基于数据场的k-means改进算法

冯应柱1

（重庆交通大学信息科学与工程学院 重庆 400074）1

**摘 要** 针对kmeans算法需要人为给出聚类个数k、孤立点难识别、聚类质量低等问题，提出一种基于数据场的kmeans改进算法。该算法通过计算每个数据点的势值，根据聚类中心的势值比周围邻居的势值大，并与其它聚类中心有相对较大距离的特点，从而确定k个聚类中心；根据孤立点的势值远远小于数据场的均值的特性，找出孤立点；最后将其它数据点按kmeans算法聚类。仿真实验表明，改进算法在不需要人为设定参数的情况下能准确找出类簇中心和孤立点，聚类效果优良。

**关键词** kmeans、聚类中心、数据场、孤立点

**中图法分类号** TP301.6 **文献标识码** A

**Clustering center selection and clustering based on data field**

ZHU Zhenguo1　FENG Yingzhu2

(College of information science and engineering, Chongqing Jiaotong University, Chongqing 400047,China)1

**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. Again by the potential value of the order of magnitude of isolated point is less than the average level, remove the isolated points and finally the other object points into larger than its potential value and nearest neighbor type of clusters, so as to realize clustering. The experimental results show that the method can accurately identify class cluster center, isolated points, and nothing to do with the shape and dimension of the data object,clustering is of good quality.

**Keywords**Cluster center, Data field, Clustering, Outlier

**1引言**

聚类是指按照事物的某些属性，将事物聚集成若干类，使得同一个簇中对象相似度尽量大，而不同簇中的对象相似度尽量小[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，广泛的应用在模式识别、数据分析、图像处理、市场研究以及生物学等领域[3]。

Kmeans[4]是一种基于划分的聚类算法，快速、简洁的特点使得其成为目前应用最为广泛的聚类算法。但原始的kmeans算法也存在一些缺陷：1）算法要求用户事先给定k值，在实际中由于缺乏经验，k值一般难以确定，2）对初始聚类中心敏感，不同的初始聚类中心，会导致不同的聚类结果 [5]。本文提出一种基于数据场的kmeans改进算法，该算法利用数据场的势函数来计算对象间的紧密程度，确定聚类个数k，k个初始聚类中心及孤立点，解决了上述问题。该算法不需要人为输入参数，仿真实验结果表明，该算法能准确找到聚类中心，聚类效果优良。

**2数据场基础**

借鉴物理场的思想，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。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[6]。

定义1：给定维空间，包含个对象的数据集 及其产生的数据场，空间任一点的势值为：

其中，表示场点与对象之间的范数距离；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。

以势值来衡量数据场中单个对象的作用范围，如图1所示。由高斯函数的“3规则”可知：每个对象的影响范围是以该对象为中心、半径等于 的邻域空间，即对象间的相互作用力程为 。

单个对象数据场影响半径R的性质：1、越小，R越小，即对象间的相互作用力程越小；2、越小，势函数衰减速度越快。



图1 不同值的势函数及其影响半径

数据场有如下性质：

独立性：每个对象点都以自己为中心，向外辐射能量而不受外界的影响，如图2，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：每个对象点的势值等于等于该空间中各个点在该点产生的能量总和，如图2，位置1、2处的势值分别为对象A、B在此处势值的叠加总和；

衰减性：势值随着距离的增加快速下降，距离场源越近势值越大，反之，势值越小，当超出一定距离后，其产生的能量可以忽略不计，如图1所示。

图2 数据场性质示意图

**3基于数据场的kmeans改进算法**

本文提出一种基于数据场的kmeans改进算法，该算法通过计算所有数据点的势值和到比它势值更大的数据点之间的最小距离，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点作为聚类中心。通过孤立点检测算法去除噪声数据，最后将其它数据点按kmeans算法聚类，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 ；

Step 2：影响因子 优化；

Step 3：计算数据对象点的势值、距离；

Step 4：确定k个聚类中心；

Step 5：去除孤立点；

Step 6：kmeans算法聚类。

**3.1标准距离计算**

为解决原始数据各维度量纲不一致,数据之间没有可操作性等问题，本文采用min-max标准化，将原始数据转换成无量纲化指标，使得数据之间具有可操作性。

定义2：设包含个对象的数据集，标准化后的值如下：

定义3：给定维空间中任意两点，它们之间的欧式距离:

具有下述的三个属性：

①非负性：；

②；

③

**3.2影响因子**

用于控制对象间的相互作用力程，其取值会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优化算法[8]。

定义4：令对象的势值为，则势熵为：

其中，为一个标准化因子。

势熵H取最小值时，对应的即为最优值。即：

此方程为单变量非线性函数的最小化问题，本文采用初始区间为[的线性探查法优化。步骤如下：

（1）置，精度

（2）计算，，

（3）若，转（4），否则转（5）

（4）若，，否则，停止计算输出

（5）若，，否则，停止计算输出

**3.3确定聚类中心**

定义5：对于包含个对象的数据集，任一对象点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势值中的距离的最小值。即：

其中，对于具有最大势值的对象，将其距离设置为距离的最大值。

定义聚类中心具有这样的特点：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较远，因此聚类中心是势值与距离都较大的点。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 与势值都比较大，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。



图3 原始数据分布与决策关系图

**4仿真实验**

为验证本文算法的可行性，进行了仿真实验，实验采用表1中的5个数据集进行测试，前4个数据集来自文献，第五个数据集为UCI数据集。前4个数据集的原始分布如图4所示。其聚类中心结果展示如图5。

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 |
| Iris | 4 | 3 | 150 |



图4原始数据分布图图5决策图

由于kmeans算法每次选取的聚类中心不一样，导致聚类结果有差异，表2为iris数据集在kmeans算法中进行5次试验的准确率及在改进kmeans算法的准确率，

表2 iris准确率

|  |  |
| --- | --- |
| 序号 | iris的准确率/(%) |
| 1  2  3  4  5 | 89.33  57.33  68.54  88.75  58.64 |
| Km算法均值  改进km算法 | 72.52  92.60 |

**5结束语**

本文提出了一种基于数据场的kmeans改进算法，具有较好的自适应性，能够对任意形状的数据集找出聚类的个数k以及聚类中心的位置，提高了kmeans算法的准确度。但由于kmeans算法本身存在“球形偏见”，对非球形形状的数据集聚类质量不高。

**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. 伍育红. 聚类算法综述[J]. 计算机科学, 2015, 42(s1).
3. 贺玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究 ,2007,01:10-

13.

1. Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.
2. Rodriguez A, Laio A. Clustering by fast search and find of density peaks.[J]. Science, 2014, 344(6191):1492-1496.
3. 李德毅,杜鹢. 不确定性人工智能[M]. 国防工业出版社, 2004.
4. 王莉军,杨炳儒,谢永红.一种基于数据场的社区发现算法[J]. 计算机应用研究,2011,11:

4142-4145.

1. 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):

258-2

1. 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

作者联系方式：

姓名：冯应柱

邮箱：fyingzhu@126.com

电话：18996688781