Journal of Computer Applications
计算机应用,2018,38(9):2511 - 2514,2542
ISSN 1001-9081
CODEN JYIIDU
2018-09-10
http://www.joca.cn

文章编号:1001-9081(2018)09-2511-04
DOI:10.11772/j.issn.1001-9081.2018010177

基于拉普拉斯中心性和密度峰值的无参数聚类算法

邱保志,程　栾*
(郑州大学 信息工程学院,郑州 450001)
(*通信作者电子邮箱 cheng_luan@163.com)

摘　要:针对聚类算法的聚类中心选取需要人工参与的问题,提出了一种基于拉普拉斯中心性和密度峰值的无参数聚类算法(ALPC)。首先,使用拉普拉斯中心性度量对象的中心性;然后,使用正态分布概率统计方法确定聚类中心对象;最后,依据对象到各个中心的距离将各个对象分配到相应聚类中心实现聚类。所提算法克服了算法需要凭借经验参数和人工选取聚类中心的缺点。在人工数据集和真实数据集上的实验结果表明,与经典的具有噪声的基于密度的聚类方法(DBSCAN)、密度峰值聚类(DPC)算法以及拉普拉斯中心峰聚类(LPC)算法相比,ALPC 具有自动确定聚类中心、无参数的特点,且具有较高的聚类精度。

关键词:聚类算法;拉普拉斯矩阵;密度峰值;无参数聚类;正态分布
中图分类号:TP391.4　　**文献标志码**:A

Parameter-free clustering algorithm based on Laplace centrality and density peaks

QIU Baozhi, CHENG Luan*
(*School of Information Engineering, Zhengzhou University, Zhengzhou Henan* 450001, *China*)

Abstract: In order to solve the problem of selecting center manually in a clustering algorithm, a Parameter-free Clustering Algorithm based on Laplace centrality and density peaks (ALPC) was proposed. Laplace centrality was used to measure the centrality of objects, and a normal distribution probability statistical method was used to determine clustering centers. The problem that clustering algorithms rely on empirical parameters and manually determine cluster centers was solved by the proposed algorithm. Each object was assigned to the corresponding cluster center according to the distance between the object and the center. The experimental results on synthetic data sets and UCI data sets show that the new algorithm can not only automatically determine cluster centers without any priori parameters, but also get better results with higher accuracy compared with the Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm, Clustering by fast search and find of Density Peaks (DPC) algorithm and Laplace centrality Peaks Clustering (LPC) algorithm.

Key words: clustering algorithm; Laplace matrix; density peak; parameter-free clustering; normal distribution

www.joca.cn

0　引言

聚类是将相似的数据对象聚集在一起,而相异的数据对象尽可能地分离[1]。聚类技术在数据挖掘、模式识别、图像处理、信息检索等领域有着广泛的应用,是数据分析的基础。在现有的聚类算法中,基于聚类中心的聚类算法由于其应用广泛而备受人们的关注,例如 K-means[2]、具有噪声的基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)[3]、密度峰值聚类(Clustering by fast search and find of Density Peaks, DPC)算法[4]等,这些算法都是以选择初始聚类中心作为聚类的起始点,通过迭代或扩展进行聚类。选择聚类中心的准确性对聚类的迭代效率、聚类个数的确定和聚类精度有着重要的影响,如何准确地选择聚类中心、提高聚类精度成为聚类技术研究的关键问题。

以 *K*-means 为代表的基于划分的聚类首先初始化聚类中心,然后通过聚类中心的反复迭代的方式进行聚类,这类算法由于初始聚类中心选择的任意性,导致聚类结果精度不高;以 DBSCAN 为代表的基于密度聚类的算法是以任一个核心点作为聚类的起始点,将密度可达的对象集合看作一个聚类[5],由于其中心点选择的随意性和密度定义的全局性,造成该类聚类算法不能对多密度数据集和高维数据集进行有效的聚类;基于密度峰值的聚类算法 DPC 以局部密度峰值为聚类中心,将包围密度峰值的低密度区域看成一个聚类[6-7],虽然能准确地确定聚类的中心,但在确定密度峰值时需人工选取相应的参数,在没有先验知识的情况下,很难确定聚类的中心[8-9]。为了解决 DPC 算法存在的问题,拉普拉斯中心峰聚类(Laplace centrality Peaks Clustering, LPC)算法[10]将数据集转换为加权完全图,通过计算节点拉普拉斯中心性评估每个节点在加权图中的重要性,并计算一个节点到更高拉普拉斯中心性的节点的相对距离,将拥有较大拉普拉斯中心性且较大相对距离的对象确定为聚类中心,以此来进行聚类,但这一过程需要建立决策图人工选择聚类中心。

