网络出版时间：2017-05-10 10:49:45
网络出版地址：http://kns.cnki.net/kcms/detail/11.2127.TP.20170510.1049.004.html

基于数据场的类簇中心选取及其聚类

朱振国，冯应柱

ZHU Zhenguo, FENG Yingzhu

重庆交通大学 信息科学与工程学院，重庆 400074
College of information science and engineering, Chongqing Jiaotong University, Chongqing 400074, China

ZHU Zhenguo, FENG Yingzhu. Clustering center selection and clustering based on data field. Computer Engineering and Applications.

Abstract: In view of the existing clustering algorithms are widespread low clustering quality, parameter dependency and outlier effects obvious, in this paper, a clustering method based on the field data is proposed. The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local potential and that they are at a relatively large distance from any points with a higher local potential. According to the characteristics of the potential value of isolated point is equal to zero, remove the outlier and finally the other object points into larger than its potential value and nearest neighbor type of clusters, so as to achieve clustering. Simulation results show that the proposed algorithm is effective and has no effect on the shape of the data set, and it can find out clusters center and outliers accurately without artificial parameters.
Key words: Cluster center; Data field; Clustering; Outlier

摘　要：针对现有聚类算法普遍存在聚类质量低、参数依赖性大、孤立点难识别等问题，提出一种基于数据场的聚类算法。该算法通过计算每个数据对象点的势值，根据类簇中心的势值比周围邻居的势值大，且与其他类簇中心有相对较大距离的特点，确定类簇中心；根据孤立点的势值等于零的特点，选出孤立点；最后将其它数据对象点划分到比自身势值大且最近邻的类簇中，从而实现聚类。仿真实验表明，该算法在不需要人为调参的情况下准确找出类簇中心和孤立点，聚类效果优良，且与数据集的形状无关。
关键词：类簇中心；数据场；聚类；孤立点
文献标识码：A　　**中图分类号：**TP301.6　　doi：10.3778/j.issn.1002-8331.1611-0427

1 引言

所谓聚类，是指按照事物的某些属性，将事物聚集成若干类，使得类内相似度尽量大，而类间相似度尽量小[1]。通过聚类，人们能够识别事物分布的不同区域，发现数据属性间的相互关系，找到潜在有使用价值的信息，并为决策提供数据依据[2]。聚类分析已经成为当前非常重要的研究方向，广泛的应用在数据分析、模式识别、图像处理、市场研究以及生物学等领域[3]。

当前聚类算法主要有以下几类：基于划分的如 k-means[4]、k-mediods[5]、CLARANS[6]；基于层次的如 BIRCH[7]、Chameleon[8]；基于密度的如 DBSCAN[9]、DENCLUE[10]；基于网格的如 STING[11]、WaveCluster[12]；基于模型

基金项目：重庆市研究生科研创新项目（No.CYS15180）。
作者简介：朱振国(1972－)，男，博士研究生，副教授，主要研究方向为智能信息处理、数据挖掘、机器学习，E-mail：zhuzhg@qq.com；冯应柱(1991－)，男，硕士研究生，主要研究方向为数据挖掘、机器学习。

的如 EM[13]。这些聚类算法各有特点，为促进聚类分析的研究起到了巨大的作用，但也存在一些不足。k-means、k-mediods 需要用户事先确定聚类数目，否则聚类结果就不准确；CLARANS、BIRCH 等存在“球形偏见”，即当数据分布为球形形状时聚类质量优良；Chameleon、WaveCluster 对噪声数据敏感，孤立点的存在严重影响聚类质量；DBSCAN、DENCLUE 对数据维度的伸缩性差；STING 聚类质量差，准确度不高；EM 算法需要优化初始值，计算量大。Rodriguez 等提出聚类算法 DPC[14]不存在上述算法存在的问题，但聚类结果严重依赖于给出的经验阈值，不能进行广泛应用。本文提出了一种基于数据场的聚类算法，该算法利用数据场的势函数来计算对象间的紧密程度，从而实现基于数据驱动的、聚类质量不依赖经验阈值的自动聚类算法。仿真实验结果表明，该算法能准确找到聚类中心，去除孤立点，聚类效果优良。

