文章编号：2095-2163(2020)07-0300-03 中图分类号：TP311.13 文献标志码：A

一种基于密度变化的无监督聚类算法

茌良生，徐 建
（江苏师范大学 智慧教育学院，江苏 徐州 221008）

摘 要：在数据挖掘中，聚类分析是一个较活跃的课题，DBSCAN 算法是最常用的基于密度聚类的算法之一，但这种算法无法对不同密度的类实现分类。本文提出一种基于密度变化的聚类算法，能自适应密度，可以对不同的密度类实现很好的分类，克服了 DBSCAN 算法在这方面的缺点。

关键词：聚类算法；密度变化；自适应

Unsupervised clustering algorithm based on density variation

CHI Liangsheng, XU Jian
(School of Wisdom, Jiangsu Normal University, Xuzhou Jiangsu 221008, China)

【Abstract】 Cluster analysis is an active topic in data mining and DBSCAN algorithm is one of the most commonly used algorithms based on density clustering . However, this algorithm cannot achieve classification for classes of different densities . This paper proposes a clustering algorithm based on density variation, which can adapt to density and classify different density classes well. It overcomes the disadvantage of DBSCAN algorithm in this aspect.

【Key words】 Clustering Algorithm; Density change; Adaptive

0 引 言

在数据挖掘中聚类分析是一个研究比较活跃的课题，指将物理或抽象的对象集合分组为由类似对象组成的多个类的分析过程，其目标是在相似的基础上收集数据来分类。聚类分析常见的算法可以分为划分法、层次法及基于密度、基于网格、基于模型的算法等[1]。其中，DBSCAN 算法就是基于密度聚类的算法，其缺点是需要指定距离和点数（密度），对密度不同的类无法实现分类。本文介绍一种密度聚类方法，能自适应类本来的密度，能处理密度不同的类。

1 算法介绍

首先对数据进行处理，剔除相等的点[2]，并将数据的每一维的数据值映射到 0~100 之间。

本算法对数据点逐个分类：

（1）计算每个点的最短距离 $pd(j)$ 为 j 点到其他点的最短的距离；

（2）确定距离最近的两点为一类；

（3）计算并定义类内平均距离 $dm(i)$ 为 i 类内所有点的 $pd(j)$ 的平均；

（4）计算并定义每个类到所有未分类的点的类点距离方法如下：

若 i 类内的数据点 $x_{ij}(j = 1 - n_i, n_i$ 为 i 类内的数据点数），第 k 个未分类的点 y_k；

$S(x,y)$ 为 x 到 y 点的距离，则类点距离 $d(i,k)$ 为最小的 m 个 $s(x_{ij},y_k)(j = 1 - n_i)$ 的平均值，m 是指定的一个参数。

（5）如果没有未分类点则结束算法；

（6）对所有的 i,k 找最小的 $s_{ik} = (d(i,k) - dm(i))/dm(i)$，如果 $s_{ik} < b$（b 是个参数），则 k 点归 i 类，重新计算类内平均距离 $dm(i)$ 及 $d(i,k)$，未分类点数减 1。否则，找所有未分类的点中距离最小的两点建立一个新类，计算新类的类内平均距离和类点距离，未分类点数减 2；如果只剩一个未分类的点，则这一个点建立一个新类，并把未分类点数减 1；

（7）如果没有未分类点则结束算法。

在第（6）步计算 s_{ik} 时，为避免在类内点数过少而使类内平均距离不具代表性，特设置 s_{ik} 乘以函数 $f(n_i)$，其中 n_i 为 i 类内的数据点数，$f(x)$ 定义公式（1）：

$$f(x) = 1 - e^{-\frac{(x-2)}{4}}. \quad (1)$$

其中：m 参数一般固定在 10 左右即可，b 参数在 0 ~ 10 之间，越小类越多，越大类越少，调整到分出类数达到想要的类数即可。

2 应用效果

将本算法应用在 UCI Aggregation 数据集上，效

作者简介：茌良生（1974-），男，硕士，讲师，主要研究方向：计算机多媒体技术；徐 建（1961-），男，学士，讲师，主要研究方向：计算机软件。

收稿日期：2020-03-26

果如图 1 所示，聚类结果如图 2 所示，正确率为 100%；应用在 UCI Aggregation 数据集的变种（将局部变为不均匀的数据集），效果如图 3 所示，聚类效果如图 4 所示；应用于 UCI Jain 数据集，效果如图 5 所示，聚类结果如图 6 所示，正确率为 100%；应用于 UCI Cancer 数据集，聚类结果的正确率为 93%。由此可证实，这一算法切实有效。