针对上述问题,本文通过引入图的拉普拉斯能量,综合考量每个节点的拉普拉斯中心性和该节点到具有更高拉普拉斯中心性的节点的相对距离,将数据集映射到由数据对象的拉普拉斯中心性和该对象的相对距离组成的二维特征空间中;

收稿日期:2018-01-21;**修回日期**:2018-04-04;**录用日期**:2018-04-10。　**基金项目**:河南省基础与前沿基金资助项目(152300410191)。
作者简介:邱保志(1964—),男,河南驻马店人,教授,博士,CCF 会员,主要研究方向:数据挖掘、人工智能;　程栾(1992—),女,河南周口人,硕士研究生,主要研究方向:数据挖掘。

进而构造决策图,依据正态分布的“3σ”准则提取聚类中心,从而自动确定聚类中心;最后将非聚类中心点指派到距离最近的数据对象所属聚类中,完成整个数据集的聚类。

1 相关定义

设数据集 $D = \{x_1, x_2, \cdots, x_n\}$,其中 $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$,$i = 1, 2, \cdots, n$,$m$ 为数据对象的维度。用 $w_{i,j}$ 表示数据对象 x_i 和 x_j 之间的欧几里得距离,首先把数据集转换为加权完全图,数据集中每个数据对象作为图 G 中的一个节点,数据对象 x_i 和 x_j 之间的欧几里得距离作为图 G 中节点 x_i 和 x_j 之间边的权值。

定义 1 相异度矩阵[10]。设加权完全图 G 是由数据集 D 转换得到的,则图 G 的相异度矩阵记为 $\boldsymbol{W}(G)$,定义如下:

$$\boldsymbol{W}(G) = \begin{bmatrix} 0 & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & 0 & \cdots & w_{2,n} \\ \vdots & \vdots & & \vdots \\ w_{n,1} & w_{n,2} & \cdots & 0 \end{bmatrix} \tag{1}$$

其中,$w_{i,j}$ 为节点 x_i 和 x_j 之间的欧几里得距离,对任意的 $1 \leqslant i, j \leqslant n$ 都有:$w_{i,j} \geqslant 0$,$w_{i,j} = w_{j,i}$,$w_{i,i} = 0$。

定义 2 和对角矩阵[10]。给定节点 $x_i \in G$,定义 Z_i 是节点 x_i 到其他节点的距离总和,即 $Z_i = \sum_{j=1}^{n} w_{i,j}$。图 G 的和对角矩阵 $\boldsymbol{Z}(G)$ 定义为:

$$\boldsymbol{Z}(G) = \mathrm{diag}(Z_1, Z_2, \cdots, Z_n) \tag{2}$$

定义 3 拉普拉斯矩阵[10]。给定加权完全图 G,它的拉普拉斯矩阵定义为:$\boldsymbol{L}(G) = \boldsymbol{Z}(G) - \boldsymbol{W}(G)$。其中:$\boldsymbol{Z}(G)$ 为图的和对角矩阵,$\boldsymbol{W}(G)$ 为图的相异度矩阵。

定义 4 拉普拉斯能量[11]。对于给定数据集 D 转换的加权完全图 G,图 G 的拉普拉斯能量 $LEL(G)$ 定义为其拉普拉斯矩阵的特征值的平方和,即:

$$LEL(G) = \sum_{i=1}^{n} \lambda^2 \tag{3}$$

其中,$\lambda_1, \lambda_2, \cdots, \lambda_n$ 是图 G 的拉普拉斯矩阵 $\boldsymbol{L}(G)$ 的特征值。

