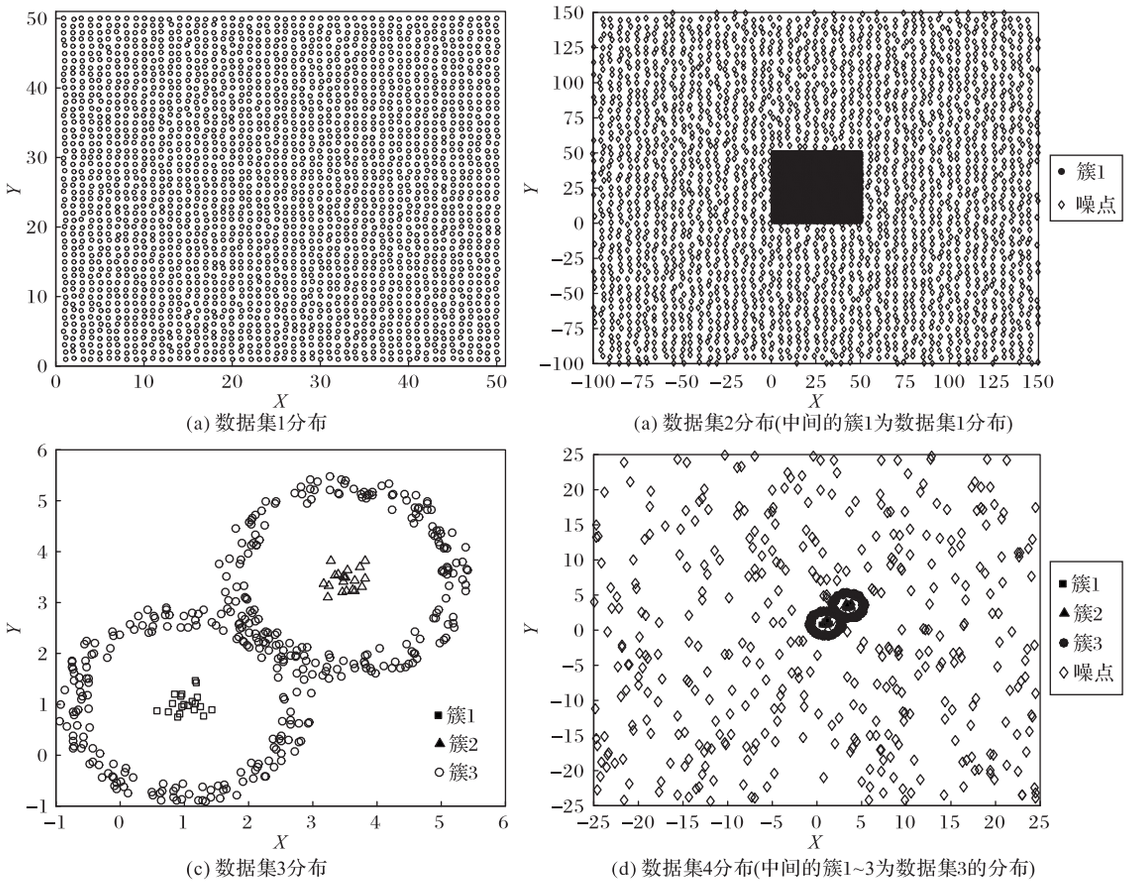
其次在MDCG-CTI算法结合DBSCAN增量聚类的运行效率方面,先比较新增数据情况下增加MDCG-CTI算法前后的时间复杂度。一次n_n个数据增量DBSCAN需要将n_n个数据依次加入到已有数据的DBSCAN聚类结果中(增量过程:先遍历已有数据查找新增数据邻域半径*Eps*内的邻域对象ON_1并判断其数量是否大于邻域对象数阈值*MinPts*,然后再针对ON_1内的每个对象做一次遍历,判断因新增数据而增加了邻域对象数的ON_1内的每个对象,其邻域半径*Eps*内的邻域对象数是否超过阈值*MinPts*),因此其时间复杂度为$O(n_n(nm+nmk_1))$,其中:*n*为已有数据量,*m*表示数据维度,k_1为ON_1的数据量。由此可见,n_n次使用增量数据MDCG-CTI算法增加的时间复杂度为$O\left(n_n\left(nm+\sum_{i=1}^{t}s_i\right)\right)$,可能相应减少一次无效的$n_n$个数据增量DBSCAN的时间复杂度$O(n_n(nm+nmk_1))$,因为增加MDCG-CTI算法后在*n*很大的情况下$O\left(\sum_{i=1}^{t}s_i\right)$会远小于$O(nmk_1)$,即增加MDCG-CTI算法后DBSCAN的时间增加量会明显小于减少量,所以MDCG-CTI

