Journal of Computer Applications ISSN 1001-9081 2019-02-10
计算机应用,2019,39(2):398-402 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2019)02-0398-05 DOI:10.11772/j.issn.1001-9081.2018061411

基于密度峰值与密度聚类的集成算法

王治和,黄梦莹*,杜 辉,秦红武
(西北师范大学 计算机科学与工程学院 兰州 730070)
(* 通信作者电子邮箱 1126380876@qq.com)

摘 要:针对快速搜索和发现密度峰值聚类(CFSFDP)算法需人工在决策图上选择聚类中心的问题,提出一种基于密度峰值和密度聚类的集成算法。首先,借鉴 CFSFDP 思想,将局部密度最大的数据作为第一个中心;接着,从该中心点出发采用一种利用 Warshall 算法求解密度相连改进的基于密度的噪声应用空间聚类(DBSCAN)算法进行聚类,得到第一个簇;最后,在尚未被划分的数据中找出最大局部密度的数据,将它作为下一个簇的中心后再次采用上述算法进行聚类,直到所有数据被聚类或有部分数据被视为噪声。所提算法既解决了 CFSFDP 选择中心需人工干预的问题,又优化了 DBSCAN 算法,即每次迭代都是从当前最好的点(局部密度最大的点)出发寻找簇。通过可视化数据集和非可视化数据集与经典算法(CFSFDP、DBSCAN、模糊 C 均值(FCM)算法和 K 均值(K-means)算法)的对比实验结果表明,所提算法聚类效果更好,准确率更高,优于对比算法。

关键词:密度峰值;密度聚类;Warshall 算法;决策图;聚类中心
中图分类号:TP301.6 **文献标志码**:A

Integrated algorithm based on density peaks and density-based clustering

WANG Zhihe, HUANG Mengying*, DU Hui, QIN Hongwu
(*College of Computer Science and Engineering, Northwest Normal University, Lanzhou Gansu* 730070, *China*)

Abstract: In order to solve the problem that Clustering by Fast Search and Find of Density Peaks (CFSFDP) needs to manually select the center on the decision graph, an Integrated Algorithm Based on Density Peaks and Density-based Clustering (IABDPDC) was proposed. Firstly, learning from the principle of CFSFDP, the data with the largest local density was selected as the first center. Then, from the first center, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm improved by Warshall algorithm was used to cluster to obtain the first category. Finally, from the data that has not been clustered, the maximum local density data was found out as the center of the next category and was clustered again by the above algorithm, until all the data was clustered or some data was considered as noise. The proposed algorithm not only solves the problem of manual center selection in CFSFDP, but also optimizes the DBSCAN algorithm, in which, every iteration starts from the current best point (the point with the largest local density). By comparing with the classical algorithms (such as CFSFDP, DBSCAN, fuzzy C-means (FCM) and K-means) on visual datasets and non-visualized datasets, the experimental results show that the proposed algorithm has better clustering effect with higher accuracy.

Key words: density peak; density-based clustering; Warshall algorithm; decision graph; cluster center

0 引言

聚类分析试图将相似的元素划分为一类、不相似的元素划分在不同类,广泛应用于天文学、生物信息学、文献计量学和模式识别[1-4]等领域。聚类分析是数据挖掘的重要模块之一,其算法主要有基于划分的算法、基于图论的算法、基于密度的算法等。在基于划分的算法中,最具代表性的是 K-means[5]和 K-medoids[6],所划分的类是数据到中心距离更小的那些数据的集合。这类算法都是先设定中心,然后通过计算目标函数和数据间距离以及不断优化中心,直到最适合的中心被找到[7];但是,这种聚类算法只是把数据分配给距离它更近的中心,只适应于凸数据。谢娟英等[8]提出了一种基于密度峰值初始化中心的 K-medoids 算法,能自动确定类的个数。在基于图论的算法中,多路谱聚类算法[9]要求对数据建立相似度矩阵和拉普拉斯矩阵,计算特征值和特征向量,然后利用 K-means 算法进行聚类;但是该算法的时间复杂度和空间复杂度都比较高,还需要进一步的改进。周林等[10]利用采样算法降低了谱聚类算法的计算复杂度,利用 Nystrōm 采样算法只计算随机采样数据点之间以及随机采样数据点与剩余数据点之间的相似度矩阵。在基于密度的聚类算法中,基于密度的噪声应用空间聚类(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)算法[11]通过建立数据的密度相连实现聚类,这种方法能发现任意形状的簇,但是对阈值 *Eps*(扫描半径)和 *MinPts*(最小包含点数)的依赖较大。Chen 等[12]针对聚类中心测量困难和参数依赖性大的问题提出了一种新的聚类中心快速确定的聚类算法。戴阳