定义 5 拉普拉斯中心性[11]。给定加权完全图 G,删除节点 x_i 及其所有相连的边之后的图记为图 G_i,给定 $x_i \in G$,节点 x_i 的拉普拉斯中心性 c_i 定义为:

$$c_i = (LEL(G) - LEL(G_i))/LEL(G) \tag{4}$$

$LEL(G_i)$ 是删除节点 x_i 之后图 G_i 的拉普拉斯能量。节点的重要性往往体现在该节点被移除之后对图的拉普拉斯能量的影响,因此可以用图中删除某节点后拉普拉斯能量的相对下降来度量一个节点的重要性(中心性),拉普拉斯中心性越高,则该节点越重要。

定义 6 相对距离[4]。节点 x_i 到具有更高拉普拉斯中心性的节点的相对距离 δ_i 定义为:当节点具有最大拉普拉斯中心性时,相对距离 δ_i 表示 x_i 与数据集中其他对象的最远距离,即 $\delta_i = \max\{w_{i,j} \mid j = 1, 2, \cdots, n\}$;否则,相对距离 δ_i 表示节点 x_i 与比其拉普拉斯中心性高的节点之间的最近距离,即 $\delta_i = \min_{j: c_j > c_i} w_{i,j}$。

当数据对象 x_i 为聚类中心时,c_i 和 δ_i 具有较大的值,令 $\gamma_i = c_i * \delta_i$,这时,$\gamma_i$ 具有较大的值。为了从数据对象中分离出聚类中心,采用式(5)对 γ_i 进行放大[12],加大聚类中心与其他数据对象的差别,从而便于识别真正的聚类中心。

$$E_i = \sum_{j \in D \setminus \{i\}} \sqrt{(\gamma_i - \gamma_j)^2} \tag{5}$$

定义 7 正态分布[13]。若随机变量 X 服从一个数学期望为 μ、标准差为 σ 的概率分布,且其概率密度函数为:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

则这个随机变量 X 就称为正态随机变量,正态随机变量服从的分布就称为正态分布。

在正态分布中,随机变量 X 落在 $(\mu - 3\sigma, \mu + 3\sigma)$ 以外的概率仅为 0.0027,基本上可以把区间 $(\mu - 3\sigma, \mu + 3\sigma)$ 看作是随机变量 X 实际可能的取值区间,这称之为正态分布的“3σ”准则。

当数据对象 x_i 为聚类中心时,该对象具有较高的拉普拉斯中心性 c_i 和较大的相对距离 δ_i 两个特征。潜在聚类中心点所对应的 E 值远远大于非聚类中心点所对应的 E 值。由于聚类中心点拥有较大的 E 值,而非聚类中心点的 E 值较小这一特点,依据正态分布“3σ”准则筛选出那些 E 值大于 $\mu + 3\sigma$ 的数据对象,将其设为聚类中心。图 1 是 Sprial 数据集所对应的分布直方图及其正态分布曲线。从图 1 中可以看出,Sprial 数据集中有 3 个数据对象的 E 值大于 $\mu + 3\sigma$,这 3 个数据对象就是数据集 Sprial 的 3 个类中心。

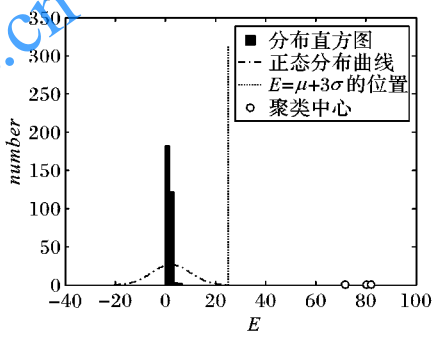

图 1 数据集 Sprial 的分布直方图和正态分布曲线

Fig. 1 Distribution histogram and normal distribution curve of Sprial dataset

2 ALPC 描述

基于拉普拉斯中心性和密度峰值的无参数聚类算法(Parameter-free Clustering Algorithm based on Laplace centrality and density peaks, ALPC)首先把数据集转换为无向完全图 G,得到图的拉普拉斯矩阵 $\boldsymbol{L}(G)$;然后根据删除某个数据对象后所引起的拉普拉斯能量变化,来得到数据对象的拉普拉斯中心性 c_i,并计算得到该数据对象的相对距离 δ_i,接着通过式(5)扩大 γ 值之间的差别,通过正态分布,选取 E 值大于 $\mu + 3\sigma$ 的数据对象为聚类中心,然后按照最近邻分类完成对整个数据集的聚类。ALPC 算法描述如算法 1 所示。