算法可以减少时间消耗、提高聚类的运行效率。

3 聚类趋势算法应用

本文算法使用 Windows 10(64 位)和 Matlab R2014b 基于一台 CPU AMD Ryzen9 3900X(12 核) 4.6 GHz,内存 32 GB 的计算机实现。

选取 7 个数据集进行测试,比较 MDCG-CTI 算法和常用聚类趋势分析算法 Hopkins、T-平方、SpecVAT 之间的运行结果及耗时情况。数据集 1~4 是自定义数据集,如图 5 所示,其中数据集 1 如图 5(a),是 2 500 个二维数据,呈现单簇无噪点的分布;数据集 2 如图 5(b),是 5 101 个二维数据,在数据集 1 的基础上增加了 2 601 个噪点,呈现单簇多噪点且噪点基本均匀的分布;数据集 3 如图 5(c),是 400 个二维数据,呈现三个簇无噪点的分布;数据集 4 如图 5(d),是 800 个二维数据,在数据集 3 的基础上增加了 400 个噪点,呈现多簇多噪点且噪点非均匀的分布。数据集 5~7 是加州大学欧文分校机器学习库 UCI 的 3 个数据集:iris、pendigits 和 avila,其中:iris 也称鸢尾花卉数据集,由 150 个花卉样本组成,每个样本包含 4 个维度;pendigits 是手写字体数据集,由 10 992 个手写体文字的图像特征组成,每个文字的图像特征有 16 个维度;avila 由阿维拉圣经中提取的 20 867 个文字的图像特征组成,每个文字的图像特征有 10 个维度。



(a) 数据集1分布　(a) 数据集2分布(中间的簇1为数据集1分布)

(c) 数据集3分布　(d) 数据集4分布(中间的簇1~3为数据集3的分布)

图 5　自定义数据集的分布

Fig. 5　Distribution of self-built datasets

表 1　自定义数据集上不同算法一次聚类的聚类趋势和运行耗时比较

Tab. 1　Comparison of clustering tendency and running time of different algorithms based on one-time clustering on self-built datasets

数据集	Hopkins			T-平方			SpecVAT			MDCG-CTI	
	抽样始点数	结果	耗时/s	抽样始点数	结果	耗时/s	抽样始点数	结果	耗时/s	结果	耗时/s
1	250	拒绝零假设(可聚类)	0.24	250	拒绝零假设(可聚类)	0.49	2	簇数 1(不可聚类)	7.74	−0.22(不可聚类)	2.61
2	500	拒绝零假设(可聚类)	0.64	500	拒绝零假设(可聚类)	1.31	2	簇数 2(可聚类)	20.82	0.28(非显著可聚类)	6.08
3	40	拒绝零假设(可聚类)	0.01	40	拒绝零假设(可聚类)	0.02	2	簇数 2(可聚类)	0.34	0.64(显著可聚类)	0.18
4	80	拒绝零假设(可聚类)	0.02	80	拒绝零假设(可聚类)	0.04	2	簇数 2(可聚类)	1.58	0.41(非显著可聚类)	0.36

注:SpecVAT 选取特征向量最大个数以内各相异度图的自动确定簇数(1 或 2)作为是否可聚类标准。

表 1 的 4 个数据集中,数据集 1 是单簇不可聚类,数据集 3 多簇显著可聚类,数据集 2 和 4 因为噪点过多(噪点量大于等于簇数据量)所以非显著可聚类。从这四种聚类趋势算法的比较中可以看出,MDCG-CTI 算法对聚类趋势的判断是完全

正确的,但其余三种算法因为只寻找数据中的自然簇而忽视了噪点过多会导致大量数据无法聚类成簇,因此聚类趋势不显著的问题,所以 Hopkins、T-平方和 SpecVAT 对数据集 2 和 4 的聚类趋势判断都不准确;此外 Hopkins 和 T-平方对数据集 1 的聚类趋势判断也是错误的,这是因为两种算法均认为单簇分布是具有聚类趋势的[22]。运行耗时方面,Hopkins 和 T-平方是具有明显优势的,这是因为抽样方法无需遍历全体数据,而其耗时主要取决于抽样始点个数的设置,设置得越多抽样密度越高,结果的准确性就相对越好,但相应的耗时也会增加。SpecVAT 和 MDCG-CTI 相比,前者的耗时较多,且当数据量越大时越明显,这是因为该算法要计算特征向量(时间复杂度为数据量的三次方),且有可视化操作步骤(需要从不同特征向量个数的 SpecVAT 相异度图中用 ADNC 算法自动确定簇数)的关系。因此,综合准确性和运行耗时来看,MDCG-CTI 算法是具有明显优势的,其适用于增量数据序列的特点不仅带来了耗时相对较少的优点,而且还可以分析全体数据(并非抽样分析)以提高聚类趋势判断的准确性。