2 数据场基础

借鉴物理学中场的思想，李德毅院士将物质粒子间的相互作用引入抽象的数域空间，提出数据场[15]。数据场把任意一个数据的状态看作是数域空间中所有数据共同作用的结果，从而把个体与整体的相互作用考虑到数据分析中[16]。

2.1 数据场

定义 1 给定 p 维空间，包含 n 个对象的数据集 $D=\{x_1,x_2,\cdots,x_n\}$ 及其产生的数据场，空间任一点 x 的势值为：

$$\varphi(x)=\sum_{i=1}^{n}\varphi_i(x)=\sum_{i=1}^{n}e^{-\left(\frac{\|x-x_i\|}{\sigma}\right)^2} \tag{1}$$

其中，$\|x-x_i\|$ 表示场点与对象之间的范数距离；σ 为影响因子；将上述势值计算公式称为势函数，可以发现其为高斯核函数。数据场有如下性质：

独立性：每个数据对象点都以自己为中心，独立的向外辐射能量而不受外界的影响，如图 1，对象 A 与 B 各自分别以自身为中心向四周辐射能量；

叠加性：每个数据对象点的势值等于该空间中各个数据点在该点产生的能量总和，如图 1，位置 1、2 处的势值分别为数据对象 A、B 在此处势值的叠加总和；

衰减性：势值随着距离的增加急剧下降，距离场源越近势值越大，反之，势值越小。

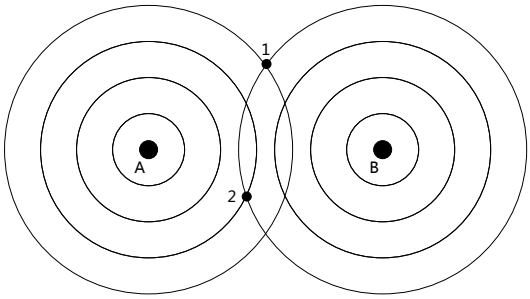

图 1 数据场性质示意图

2.2 数据场改进

考虑数据场中单个对象的作用范围，如图 2 所示。势值随着距离的增加而减小,σ 越小，势函数衰减速度越快；且 σ 越小，距离 R 越短，表示对象间的相互作用范围越小；

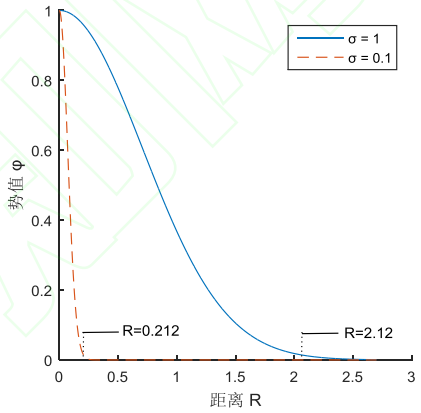

图 2 不同σ值的势函数及其影响半径

公式(1)为高斯核函数，由高斯函数的“3σ 规则”-在 $\pm3\sigma$ 范围内包含 99.73%的数据对象，可知：数据场中每个数据对象的影响范围是以该对象为中心、半径 R 等于 $\frac{3}{\sqrt{2}}\sigma\approx2.12\sigma$ 的邻域空间，即对象间的相互作用范围为 2.12σ。当任意两个对象之间的距离大于 2.12σ 时，相互之间的作用可以忽略不计。由此，改进的数据场公式如下：

$$\varphi(x)=\sum_{i=1}^{n}\varphi_i(x) \tag{2}$$

其中，

$$\varphi_i(x)=\begin{cases}e^{-\left(\frac{\|x-x_i\|}{\sigma}\right)^2}, & \|x-x_i\|\le R\\ 0, & \|x-x_i\|\le R\end{cases}$$

改进的数据场只考虑 2.12σ 邻域内所有

对象间的相互影响，较全局而言，局部势值更能体现数据分布的紧密程度，且减少计算量，降低了时间复杂度。改进后的数据场公式降低了孤立点的势值，使得孤立点更容易辨认。

3 一种基于数据场的聚类算法

本文提出一种基于数据场的聚类算法（clustering algorithm based on data field，DFC），该算法通过计算所有数据对象点的势值和到比它势值更大的数据对象点之间的最小距离，根据势值和距离的分布决策图，将势值较大且与比它势值更大的数据点有较大距离的数据对象点作为类簇中心。去除孤立点后，对剩余数据对象点按势值大小排序，将其划分到比自身势值大且最近邻的类簇中，得到最终的聚类结果。