算法 1 ALPC。

输入 数据集 D;

输出 聚类结果。

步骤 1 将数据集 D 转换为带权完全图根据定义 1 ~ 3 计算 $\boldsymbol{W}(G)$、$\boldsymbol{Z}(G)$ 和 $\boldsymbol{L}(G)$。

步骤 2 根据式(3)、式(4)计算各数据对象的拉普拉斯中心性。

步骤 3 提取数据集聚类中心:

www.joca.cn

a)根据定义 6 计算各数据对象的相对距离 δ_i;

b)根据式(5)计算各数据对象的 E 值;

c)运用正态分布的 3σ 准则,将 E 值大于 $\mu+3\sigma$ 的数据对象作为聚类中心。

步骤 4　按照最近邻分类完成对整个数据集的聚类。

在 ALPC 中,首先需要将待聚类的数据集转换为带权的无向完全图,图中的每个节点代表数据集中每个对象,利用拉普拉斯中心性来度量对象的重要性。拉普拉斯中心性基本思想是:通过删除某个节点后所引起的图的拉普拉斯能量的相对下降来度量该节点的中心性。然后计算该对象到具有更高拉普拉斯中心性的对象的相对距离,在聚类中心选取时,ALPC 不再手动选取那些具有较大拉普拉斯中心性和较大相对距离的数据对象作为聚类中心,而是采用概率统计中正态分布理论选取聚类中心,避免了手动选取聚类中心时因视觉误差所带来的聚类结果的不准确问题。ALPC 综合考虑对象的拉普拉斯中心性和该对象到更高拉普拉斯中心性的对象的相对距离,通过式(5)加大潜在聚类中心与待聚类数据集中其他对象的差别,利用概率统计中的正态分布理论选取具有较大 E 值,即 E 值大于 $\mu+3\sigma$ 的数据对象为聚类中心,然后按照最近邻分类原则将其余的点分配到相应的聚类中心,从而完成聚类并输出聚类结果。

3　实验结果与分析

该算法的实验环境为:CPU 为 AMD 3.4 GHz,4 GB 内存,操作系统为 Windows 7(64 bit),算法的编写使用 Matlab R2014a 软件。

3.1　实验数据集及评价指标

实验选用 10 个数据集,其中 4 个人工数据集分别用于验证 ALPC 识别桥接簇、螺旋形、球形簇和非球形簇的能力;6 个真实数据集来源于 UCI 机器学习数据库,用于验证 ALPC 对高维数据处理的有效性。表 1 详细描述了实验数据集的基本信息。

表 1　实验数据集基本信息

Tab. 1　Basic information of experimental datasets

数据集类别	数据集名称	样本数	维度	类数
人工合成	4k2_far	400	2	4
	Aggregation	788	2	7
	Spiral	312	2	3
	Ring	556	2	5
真实	Iris	150	4	3
	Wine	178	13	3
	Seed	210	7	3
	Glass	214	9	7
	Sonar	208	60	2
	Wdbc	569	31	2

实验的评价指标采用准确率(ACCuracy, ACC)[10]、调整兰德系数(Adjusted Rand Index, ARI)[5]和归一化互信息(Normalized Mutual Information, NMI)[5]度量 ALPC 聚类的有效性。准确率衡量聚类结果与实际结果的符合程度,其取值范围为[0,1],当算法的聚类结果与真实情况完全一致时准确率为 1,两者越不一致,准确率越低;调整兰德系数的计算需要待聚类数据集的真实分类结果,其取值范围为[-1,1],调整兰德系数越高意味着算法的聚类结果与数据集的实际分类情况越吻合,从广义的角度看,调整兰德系数衡量的是算法聚类结果与数据集实际分类情况的吻合程度;归一化互信息是衡量聚类结果与数据集实际分类情况的相近程度,同样需要待聚类数据集的真实分类结果,归一化互信息的取值范围为[0,1],值越大意味着聚类结果与真实情况越相近,聚类效果越好。