收稿日期:2018-07-10;修回日期:2018-09-06;录用日期:2018-09-11。 基金项目:国家自然科学基金资助项目(61662068)。
作者简介:王治和(1965—),男,甘肃武威人,教授,硕士,主要研究方向:数据挖掘; 黄梦莹(1990—),女,河南许昌人,硕士研究生,CCF 会员,主要研究方向:数据挖掘、机器学习; 杜辉(1976—),女,甘肃兰州人,副教授,博士,主要研究方向:数据挖掘、智能计算; 秦红武(1978—),男,甘肃兰州人,副教授,博士,主要研究方向:大数据、社会计算。

阳等[13]提出了初始点优化与参数自适应的 DBSCAN 算法,解决了阈值对密度不均匀数据聚类影响的问题。快速搜索和发现密度峰值聚类(Clustering by Fast Search and Find of Density Peaks, CFSFDP) 算法[14]是基于局部密度和相对距离的算法,利用决策图人工识别中心进行聚类,运算效率高,但并不是所有的数据集都能通过决策图准确地找到中心并进行聚类。因此,利用 CFSFDP 算法寻找中心的方法不断地被学者们研究,提出了各种各样的方法。比如,文献[15]提出了一种自动确定中心的 CFSFDP 算法,是基于基尼指数的自适应截断距离和自动获取聚类中心的方法,避免了决策图人工选择中心带来的误差。马春来等[16]提出了一种自动选择中心的密度峰值算法,根据中心权值的变化趋势选择“拐点”,以“拐点”之前的一组数据作为中心,避免了决策图人工找中心的误差。周世波等[17]提出了一种基于决策图和相对密度的聚类算法,实现了快速寻找聚类中心并确定有效的类数。谢国伟等[18]提出了一种非参数核估计的 CFSFDP 算法,用非参数核估计的方法计算数据的局部密度并选择潜在中心进行归类,最后对相邻类合并得到聚类结果。

通过上述研究,本文针对 CFSFDP 需要人工从决策图上选择中心的问题,提出了一种基于 CFSFDP 和 DBSCAN 的集成算法(Integrated Algorithm Based on Density Peaks and Density-based Clustering, IABDPDC) 。在找到中心之后,对数据集的聚类设计了一种利用传递闭包的 Warshall 算法[19]求解密度相连的密度聚类,以便每次聚类都是从最优点出发,而不是从传统密度聚类的核心点出发。IABDPDC 既解决了 CFSFDP 在确定中心时需人工在决策图上选择的问题,又优化了 DBSCAN 算法。

1　相关工作

CFSFDP 算法和 DBSCAN 算法都需要计算数据的局部密度,其中: CFSFDP 算法是一种快速、高效的聚类算法,但决策图需人工选择中心; DBSCAN 算法相比 K-means 算法,能对非凸数据聚类,但算法效率不高。Warshall 算法则是一种计算相似关系的算法,本文拟利用 Warshall 算法实现 DBSCAN 算法的密度相连,以优化 DBSCAN 算法。

1.1　CFSFDP 算法

CFSFDP 算法是一种快速搜索查询的密度峰值算法,它认为局部密度较高的数据被局部密度比它低的数据包围,即中心的局部密度是这一类中局部密度最大的数据,通过决策图人工选择局部密度更高且距离更大的数据作为聚类中心。对于每一个数据 i,该算法需要计算局部密度 ρ_i 和相对距离 δ_i,具体见式(1) 和式(2) 。

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

其中: 如果 $\chi(a) = 1$,当且仅当 $a < 0$; 如果 $a \geqslant 0$,$\chi(a) = 0$。d_c 为截断距离,d_{ij} 是数据 i 和 j 之间的欧氏距离。由此可知,ρ_i 是数据点 i 与其他数据点的距离中比 d_c 小的点的个数,称为数据点 i 的局部密度。