算法流程描述如下：

Step 1：计算任意两个数据对象点之间的标准距离 d_{ij}；

Step 2：影响因子 σ 优化；

Step 3：计算数据对象点的势值 φ、距离 δ；

Step 4：确定类簇中心及孤立点；

Step 5：将其它数据对象点划分到最近邻的高势值点类簇中。

3.1 标准距离计算

为了解决原始数据存在着各维度量纲不一致，数据之间没有可比性，数据范围大容易造成大数吃小数等问题，本文采用 min-max 标准化，将原始数据转换成无量纲化指标，使得各指标值都处于同一个数量级别上，让数据之间具有可比性。min-max 标准化方法是将原始数据映射在区间[0,1]的一种线性变换。

定义 2 含 n 个对象 $D=\{x_1,x_2,\cdots,x_n\}$ 的数据集，标准化后的值如下：

$$x_i'=\frac{x_i-\min(D)}{\max(D)-\min(D)} \tag{3}$$

基于数据场的特点，可以用距离来衡量数据对象点之间的紧密程度。从聚类看，距离越小对象间的相似性越大，反之对象间的相似性就越小，本文采用欧式距离来度量对象间的相似性。

给定n维空间中任意两点 x_i,x_j，它们之间的欧式距离：$d_{ij}=\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2}$

d_{ij} 具有下述的三个属性：①非负性：$d_{ij}\geq 0$；②$d_{ij}=0\Leftrightarrow i=j$；③$d_{ij}=d_{ji}$

3.2 影响因子σ

σ 用于控制数据对象间的相互作用力程，其取值会严重影响数据场的空间分布。本文采用淦文燕等提出的基于最小势熵的 σ 优化算法[17]。

定义 3 设对象 x_i 的势值为 φ_i，则势熵为：

$$H=-\sum_{i=1}^{n}\frac{\varphi_i}{Z}\log\left(\frac{\varphi_i}{Z}\right) \tag{4}$$

其中 $Z=\sum_{i=1}^{n}\varphi_i$ 为一个标准化因子。

当势熵 H 取最小值时，对应的 σ 即为最优值。即：

$$\min_{\sigma\geq 0}H(\sigma)=\min_{\sigma\geq 0}-\sum_{i=1}^{n}\frac{\varphi_i}{Z}\log\left(\frac{\varphi_i}{Z}\right) \tag{5}$$

此为单变量非线性函数的最小化问题，本文采用初始区间为 $\left[\min_{i\neq j}\left\|x_i-x_j\right\|,\max_{i\neq j}\left\|x_i-x_j\right\|\right]$ 的黄金分割法优化 σ，算法时间复杂度为 $O\left(n^2\right)$。步骤如下：

（1）置 $a=\min_{i\neq j}\left\|x_i-x_j\right\|$，$b=\max_{i\neq j}\left\|x_i-x_j\right\|$，精度 ε；

（2）计算 $\sigma_1=a+(1-t)(b-a)$，$\sigma_2=a+t(b-a)$，$t=0.618$；

（3）若 $H(\sigma_1)-H(\sigma_2)>0$，转（4），否则转（5）；

（4）若 $b-a>\varepsilon$，$a=\sigma_1$，$\sigma_1=\sigma_2$，$\sigma_2=a+t(b-a)$，否则，停止计算输出 $\sigma=\sigma_2$；

（5）若 $b-a>\varepsilon$，$b=\sigma_2$，$\sigma_2=\sigma_1$，$\sigma_1=a+(1-t)(b-a)$，否则，停止计算输出 $\sigma=\sigma_1$。