3.2　聚类结果分析

首先在二维数值属性数据集上与 DBSCAN、DPC 和 LPC 算法进行聚类有效性的比较,然后在高维数据集上与 DBSCAN、DPC 和 LPC 算法进行聚类有效性的比较。

图 2 依次是 DBSCAN、DPC、LPC 和 ALPC 在 4k2_far、Aggregation、Spiral 和 Ring 数据集上的聚类结果,不同的灰度值代表不同的聚类。

图 2(a)是 DBSCAN 算法的聚类结果,它在 Aggregation 和 Ring 数据集上的聚类结果不正确,说明 DBSCAN 算法不能识别桥接簇和非球形簇,这是因为 DBSCAN 算法的半径阈值和密度阈值定义的全局性,当待聚类数据集的密度不均匀、聚类间距差很大时,参数选取困难,半径阈值和密度阈值选取不合适会造成聚类结果的不准确;图 2(b)是 DPC 算法在 4 个人工数据集上的聚类结果,它在 Ring 数据集上的聚类结果不正确,这是因为 DPC 算法需要人工输入截断距离和手动选取聚类中心,截断距离定义的全局性和视觉误差造成的聚类中心选取不准确,就会影响聚类结果的准确性;图 2(c)是 LPC 算法的聚类结果,它在 4k2_far 数据集上聚类结果不正确,这是因为 LPC 算法人为分析决策图,手动选取聚类中心时视觉误差造成的聚类中心选取不准确,使聚类结果不准确;图 2(d)是算法 ALPC 的聚类结果,表明 ALPC 可以准确识别出数据集中的球形簇、桥接簇、螺旋形和非球形簇。上述四个二维数据集上的聚类结果验证了 ALPC 的有效性。

表 2 给出了 4 种算法在人工数据集和真实数据集上的实验结果。从表 2 中可以看出:与其他三个算法相比,ALPC 的 ACC 值有 10 项为最高,NMI 值有 7 项为最高,ARI 值有 7 项为最高;虽然 ALPC 在 Iris、Seed 和 Sonar 这三个数据集上的归一化互信息和调整兰德系数不是最高值,但与评价指标最高值接近。

ALPC 对于 Spiral、Ring 和 4k2_far 这三个数据集的执行能力最好,其聚类结果完全正确;与 LPC 算法相比较,ALPC 在不需要人为干预的情况下对于 Spiral、Aggregation、Ring、Wine、Seed 和 Wdbc 这六个数据集自动划分的效果与人为划分效果一致,实现了自动聚类的目标;ALPC 对 4k2_far、Iris、Glass 和 Sonar 这四个数据集自动划分的聚类结果的精度比 LPC 算法人为划分的聚类结果的精度要高,说明 ALPC 在应对一些数据集时,具有更优秀的性能,能得到较好的聚类效果。综上所述,ALPC 对人工数据集和真实数据集均有良好的聚类结果。

3.3　时间复杂度分析

对于对象数量为 n 的数据集 D:1)DPC 算法的时间复杂度主要来自于距离矩阵、局部密度和相对距离的计算过程,其中计算距离矩阵和相对距离的时间复杂度为 $O(n^2)$,计算局部密度需要依次寻找与该对象相对距离小于截断距离的对象的时间复杂度最坏为 $O(n^2)$,所以 DPC 算法的时间复杂度为 $O(n^2)$。2)DBSCAN 算法的时间复杂度为 $O(n^2)$;LPC 算法

www.joa.cn

的时间复杂度主要来自于距离矩阵、拉普拉斯中心性和相对距离的计算过程，其中计算距离矩阵和相对距离的时间复杂度为 $O(n^2)$，计算对象拉普拉斯中心性需要依次删除该对象并计算拉普拉斯能量的相对下降，时间复杂度最坏为 $O(n^3)$，所以 LPC 算法的时间复杂度为 $O(n^3)$。3) ALPC 与 LPC 算法类似，所以 ALPC 算法的时间复杂度同样为 $O(n^3)$。

LPC 算法不需要人工输入参数，但仍然需要人工选取聚类中心，虽然 ALPC 和 LPC 算法的时间复杂度相同，但 ALPC 可以自动选取聚类中心，避免了人工选取聚类中心的不准确问题。