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij} - d_c) \tag{2}$$

由式(2) 可知,数据 i 的相对距离 δ_i 是它与比它局部密度大的那些数据间距离的最小值。

CFSFDP 算法认为 δ_i 和 ρ_i 值都较大的数据 i 是聚类中心,将其余数据划分到比它的局部密度大且最近邻的中心所属的类别。通常,CFSFDP 的第一个中心是局部密度最大的数据,剩余中心很难抉择,易造成选择的中心不是被局部密度比它低的数据包围。

1.2　DBSCAN 算法

DBSCAN 算法是一种基于密度聚类的典型算法,可以实现对任意数据的聚类并识别噪声,在 *Eps* 邻域、核心点、直接密度可达、密度可达的基础上,建立密度相连[20],通过实现最大密度相连的集合对数据集进行聚类。DBSCAN 算法需要识别数据集的噪声点、边界点和核心点,然后进行聚类。识别噪声点之后,将所有在核心点 *Eps* 邻域内的数据与该核心点划分为一簇,并对簇中未建立 *Eps* 邻域的核心点建立 *Eps* 邻域,建立最大密度相连集合,实现对簇的划分。

DBSCAN 算法中的密度相连也是一种数据间的传递关系,为使其密度相连过程更加简单,设计一种改进的 DBSCAN 算法,即用 Warshall 算法求解传递闭包的过程代替 DBSCAN 算法的求解最大密度相连的过程。基于 Warshall 的 DBSCAN 算法在利用局部密度最大的数据是中心的想法找到中心后,再从中心点出发查找所有和中心点同一类的数据进行聚类,不需要识别数据集的噪声点、边界点和核心点,降低了程序的复杂性。

1.3　Warshall 算法

Warshall[19]于 1962 年提出了求关系传递闭包的算法,并将它命名为 Warshall 算法,因其使复杂的二元关系的传递变得简单化而被人广泛学习和应用。该算法通过分析数据间的相似关系,求出相似关系的传递闭包。Warshall 算法如下:

步骤 1　置矩阵 $\boldsymbol{M}$。

步骤 2　置 $a = 1$。

步骤 3　对所有的 b,如果 $M[b,a] = 1$,则对 $c = 1,2,3,\cdots,n$,$M[b,c] := M[b,c] + M[a,c]$。

步骤 4　a 加 1。

步骤 5　如果 $a \leqslant n$,则转到步骤 3; 否则,算法结束。

最后所得矩阵是矩阵 $\boldsymbol{M}$ 的关系矩阵。Warshall 算法中,可达矩阵利用的是逻辑加运算,如果 $M[b,c] = 0$,$M[a,c] = 0$,则 $M[b,c] = 0$; 否则,$M[b,c] = 1$。

2　IABDPDC

CFSFDP 算法确定 δ_i 和 ρ_i 值都较大的数据 i 需要人工在决策图(如图 1) 上选择,第一个中心就是最大局部密度的数据,它在决策图的最右上方,第二或其他的中心为决策图中 δ_i 和 ρ_i 值相对都较大的数据,也分布在决策图的右上方,但应该选择右上方的哪些数据作为中心是不好解决的问题。例如图 1,最右上方的数据 1 的局部密度是最大的,将它作为第一个中心,数据 2、3、4 的 δ_i 和 ρ_i 值相差不大,选择哪个数据作为下一个中心是很难抉择的问题。尽管有时通过计算 $\gamma_i = \delta_i\rho_i$ 选择 γ_i 值较大的数据作为中心,但选取多少个 γ_i 较大的数据点作为中心是不容易确定的。基于 CFSFDP 与 DBSCAN 的集成算法,可以准确地找出类簇的中心,避免了中心选择不当造

成的聚类错误。

从上述寻找到的中心(最好的点)出发利用密度聚类算法进行聚类,然而,传统的密度聚类无法从最好的点出发聚类,而是需要找数据的核心点、直接密度可达点、密度可达点,进而找出数据的最大密度相连对数据进行聚类,聚类过程比较复杂,给程序造成了负担。利用 Warshall 算法求解 DBSCAN 算法的最大密度相连,从聚类中心出发利用 Warshall 算法找出所有和该点相连的数据并将它们划分为同一类降低了算法的复杂性。