从表 2 可以看出,Hopkins 和 T-平方在递增数据序列上的表现是 100 次运行的平均可聚类比率 *AR*(可聚类比率指数据在递增过程中可聚类判断次数占总聚类趋势判断次数的比值)大于 94%,SpecVAT 则在 87% 以上,这都明显高于 MDCG-CTI 的平均显著可聚类比率,究其原因:一方面是因为误认为单簇分布是可聚类的,另一方面则是将含大量噪点的数据集判断为可聚类。所以相比之下,MDCG-CTI 对聚类趋势的判断要更准确,从而其相对较低的可聚类比率在后续的批量增量 DBSCAN 中便可以更有效地减少无效聚类,降低聚类耗时,同时提高聚类结果的有效性。而从算法各运行 100 次的可聚类比率标准差 *ER* 上来看,Hopkins 和 T-平方基本上在 0.01 左右,SpecVAT 在 0.006 左右,MDCG-CTI 则约为 0.003,可见 MDCG-CTI 算法的波动幅度最小,稳定性最高,而 Hopkins 和 T-平方则因为抽样的关系导致结果相对片面和不稳定。

此外平均累计耗时 *AT* 方面,Hopkins 和 T-平方较低,MDCG-CTI 居中,SpecVAT 最高(SpecVAT 采用的谱方法每增一次数据就需要重新计算一次特征向量和自动确定一次簇数),这和单次的聚类趋势计算耗时情况(如表 1)是一致的,尤其在大数据集 pendigits 和 avila 上,MDCG-CTI 相较 SpecVAT 数据递增平均累计耗时降低了 38% 和 42%。

表 3 列出了四种聚类趋势算法结合批量增量 DBSCAN[18] 的聚类结果,其中 Hopkins、T-平方、SpecVAT、MDCG-CTI 四种算法的参数如表 2,DBSCAN 参数 *Eps* 和 *MinPts* 采用自适应算法[23]自动生成,平均准确率 *CR* 计算如下:

$$CR = \frac{1}{N_{\text{all}} - N_{\text{nc}} - 99} \sum_{i=1}^{N_{\text{all}} - N_{\text{nc}} - 99} C_i / N_i$$

其中:N_i 表示第 i 次随机递增且可聚类的累计数据量,C_i 表示 N_i 中被 DBSCAN 正确划分类别的数据量,N_{nc} 表示不可聚类次数。

表 3 中在数据按照数据量 N_{add} 随机递增的过程中,由于增加聚类趋势算法所减少的若干次不具有聚类趋势(不可聚类)的 DBSCAN 聚类,其结果大多是聚类成单簇或噪点过多而导致准确率不高,所以用了聚类趋势分析算法以后聚类的 *CR* 便会提高,而因 MDCG-CTI 的平均可聚类比率 *AR* 是最低的(见表 2),所以其后续 DBSCAN 的聚类 *CR* 也就相应最高,相较 SpecVAT+DBSCAN 在数据集 pendigits 和 avila 上分别提高了 6 和 11 个百分点。此外聚类累计耗时 *T* 方面,iris 数据集由于数据量小而聚类耗时很短,所以使用聚类趋势算法后耗时均比不使用要多,且以 SpecVAT 最为突出;但是在 pendigits 和 avila 两个大数据集上,聚类趋势算法减少 DBSCAN 聚类耗时的优点则显露了出来,其中 Hopkins 由于 100% 可聚类的关系(见表 2)所以耗时比只用 DBSCAN 多,而其余三种聚类趋势算法结合 DBSCAN 后累计耗时 *T* 都有所减少,且 MDCG-CTI 减少得最为明显,相较 SpecVAT+DBSCAN 的聚类累计耗时分别降低了 7% 和 8%。

表 2　UCI 数据集上不同算法的平均可聚类比率分布和运行耗时比较(算法各运行 100 次)

Tab. 2　Comparison of clusterable rate distribution and running time of different algorithms on UCI datasets (100 runs for each algorithm)