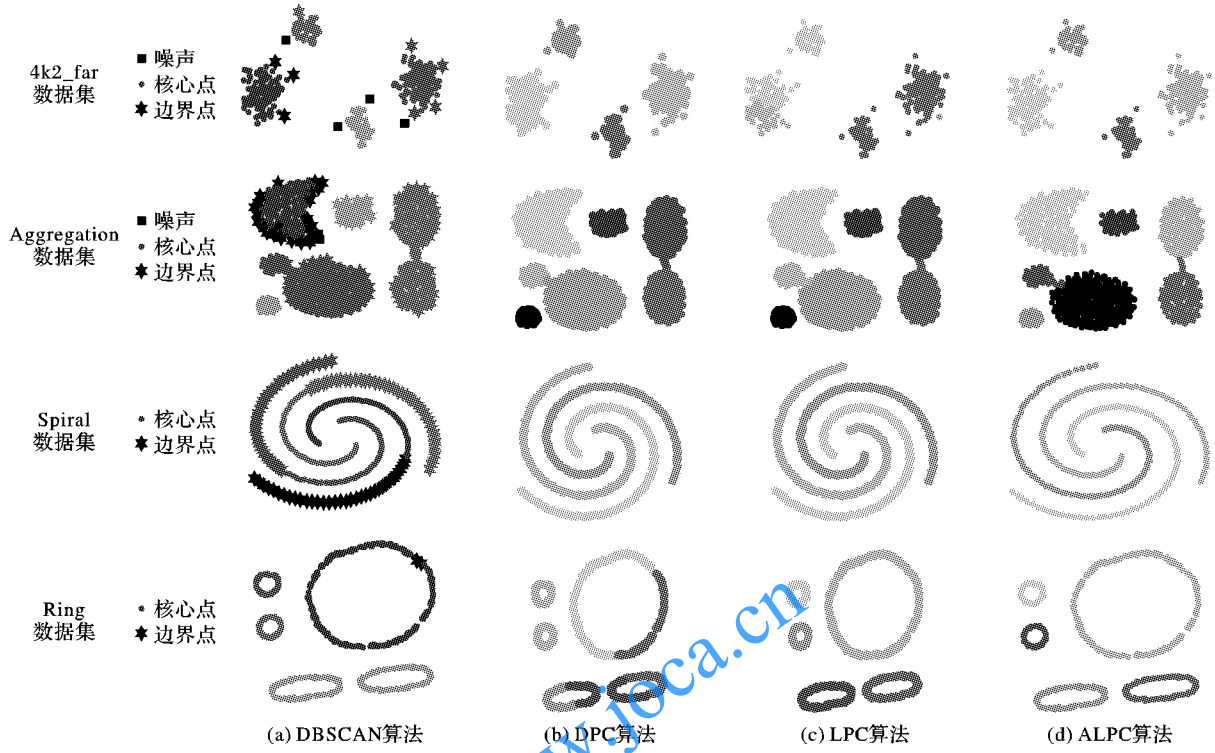

图 2 4 种算法在人工数据集上聚类结果比较

Fig. 2 Comparison of clustering results on artificial datasets by four algorithms

www.joca.cn

表 2 不同算法在不同数据集上的聚类指标对比

Tab. 2 Clustering index results of different algorithms on different datasets

数据集类别	数据集名称	DBSCAN			DPC			LPC			ALPC		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
人工合成	4k2_far	0.9900	0.9500	0.9100	1.0000	1.0000	1.0000	0.6675	0.6648	0.5202	1.0000	1.0000	1.0000
	Aggregation	0.8274	0.8690	0.8539	0.9429	0.9358	0.9214	0.9873	0.9700	0.9740	0.9873	0.9700	0.9740
	Spiral	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Ring	0.6861	0.5656	0.3231	0.5755	0.6749	0.4962	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
真实	Iris	0.6667	0.7612	0.5681	0.6667	0.6586	0.4531	0.6533	0.6409	0.5252	0.6933	0.7229	0.5609
	Wine	0.5337	0.2100	0.0154	0.5843	0.2802	0.1794	0.6461	0.3633	0.3380	0.6461	0.3633	0.3380
	Seed	0.3905	0.0283	0.0042	0.6206	0.7560	0.7340	0.7476	0.5032	0.4615	0.7476	0.5032	0.4615
	Glass	0.3832	0.2239	0.1055	0.4579	0.2419	0.0924	0.3505	0.1434	0.0525	0.4860	0.2938	0.1552
	Sonar	0.5577	0.0431	0.0569	0.4524	0.1056	0.0191	0.5385	0.0061	0.0011	0.5673	0.0181	0.0134
	Wdbc	0.6274	0.0000	0.0000	0.6130	0.0090	−0.011	0.6327	0.0112	0.0109	0.6327	0.0112	0.0109