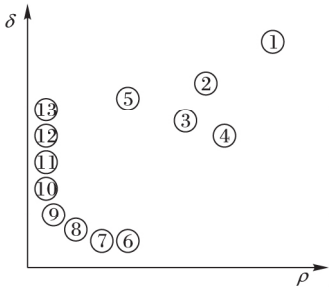


图 1 决策图

Fig. 1 Decision graph

针对上述问题,提出一种基于 CFSFDP 和 DBSCAN 的集成算法 IABDPDC,每次从最大局部密度数据出发用改进的 DBSCAN 算法进行聚类。IABDPDC 的主要步骤如下:

步骤 1 输入数据集。

步骤 2 利用式(1)计算所有数据的局部密度。

步骤 3 从剩余没有被聚类的数据出发选出最大局部密度的数据作为中心。

步骤 4 通过改进的密度聚类将所有和该中心同一类的数据划分出来。

步骤 5 执行步骤 3。

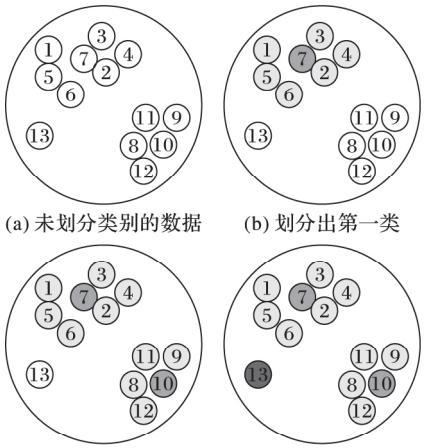
步骤 6 用步骤 4 的方法对剩余的数据聚类。

步骤 7 以此类推,直到所有数据被聚类或部分未被聚类的数据作噪声处理。

如图 2 所示,首先计算数据点之间的距离,然后利用式(1)计算所有数据的局部密度,如图 2(a)所示,为一组未划分类别的数据集;从未划分类别的数据出发计算局部密度,得到最大局部密度的数据 7,将它作为第一个中心,然后利用改进的 DBSCAN 算法将和该中心同属一类的数据划分出来,将数据 1、2、3、4、5、6、7 划分为一类,聚类结果如图 2(b)所示;剩余没有被划分类别的数据分别是 8、9、10、11、12、13,再次计算它们的局部密度,得到最大局部密度的数据 10,将它作为下一个中心,然后再利用改进的 DBSCAN 算法聚类,聚类结果如图 2(c)所示;判断是否还有数据没有被聚类,如果有,则继续对剩余的数据进行上述操作,直到所有数据被划分类别或有部分数据作噪声处理,如图 2(d)所示,没有被聚类的数据被视为噪声。

3 实验结果与分析

IABDPDC 通过 C 语言编写,用 Matlab 工具显示实验结果。通过对可视化数据集和非可视化数据集进行实验,从而评估和分析所提算法,对比算法包括: CFSFDP、DBSCAN、FCM 和 K-means 算法。其中: CFSFDP 算法是一种快速搜索查询的利用决策图确定中心的算法; DBSCAN 算法是基于密度的典型聚类算法; FCM 算法[21]是一种基于拉普拉斯矩阵的聚类算法; K-means 是一种只适应于凸数据的聚类算法; IABDPDC 则是一种基于密度分析的聚类算法,不仅可以自动确定中心,且利用 Warshall 算法建立矩阵实现了对任意形状数据的聚类。



(a) 未划分类别的数据 (b) 划分出第一类

(c) 划分出第二类 (d) 完成聚类

图 2 IABDPDC 应用示例

Fig. 2 IABDPDC application example

3.1 可视化数据集实验结果分析

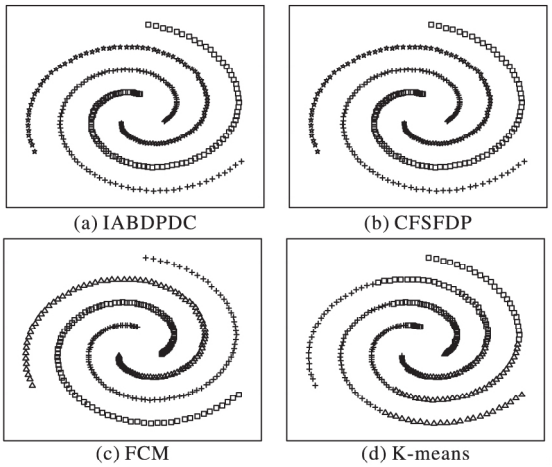
利用四组二维人造数据进行实验,实验数据集见表 1,聚类结果如图 3~6 所示。

图 3 是四种聚类算法在 Spiral 数据集上的聚类结果,可以看出,K-means 聚类效果不理想,IABDPDC、FCM 和 CFSFDP 都能将数据正确地划分类别。