3.3 类簇中心及孤立点

定义 4：对于包含 n 个对象的数据集，任一数据对象点 i 到比自身势值更大的对象

点之间的距离设为势值大于当前对象点势值中的距离的最小值。即：

$$\delta_i = \min_{j:\varphi_i>\varphi_j}\left(d_{ij}\right) \tag{6}$$

其中，对于具有最大势值的对象，将 δ_i 设置为距离的最大值 $\delta_i = \max_j\left(d_{ij}\right)$[18]。

定义类簇的中心是这样的一类点：它们被很多点围绕，导致局部势值大，且与局部势值比自己大的点之间的距离也较远，因此类簇中心是势值与距离都大的对象。如图 3（a）数据分布图所示，通过观察得知，数据点 1、6 为聚类中心，21、22 为孤立点。在图 3（b）决策图中，数据点 1、6 距离 δ 与势值 φ 都比较大，数据点 21、22 距离较大，但势值很小，其它数据点的距离都较小。

根据上述性质，文献[14]作者提供的代码中，聚类中心是在决策图上由人工选定的，具体方法是在决策图上框选某个区域，将区域内的点作为聚类中心，并将区域内点的数目作为聚类数目。本文也可采用上述方法，但没有自动选出最佳聚类数及聚类中心。由此提出一种方法自动确定最佳聚类数及聚类中心。令 $\tau_i = \varphi_i \times \delta_i$，在 τ_i 值计算之前，先分别对 φ_i 和 δ_i 用公式(3)做归一化处理，使得两者参与决策的权重相当。图 3(c)所示为将 τ_i 值从大到小排序得到的结果图，对于整体而言，τ_i 值是相近的，差异较大的就是那几个聚类中心。本文从异常检测的角度去寻找这些跳跃点，思想如下：从前往后依次计算相邻两个 τ 值的比值，直到连续两个比值都小于某个阈值时结束，则将该点前 k 个 τ 值对应的数据对象作为聚类中心，k 为最佳聚类数目。其数学形式为：

$$\frac{\tau_i}{\tau_{i+1}} < \mu, \mu \in (1,2)$$

对于非聚类中心点， 相邻 τ 值的比值趋近于 1， 而聚类中心点，相邻 τ 值的比值近似介于（2，10）之间。本文中 μ 值为 1.2，实验显示该取值是合理且普遍适用的。

推论 1：距离最大的对象一定是聚类中心；

证明：由定义 4，距离最大，则势值最大。即在对象 X 的 2.12σ 邻域内数据分布最密集，具有最大密度，故其为一个聚类中心。

推论 2:势值为零的对象一定是孤立点。

证明：设 $\varphi(x)=0$，由公式(2)可知，$\|x-x_i\|>2.12\sigma$，即在 X 的 2.12σ 领域内，无其他数据点存在，故其为孤立点。

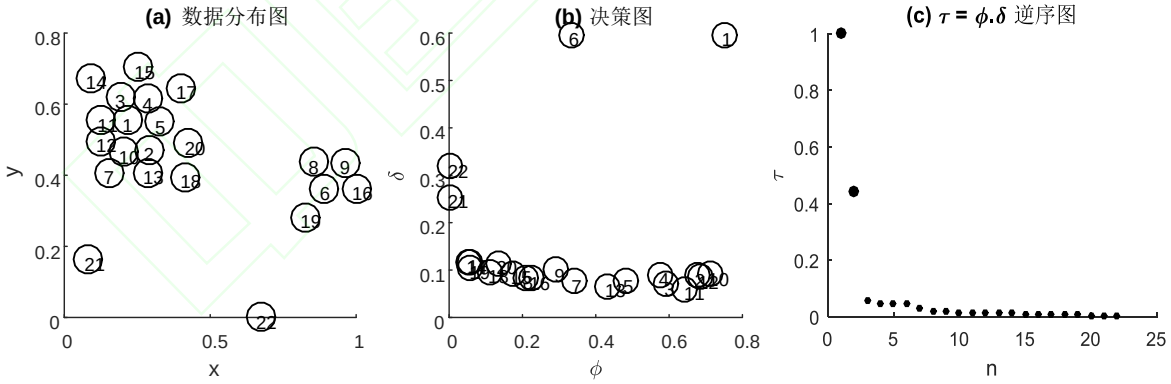

图 3　原始数据分布与决策关系图

4　仿真实验与结果分析

为比较本文 DFC 聚类算法与文献[14]中 DPC 聚类算法的差异，进行了仿真实验。实验中的操作系统为 Windows7，集成开发环境为 Matlab2015。