4 结语

本文应用拉普拉斯理论和正态分布的“3σ”准则，提出了一种无参数聚类算法，实现了聚类中心的自动选择和无参数聚类，解决了 LPC 算法人工选取聚类中心的不准确性问题，为基于中心的聚类技术的聚类中心选取提供了理论基础，提高了聚类精度。

参考文献（References）

[1] HAN J W, KAMBER M. 数据挖掘：概念与技术[M]. 范明，孟小峰，译. 2 版. 北京：机械工业出版社，2007：263 − 266.（HAN J W, KAMBER M. Data Mining: Concepts and Techniques [M]. FAN M, MENG X F, translated. 2nd ed. Beijing: China Machine Press, 2007: 263 − 266.）

[2] BRODER A, GARCIA-PUETO L, JOSIFOVSKI V, et al. Scalable K-means by ranked retrieval [C]// WSDM '14: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM, 2014: 233 − 242.

[3] MIYAHAR S, MIYANOTO S. A family of algorithms using spectral clustering and DBSCAN [C]// Proceedings of the 2014 IEEE International Conference on Granular Computing. Washington, DC: IEEE Computer Society, 2014: 196 − 200.

（下转第 2542 页）

2005,26(5):1250 – 1252.(LIU Y J. Study of network service quality evaluation system [J]. Computer Engineering and Design, 2005, 26(5): 1250 – 1252.)

[17] TMF. GB923 wireless service measurements handbook approved version 3.0 [S]. Suffolk: TMF, 2004: 2 – 13.

[18] ITU-T. E. 800 terms and definitions related to quality of service and network performance including dependability [S]. Geneva: ITU-T, 2008: 1 – 21.

[19] ITU-T. E. 803 quality of service parameters for supporting service aspects [S]. Geneva: ITU-T, 2011: 5 – 15.

[20] ITU-T. How to increase QoS/QoE of IP-based platform(s) to regionally agreed standards [S]. Geneva: ITU-T, 2013: 1 – 20.

[21] ITU-T. G. 1020 performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks[S]. Geneva: ITU-T, 2006: 1 – 17.

[22] ITU-T. E. 802 framework and methodologies for the determination and application of QoS parameters [S]. Geneva: ITU-T, 2007: 8 – 14.

[23] 刘路,刘志民,罗英姿.欧洲 MOOC 教育质量评价方法及启示[J].开放教育研究,2015,21(5):57 – 65.(LIU L, LIU Z M, LUO Y Z. Analysis of three European assessment tools for measuring MOOCs' quality in higher education and its implications [J]. Open Education Research, 2015, 21(5): 57 – 65.)

[24] 肖静华,宛小伟,谢康.企业人力资源管理质量评价模型及实证分析[J].管理评论,2011,23(8):143 – 151.(XIAO J H, WAN X W, XIE K. Human resource management quality evaluation model and empirical analysis[J]. Management Review, 2011, 23(8): 143 – 151.)

[25] 杨阳,汪海涛,姜瑛,等.基于模糊层次分析法的软件质量评价模型的研究[J].计算机与数字工程,2017,45(4):620 – 623.(YANG Y, WANG H T, JIANG Y, et al. Software quality evaluation model based on fuzzy analytic hierarchy process [J]. Computer and Digital Engineering, 2017, 45(4): 620 – 623.)

[26] 谭跃进.定量分析方法[M].北京:中国人民大学出版社,2012: 103 – 105.(TAN Y J. Method of Quantitative Analysis [M]. Beijing: China Renmin University Press, 2012: 103 – 105.)