表 1 可视化实验数据集

Tab. 1 Visual experiment datasets

数据集	实例数	维数	分类数
Spiral	312	2	2
Lineblobs	266	2	3
Path-based1	300	2	3
Aggregation	788	2	7



(a) IABDPDC (b) CFSFDP

(c) FCM (d) K-means

图 3 四种聚类算法对数据集 Spiral 聚类的结果

Fig. 3 Clustering results of four clustering algorithms on dataset Spiral

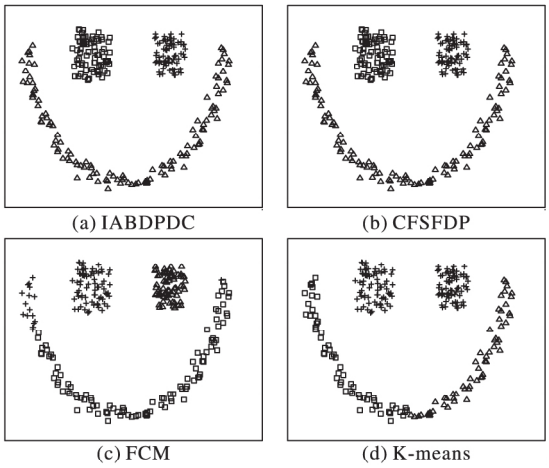


图 4　四种聚类算法对数据集 Lineblobs 聚类结果

Fig. 4　Clustering results of four clustering algorithms on dataset Lineblobs

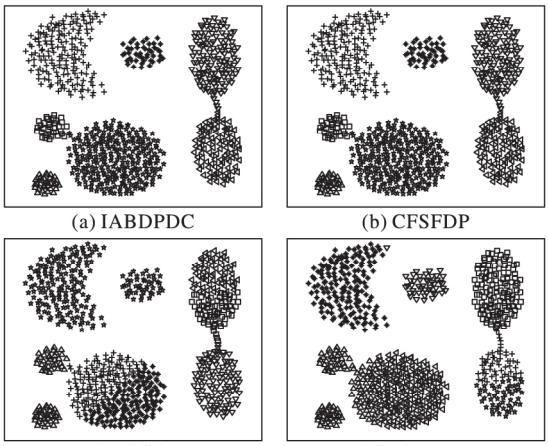


图 5　四种聚类算法对数据集 Aggregation 聚类结果

Fig. 5　Clustering results of four clustering algorithms on dataset Aggregation

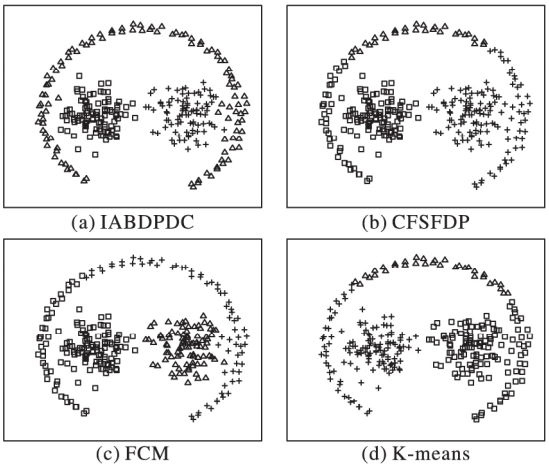


图 6　四种聚类算法对数据集 Path-based1 聚类结果

Fig. 6　Clustering results of four clustering algorithms on dataset Path-based1

图 4 是四种聚类算法在 Lineblobs 数据集上的聚类结果，可以看出，FCM 和 K-means 算法聚类效果不理想，IABDPDC 和 CFSFDP 算法聚类效果较好。

图 5 是四种聚类算法在 Aggregation 数据集上的聚类结果，可以看出，FCM 算法和 K-means 算法聚类效果不理想，IABDPDC 和 CFSFDP 算法将数据正确地划分了类别。