4.1　仿真实验

实验采用表 1 中的 4 个数据集进行测试，数据集来自文献[14]和文献[18]，4 个数据集的原始分布如图 4 所示。其聚类结果展示如图 5、图 6。

表 1　数据集

数据集名称	维数	类别数	数据量
Aggregation	2	7	788
Flame	2	2	240
Spiral	2	3	312
Jain	2	2	373

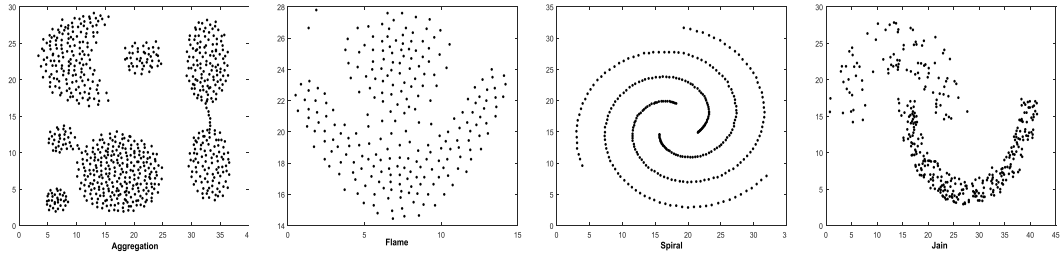

图 4　原始数据分布图

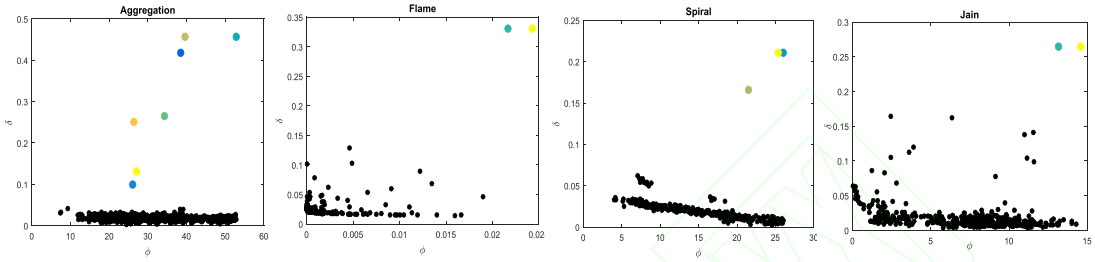

图 5　DFC 决策图

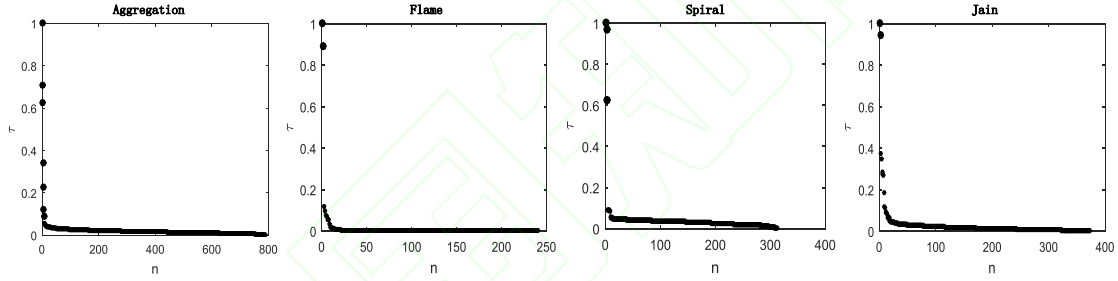

图 6　DFC $\tau = \varphi \cdot \delta$ 逆序图

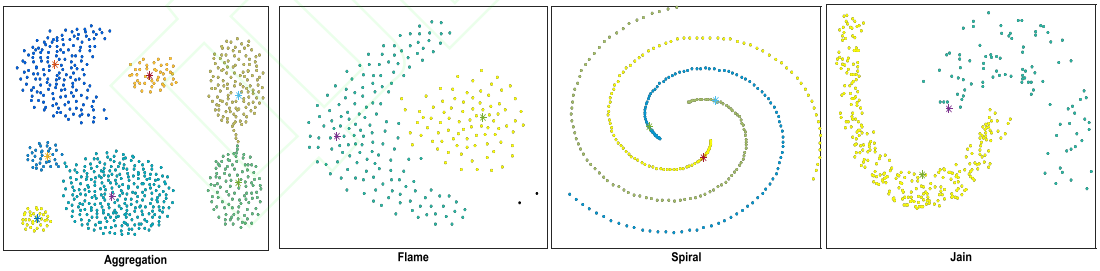

图 7　DFC 聚类结果分布图

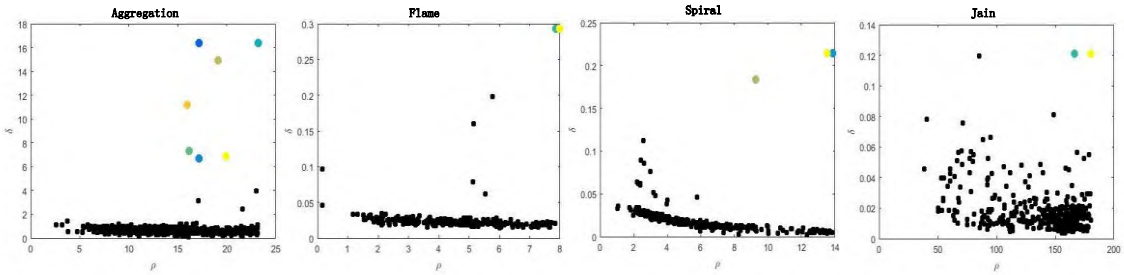

图 8　DPC 决策图

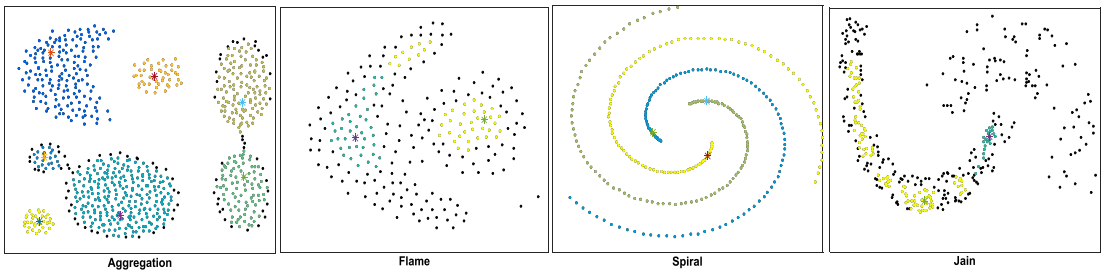

图 9　DPC 聚类结果分布图