[27] 张潮.移动通信网络管理系统质量评价体系和方法的研究与应用[D].北京:北京邮电大学,2011:20 – 22.(ZHANG C. Study and application of mobile network management quality evaluation system and method [D]. Beijing: Beijing University of Posts and Telecommunications, 2011:20 – 22.)

[28] 张旭.移动运营商运营质量评估系统的研究[D].北京:北京交通大学,2012:4 – 5.(ZHANG X. Research on the mobile operator's operation quality assessment system [D]. Beijing: Beijing Jiaotong University, 2012:4 – 5.)

[29] BRANDES U. A faster algorithm for betweenness centrality [J]. Journal of Mathematical Sociology, 2001, 25(2): 163 – 177.

This work is partially supported by the Natural Science Foundation of Jiangsu Province (BK20171447), the Natural Science Foundation for Colleges and Universities of Jiangsu Province (17JKB520024), the Ministry of Education-China Mobile Research Fund (MCM20150510), the Introduction of Talent Research Start-up Foundation of Nanjing University of Posts and Telecommunications (NY215045).

WU Muyang, born in 1995, M. S. candidate. Her research interests include log analysis, data mining.

LIU Zheng, born in 1980, Ph. D., lecturer. His research interests include graph data mining and query, network data mining.

WANG Yang, born in 1983, Ph. D., senior engineer. His research interests include mobile communication big data analysis, centralized monitoring, network management system.

LI Yun, born in 1975, Ph. D., professor. His research interests include machine learning, data mining, distributed computing, pattern recognition.

LI Tao, born in 1975, Ph. D., professor. His research interests include data mining, machine learning, information retrieval.

www.joca.cn

(上接第 2514 页)

[4] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492 – 1496.

[5] 邱保志,唐雅敏.快速识别密度骨架的聚类算法[J].计算机应用,2017,37(12):3482 – 3486.(QIU B Z, TANG Y M. Efficient clustering algorithm for fast recognition of density backbone [J]. Journal of Computer Application, 2017, 37(12): 3482 – 3486.)

[6] 蒋礼青,张明新,郑金龙,等.快速搜索与发现密度峰值聚类算法的优化研究[J].计算机应用研究,2016,33(11):3251 – 3254.(JIANG L Q, ZHANG M X, ZHENG J L, et al. Optimization of clustering by fast search and find of density peaks [J]. Application Research of Computers, 2016, 33(11): 3251 – 3254.)

[7] 马春来,单洪,马涛.一种基于簇中心点自动选择策略的密度峰值聚类算法[J].计算机科学,2016,43(7):255 – 258.(MA C L, SHAN H, MA T. Improved density peaks based clustering algorithm with strategy choosing cluster center automatically [J]. Computer Science, 2016, 43(7): 255 – 258.)

[8] BIE R, MEHMOOD R, RUAN S S, et al. Adaptive fuzzy clustering by fast search and find of density peaks [J]. Personal and Ubiquitous Computing, 2016, 20 (5): 785 – 793.

[9] XU J, WANG G, DEND W. DenPEHC: density peak based efficient hierarchical clustering [J]. Information Sciences, 2016, 373 (12): 200 – 218.

[10] YANG X H, ZHU Q P, HUANG Y J, et al. Parameter-free Laplacian centrality peaks clustering [EB/OL]. [2017-12-05]. https://www.researchgate.net/publication/320602767_Parameter-free_Laplacian_centrality_peaks_clustering.

[11] QI X, FULLER E, WU Q, et al. Laplacian centrality: a new centrality measure for weighted networks [J]. Information Sciences, 2012, 194(5): 240 – 253.

[12] YE X, LI D, HE X. An algorithm for automatic recognition of cluster centers based on local density clustering [C]// Proceedings of the 2017 29th Chinese Control and Decision Conference. Piscataway, NJ: IEEE, 2017: 1347 – 1351.

[13] STEIN C M. Estimation of the mean of a multivariate normal distribution [J]. Annals of Statistics, 1981, 9(6): 1135 – 1151.

This work is partially supported by the Basic and Advanced Technology Research Project of Henan Province (152300410191).

QIU Baozhi, born in 1964, Ph. D., professor. His research interests include data mining, artificial intelligence.

CHENG Luan, born in 1992, M. S. candidate. Her research interests include data mining.