数据集	Hopkins				T-平方				SpecVAT				MDCG-CTI			
	抽样始点数	*AR*/%	*ER*	*AT*/h	抽样始点数	*AR*/%	*ER*	*AT*/h	特征向量最大数	*AR*/%	*ER*	*AT*/h	*AR*/%	*ER*	*AT*/h
iris	$N_{\text{add}}\times 10\%$ 取整	96	0.012	0.000 08	$N_{\text{add}}\times 10\%$ 取整	94	0.010	0.000 16	2	87	0.006	0.000 46	77	0.003	0.000 2
pendigits		100	0.009	0.69		99	0.009	1.57		93	0.006	6.29	90	0.004	3.87
avila		100	0.010	4.97		99	0.009	11.39		99	0.005	27.35	97	0.003	15.84

注:每次递增的数据量 $N_{\text{add}} = \{100, 101, 102, \cdots, N_{\text{all}}\}$,$N_{\text{all}}$ 为总数据量。

表 3　UCI 数据集上不同算法结合 DBSCAN 的聚类结果比较

Tab. 3　Comparison of clustering results of different algorithms combined with DBSCAN on UCI datasets

数据集	DBSCAN		Hopkins+DBSCAN		T-平方+DBSCAN		SpecVAT+DBSCAN		MDCG-CTI+DBSCAN	
	CR/%	*T*/h	*CR*/%	*T*/h	*CR*/%	*T*/h	*CR*/%	*T*/h	*CR*/%	*T*/h
iris	69	0.000 3	77	0.000 4	79	0.000 5	85	0.000 75	96	0.000 4
pendigits	61	12.2	62	12.9	70	11.8	82	10.5	88	9.8
avila	53	54.9	53	59.8	69	52.3	74	52.7	85	48.6

4　结语

本文提出的 MDCG-CTI 算法能够利用动态递增数据的特点,基于数据的最小距离连通图计算出聚类趋势指数 R_{MDCGCTI},并通过实验分析得出以下结论:$R_{\text{MDCGCTI}} \leqslant 0$ 时不可聚类;$0 < R_{\text{MDCGCTI}} \leqslant 0.5$ 时非显著可聚类;$R_{\text{MDCGCTI}} > 0.5$ 则显著可聚类。时间复杂度分析及实验表明,MDCG-CTI 算法可以提高后续批量增量 DBSCAN 的结果有效性及运行效率;同时在多个数据集上的比较则表明,MDCG-CTI 算法在损失有限运算时间的前提下可以有效提高聚类趋势判断的准确性,

特别是在处理大数据集时具有不错的性能表现。

未来的研究将进一步提高 MDCG-CTI 算法在处理大数据集上的效率优势，计划从两个方面进行算法的并行化：一是利用 CPU 的多核运算能力同时计算以顶点 id_i 为出发点的多条深度优先遍历的最邻近路径；二是把算法分为在线和离线两部分，在线生成最小距离连通图 MDCG[ID,E]，离线统计聚类趋势指数 R_{MDCGCTI}，实现按需判断多批量增量数据的聚类趋势。

参考文献（References）

[1] 褚娜，马利庄，王彦. 聚类趋势问题的研究综述[J]. 计算机应用研究，2009，26(3)：801-803，822.（CHU N，MA L Z，WANG Y. Research for clustering tendency [J]. Application Research of Computers，2009，26(3)：801-803，822.）

[2] HOPKINS B，SKELLAM J G. A new method for determining the type of distribution of plant individuals [J]. Annals of Botany，1954，18(2)：213-227.

[3] PANAYIRCI E，DUBES R C. Extension of the Cox-Lewis method for testing multidimensional data[J]. Pattern Recognition，1988，7(1)：1-8.

[4] BESAG J E，GLEAVES J T. On the detection of spatial pattern in plant communities [J]. Bulletin of the International Statistical Institute，1973，45：153-158.

[5] 高新波，裴继红，谢维信. 基于统计检验指导的聚类分析方法[J]. 电子科学学刊，2000，22(1)：6-12.（GAO X B，PEI J H，XIE W X. A novel cluster analysis method supervised by statistical tests[J]. Journal of Electronics，2000，22(1)：6-12.）

[6] 曾广周. 聚类趋势的 Monte-Carlo 检验[J]. 计算机应用与软件，1989，6(1)：35-40.（ZENG G Z. Monte-Carlo test of clustering tendency[J]. Computer Applications and Software，1989，6(1)：35-40.）

[7] SILVA H B，BRITO P，DA COSTA J P. A partitional clustering algorithm validated by a clustering tendency index based on graph theory[J]. Pattern Recognition，2005，39(5)：776-788.