表 2　孤立点数目

数据集名称	类别数	DFC 算法孤立点数目	DPC 算法孤立点数目
Aggregation	7	0，0，0，0，0，0，0	0，12，35，20，18，0，0
Flame	2	2，0	63，95
Spiral	3	0，0，0	0，0，0
Jain	2	0，0	128，125

图 5 为 DFC 算法在 4 个数据集中的类簇中心决策图，通过图 $6\tau=\varphi.\delta$ 逆序图找出最佳类簇中心及簇的数目后，最终聚类结果如图 7。图 8 为 DFC 算法在 4 个数据集中的类簇中心决策图，由人工确定类簇中心后聚类结果如图 9。由图 5 和图 8 对比得知，DFC 决策图发现的类簇中心明显优于 DPC 决策图，DPC 决策图中难以识别最佳类簇中心，需要人工经验来获取，而 DFC 决策图类簇中心与其他对象分离明显。再将图 7、图 9 与原始数据分布图 4 比较得知，DFC 算法识别的孤立点较准确，而 DPC 算法在数据集不存在孤立点的情况下错误的识别出孤立点，导致聚类质量低下，二者识别的孤立点数目如表 2 所示。故仿真实验表明，本文提出的 DFC 算法能够准确的找出任意形状数据集的类簇中心，检测出孤立点，具有良好的聚类质量。