图 6 是四种聚类算法在 Path-based1 数据集上的聚类结果，可以看出，CFSFDP、FCM 和 K-means 的聚类效果都不理想，IABDPDC 聚类效果理想。

通过图 3 ~ 6 四组实验的可视化对比说明 IABDPDC 比 FCM、CFSFDP 和 K-means 算法聚类效果好。

3.2　非可视化数据实验结果分析

为了测试 IABDPDC 的性能，随机从 UCI 数据库中选取五组数据集分别用 IABDPDC、DBSCAN、CFSFDP、FCM 和 K-means算法进行实验，采用评价指标 ACC (accuracy) [23] 用于评价算法的正确率，计算公式如式(3) 所示，实验数据集见表 2，聚类结果见表 3。

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\delta(\hat{C}_i, map(C_i)) \tag{3}$$

其中: C_i 是所提算法的类标签; $\hat{C}_i$ 是数据真实的类标签; $\delta(x, y)$ 表示函数，当 $x = y$ 时 $\delta(x,y)$ 的值取 1，否则取 0; $map(x)$ 作为最好的映射函数使用了匈牙利算法进行映射，对获得的的中心和真实的中心进行映射。ACC 的值越大表示聚类结果越正确。

表 2　非可视化实验数据集

Tab. 2　Non-visualized experimental datasets

数据集	实例数	维数	分类数
Iris	150	4	3
Wine	178	13	3
Cmc	1 473	9	3
Seeds	210	7	3
Tae	151	5	3

表 3　五种算法在 UCI 数据库的聚类准确率对比

Tab. 3　Clustering accuracy comparison of five algorithms on UCI dataset

数据集	IABDPDC	DBSCAN	CFSFDP	FCM	K-means
Iris	0.96	0.96	0.86	0.89	0.78
Wine	0.92	0.88	0.88	0.82	0.70
Cmc	0.52	0.56	0.65	0.46	0.35
Seeds	0.83	0.86	0.86	0.87	0.74
Tae	0.71	0.62	0.55	0.47	0.36

表 3 是五种聚类算法在五个数据集上的聚类结果，表中加粗字体表示聚类效果最好，可以看出，数据集 Iris、Wine 和 Tae 在 IABDPDC 算法上的聚类效果最好，而数据集集 Seeds 的聚类效果次之，数据集 Cmc 的聚类结果不好是因为该数据集的 9 个属性里的一个属性值偏大对整体聚类结果产生了较大的影响。通过上述分析可以得出，IABDPDC 聚类效果更好，聚类准确率优于对比算法。

4　结语

针对 CFSFDP 算法确定中心需要人工在决策图上选择的问题，提出了一种集成算法，将 CFSFDP 的思想(局部密度最大的数据是中心) 和 DBSCAN 算法结合为一种新的算法 IABDPDC。为降低算法的复杂性，还将 DBSCAN 算法求密度相连的过程设计为利用二元关系传递闭包的 Warshall 算法来

求解。IABDPDC 每次从未被聚类的数据出发将局部密度最大的数据作为中心，利用改进的 DBSCAN 算法从最优点出发进行聚类，既解决了 CFSFDP 算法选择中心需人工干预的问题，又优化了 DBSCAN 算法。实验对比与分析表明，IABDPDC 能得到较好的聚类结果，但部分聚类结果仍受到阈值的影响，下一步将研究自适应的聚类算法，减少阈值对聚类结果的影响，使算法的聚类结果达到最优。

参考文献：

[1] HE H, TAN Y. Automatic pattern recognition of ECG signals using entropy-based adaptive dimensionality reduction and clustering [J]. Applied Soft Computing, 2017, 55: 238 - 252

[2] PENG C, KANG Z, XU F, et al. Image projection ridge regression for subspace clustering [J]. IEEE Signal Processing Letters, 2017, 24(7): 991 - 995.

[3] R? THLISBERGER V, ZISCHG A P, KEILER M. Identifying spatial cluster of flood exposure to support decision making in risk management [J]. Science of the Total Environment, 2017, 598: 593 - 603.

[4] SUZUKI S, KAKUTA M, ISHIDA T, et al. Faster sequence homology searches by clustering subsequences [J]. Bioinformatics, 2015, 31(8): 1183 - 1190.