[8] BEZDEK J C，HATHAWAY R J. VAT：a tool for visual assessment of (cluster) tendency [C]// Proceedings of the 2002 International Joint Conference on Neural Networks. Piscataway：IEEE，2002：2225-2230.

[9] HATHAWAY R J，BEZDEK J C，HUBAND J M. Scalable visual assessment of cluster tendency for large data sets [J]. Pattern Recognition，2006，39(7)：1315-1324.

[10] HUBAND J M，BEZDEK J C，HATHAWAY R J. bigVAT：visual assessment of cluster tendency for large datasets [J]. Pattern Recognition，2005，38(11)：1875-1886.

[11] WANG L，GENG X，BEZDEK J，et a1. SpecVAT：enhanced visual cluster analysis [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Piscataway：IEEE，2008：638-647.

[12] REDDY B E，PRASAD K R. Improving the performance of visualized clustering method [J]. International Journal of System Assurance Engineering and Management，2016，7(S1)：102-111.

[13] 赖家文，彭显刚，王洪森，等. 霍普金斯统计在短期负荷预测中的应用探讨[J]. 广东电力，2013，26(8)：89-93，98.（LAI J W，PENG X G，WANG H S，et a1. Discussion on application of Hopkins statistics in short-term load prediction [J]. Guangdong Electric Power，2013，26(8)：89-93，98.）

[14] 李晔，陈奕延，张淑芬. 基于密度峰值的混合型数据聚类算法设计[J]. 计算机应用，2018，38(2)：483-490，496.（LI Y，CHEN Y Y，ZHANG S F. Design of mixed data clustering algorithm based on density peak [J]. Journal of Computer Applications，2018，38(2)：483-490，496.）

[15] 孙谦，姚建刚，李欣然，等. 基于聚类趋势分析与逐步回归的电铁牵引负载负序源模型研究[J]. 中国电机工程学报，2012，32(34)：120-128.（SUN Q，YAO J G，LI X R，et a1. Study on negative sequence source model of electrified railway traction load based on clustering tendency analysis and stepwise regression[J]. Proceedings of the CSEE，2012，32(34)：120-128.）

[16] ESTER M，KRIEGEL H P，SANDER J，et a1. 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. Palo Alto，CA：AAAI Press，1996：226-231.

[17] 毛睿，张贺，陆敏华，等. 一种基于密度的增量聚类数据挖掘方法及系统：CN201610055222.2[P]. 2016-07-06.（MAO R，ZHANG H，LU M H，et a1. A density-based incremental clustering data mining method and system：CN201610055222.2[P]. 2016-07-06.）

[18] 田路强. 基于 DBSCAN 的分布式聚类及增量聚类的研究与应用[D]. 北京：北京工业大学，2016：29-34.（TIAN L Q. Research and application on distributed clustering and incremental clustering based on DBSCAN[D]. Beijing：Beijing University of Technology，2016：29-34.）

[19] 任群. 基于增量聚类的协流识别系统研究[J]. 西安文理学院学报(自然科学版)，2018，21(1)：63-67.（REN Q. Research on the co-flow recognition system based on incremental clustering[J]. Journal of Xi'an University (Natural Science Edition)，2018，21(1)：63-67.）

[20] JUNIOR M，OLIVIERI B，ENDLER M. DG2CEP：a near real-time on-line algorithm for detecting spatial clusters large data streams through complex event processing[J]. Journal of Internet Services and Applications，2019，10(1)：No. 8.

[21] 朱迪，陈丹伟. 基于密度聚类和随机森林的移动应用识别技术[J]. 计算机工程与应用，2020，56(4)：63-68.（ZHU D，CHEN D W. Technology of mobile application identification based on density-based clustering and random forest [J]. Computer Engineering and Applications，2020，56(4)：63-68.）

[22] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社，2004：134-136.（GAO X B. Fuzzy Clustering Analysis and its Applications[M]. Xi'an：Xidian University Press，2004：134-136.）

[23] 王光，林国宇. 改进的自适应参数 DBSCAN 聚类算法[J/OL]. 计算机工程与应用.[2020-01-20]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20191230. 1437. 008. html.（WANG G，LIN G Y. Improved adaptive parameter DBSCAN clustering algorithm [J/OL]. Computer Engineering and Applications.[2020-01-20]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20191230. 1437. 008. html.）

This work is partially supported by the National Key Research and Development Program of China (2018YFC1505804).

FAN Zhongxin，born in 1981，M. S.，engineer. His research interests include data mining，neural network.