4.2　实验结果分析

本文 DFC 算法实验结果优于文献[14]中的 DPC 算法，原因有二。其一：文献[14]对于每个点的密度 $\rho_i=\sum_j \chi\left(d_{ij}-d_c\right)$，其中 d_c 称为截断距离，其值取为数据集中样本数目的 1–2%，本文取 1.5%，然而这个经验值对于不同的数据集并不能普遍适用。本文引入数据场中的势函数并加以改进，考虑局部对象间的相互影响，根据势函数的叠加性和衰减性，求得每个数据对象点的势值，无人为输入参数，从根本上解决了经验阈值的问题，普遍适用于不同的数据集，因此计算出的每个点的密度更准确，找出的聚类中心更接近于实际。其二：文献[14]对于孤立点的判别是对每个簇取平均密度作为阈值，若该簇中数据点的密度小于平均密度则将其置为孤立点，此方法将密度较小的数据点都视作孤立点，增加了孤立点的数目。本文改进后的数据场，孤立点的势值为零，很好的解决了孤立点问题，使得聚类质量较高。

4.3　时间复杂度分析

假设由 n 个数据（样本）组成的数据集，则 DFC 算法的时间复杂度主要由计算每个数据对象点的势值、距离及影响因子 σ 构成，该过程的计算代价分别为 $O\left(n^2\right)$、$O\left(\left(n^2-n\right)/2\right)$ 与 $O\left(n^2\right)$，类簇中心确定后，算法只需经过一次划分就能完成聚类，与其它算法相比，本文算法的时间复杂度较高，主要消耗在线性探查寻找最小势熵的过程中。但是其优势在于能够在不需人为输入参数的情况下计算每个数据对象点的势值，找出类簇中心和孤立点，并且对于任意形态分布的数据集都能得到较高的聚类质量，因此在一定程度上可以弥补其时间复杂度较高的缺陷。

5 结束语

本文提出了一种基于数据场的聚类算法，具有较好的自适应性，能够处理任意形状的数据集。实验验证了本算法的可行性和有效性，获得比较好的聚类质量。文中影响因子 σ 的选取对实验结果影响较大，算法时间复杂度较高，接下来将重点研究如何优化算法降低时间复杂度，并进一步优化 σ 以达到更好的聚类质量。

参考文献：

[1] 郭锋. 基于数据场的聚类方法研究[D].哈尔滨工程大学,2009.

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1):48−61.

[3] 周涛, 陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用, 2012, 48(12):100−111.

[4] Zha H, He X, Ding C, et al. Spectral Relaxation for K-means Clustering[J]. Advances in Neural Information Processing Systems, 2002, 14:1057-1064.

[5] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2):3336-3341.

[6] Ng R T, Han J. A Method for Clustering Objects for Spatial Data Mining[J]. IEEE Transactions on Knowledge & Data Engineering, 2002, 14(5):1003-1016.

[7] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[J]. Acm Sigmod Record, 1999, 25(2):103-114.

[8] Karypis G, Han E H, Kumar V. Chameleon: Hierarchical Clustering Using Dynamic Modeling[J]. Computer, 1999, 32(8):68-75.

[9] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// 2008:226-231.

[10] YUAN, Hanning, WANG, et al. Feature Selection with Data Field[J]. Chinese Journal of Electronics, 2014, 23(4):661-665.

[11] Wang W, Yang J, Muntz R R. STING: A Statistical Information Grid Approach to Spatial Data Mining[C]//Vldb'97, Proceedings of, International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece. 1997:186-195.

[12] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases[C]// Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 2010:428-439.

[13] Mclachlan G J, Krishnan T. The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)[J]. Journal of Classification, 2007, 15(1):154-156.

[14] Rodriguez A, Laio A. Clustering by fast search and find of density peaks.[J]. Science, 2014, 344(6191):1492-1496.

[15] Wang S, Wang D, Caoyuan L I, et al. Clustering by Fast Search and Find of Density Peaks with Data Field[J]. Chinese Journal of Electronics, 2016, 25(3):397-402.

[16] 李德毅,杜鹢. 不确定性人工智能[M]. 国防工业出版社, 2004:197−212.

[17] 淦文燕,李德毅,王建民. 一种基于数据场的层次聚类方法[J]. 电子学报,2006,34(2):258−262

[18] 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. 自动化学报,2015, 41(10):1798−1813.