[5] MACQUEEN J B. Some methods for classification and analysis of multivariate observations [C] // Proceedings of the 1967 5th Berkeley Symposium on Mathematical Statistics and Probability. Berlin: Springer, 1967: 281 - 297.

[6] KAUFMAN L. ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis [M]. New York: John Wiley & Sons, 1990: 74.

[7] HOPPNER F, KLAWONN F, KRUSE R, et al. Fuzzy cluster analysis-methods for classification [J]. Journal of the Operational Research Society, 1998, 51(6): 769 - 770.

[8] 谢娟英，屈亚楠. 密度峰值优化初始中心的 K-medoids 聚类算法 [J]. 计算机科学与探索，2016，10(2)：230 - 247. (XIE J Y, QU Y N. K-medoids clustering algorithms with optimized initial seeds by density peaks [J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(2): 230 - 247.)

[9] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C] // Proceedings of the 2001 International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA: MIT Press, 2001: 849 - 856.

[10] 周林，平西建，徐森，等. 基于谱聚类的聚类集成算法 [J]. 自动化学报，2012，38(8)：1335 - 1342. (ZHOU L, PING X J, XU S, et al. Cluster ensemble based on spectral clustering [J]. Acta Automatica Sinica, 2012, 38(8): 1335 - 1342.)

[11] ESTER M, KRIEGEL H-P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996: 226 - 231.

[12] CHEN J, LIN X, ZHENG H, et al. A novel cluster center fast determination clustering algorithm [J]. Applied Soft Computing, 2017, 57: 539 - 555.

[13] 戴阳阳，李朝锋，徐华，等. 初始点优化与参数自适应的密度聚类算法 [J]. 计算机工程，2016，42(1)：203 - 209. (DAI Y Y, LI C F, XU H, et al. Density spatial clustering algorithm with initial point optimization and parameter self-adaption [J]. Computer Engineering, 2016, 42(1): 203 - 209.)

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

[15] 王洋，张桂珠. 自动确定聚类中心的密度峰值算法 [J]. 计算机工程与应用，2017，2018，54(8)：137 - 141. (WANG Y, ZHANG G Z. Automatically determine density of cluster center of peak algorithm [J]. Computer Engineering and Applications, 2017, 2018, 54(8): 137 - 141.)

[16] 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.)

[17] 周世波，徐维祥. 一种基于相对密度和决策图的聚类算法 [J]. 控制与决策，2018，33(11)：1921 - 1930. (ZHOU S B, XU W X. A novel clustering algorithm based on relative density and decision graph [J]. Control and Decision, 2018, 33(11): 1921 - 1930.)

[18] 谢国伟，钱雪忠，周世兵. 基于非参数核密度估计的密度峰值聚类算法 [J/OL]. 计算机应用研究，2018 [2018-04-06]. http://www.arocmag.com/article/02-2018-10-018.html. (XIE G W, QIAN X Z, ZHOU S B. Density peak clustering algorithm based on non-parametric kernel density estimation [J]. Application Research of Computers, 2018 [2018-04-06]. http://www.arocmag.com/article/02-2018-10-018.html.)

[19] WARSHALL S. A theorem on Boolean matrices [J]. Journal of the ACM, 1962, 9(1): 11 - 12.

[20] 刘淑芬，孟冬雪，王晓燕. 基于网格单元的 DBSCAN 算法 [J]. 吉林大学学报(工学版)，2014，44(4)：1135 - 1139. (LIU S F, MENG D X, WANG X Y. DBSCAN algorithm based on grid cell [J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(4): 1135 - 1139.)

[21] PAL N R, BEZDEK J C. On cluster validity for the fuzzy c-means model [J]. IEEE Transactions on Fuzzy Systems, 1995, 3(3): 370 - 379.

[22] PAPADIMITRIOU C H, STEIGLITZ K. Combinatorial optimization: algorithms and complexity [M]. Upper Saddle River, NJ: Prentice-Hall, 1982.

This work is partially supported by the National Natural Science Foundation of China (61662068).

WANG Zhihe, born in 1965, M. S., professor. His research interests include data mining.

HUANG Mengying, born in 1990, M. S. candidate. Her research interests include data mining, machine learning.

DU Hui, born in 1976, Ph. D., associate professor. Her research interests include data mining, intelligent computing.

QIN Hongwu, born in 1978, Ph. D., associate professor. His research interests include big data, social computing.