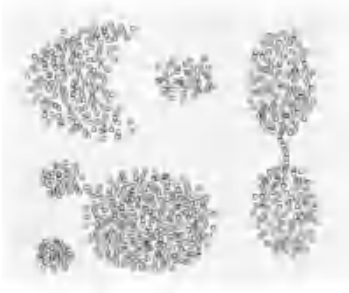

图 1　UCI Aggregation 数据集

Fig. 1　UCI datasets

图 2　UCI Aggregation 数据集聚类结果

Fig. 2　The clustering result of UCI datasets

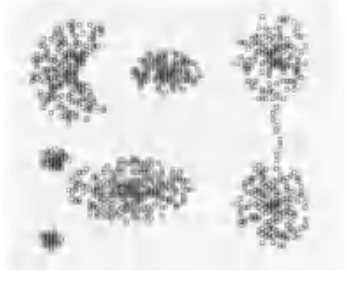

图 3　UCI Aggregation 数据集的变种

Fig. 3　The variant of UCI datasets

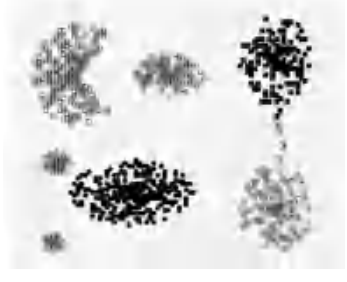

图 4　UCI Aggregation 数据集的变种聚类结果

Fig. 4　The clustering result of the UCI datasets variant

图 5　UCI Jain 数据集

Fig. 5　UCI Jain dataset

图 6　UCI Jain 数据集聚类结果

Fig. 6　The clustering result of UCI Jain dataset

3　与 DBSCAN 算法比较

如图 7、图 8 所示的不同密度数据集，DBSCAN 算法根本无法正确分类。但使用本文算法后，可得出分类结果如图 9、图 10 所示，均可实现正确分类。

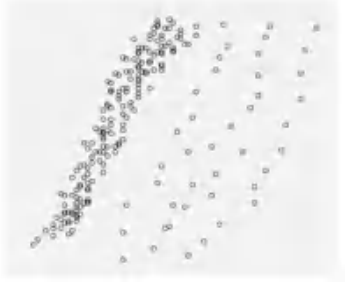

图 7　分散不同密度数据集

Fig. 7　Scattered datasets with different densities

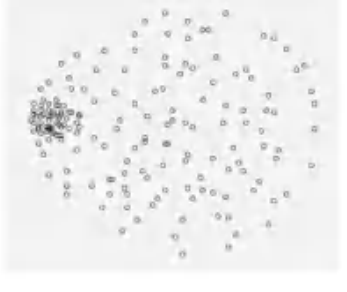

图 8　连续不同密度数据集

Fig. 8　Continuous datasets with different densities

图 9 分散不同密度数据集聚类结果

Fig. 9 The clustering result of scattered datasets with different densities

图 10 连续不同密度数据集聚类结果

Fig.10 The clustering result of continuous datasets with different densities

4 结束语

本文针对 DBSCAN 算法的固有缺点,如需要指定距离和点数(密度),对密度不同的类无法实现分类等,提出一种算法,能够自适应密度,能够分类不同密度的类。将本算法应用在 UCI Aggregation、UCI Jain 等数据集上,均可实现正确分类。实验证明,本文算法是切实有效的。

参考文献

[1] 崔尚卿 ,马秀莉,唐世渭,等. 基于不均匀密度的自动聚类算法[J]. 计算机工程, 2008,34(23):86-88

[2] 张鲁营, 赵晓凡. 一种有效的均值聚类初始化方法[J]. 智能计算机与应用, 2016(3):17-20

(上接第 299 页)

(5)文化产业链及文化产业影响因素研究。主题群 5 的核心关键词包括文化产业园、文化产业链、民营文化产业、动漫产业、文化消费、竞争力等,该部分研究集中在文化产业链上文化生产、内容衍生、营销推广、获取效益等环节,以及文化产业发展的影响因素。

4 结束语

通过对样本文献的外部特征及内容分析,配合分析结果,得出长三角文化产业研究的以下特点:

第一、文化产业政策对长三角文化产业研究具有显著的导向作用。这种导向作用体现于两点:①文化产业政策直接催生新的研究内容;②文化产业政策对已有的研究产生影响。

第二、长三角文化产业研究力量需进一步加强。目前,研究力量多分布于开设文化产业相关专业的高校,各级党委和各级研究院也有一定贡献,各个机构具有自己的核心作者。随着文化产业在国民经济、文化走出去中发挥的作用越来越大,以及文化产业管理等学科在长三角地区高校中的设立,相关的研究力量会越来越强。

第三、长三角文化产业研究的开放性有待增强。长三角文化产业研究存在着一定的封闭性,表现在 3 个方面:①研究者、研究机构之间的科研合作并不频繁,未形成具有代表性的合作网络;②“政学研”在已有的合作关系中,绝大多数存在于同类型机构之间;③地域集聚性较强,即使是不同机构之间、不同机构作者之间的合作,也一般同属于一个地区,跨越行政壁垒进行合作现象较为少见。

因此,在未来的研究中需要丰富研究主体,加强合作;拓宽研究内容,紧跟时代步伐;关注政策文件;寻找文化共性;增加长三角文化产业的整体性研究,并且长三角要承担好自身的责任,推动中华文化走出去,提升国际影响力。

参考文献

[1] 李宇可,刘遵月. 浅析文化衍生品的常见设计模式[J]. 戏剧之家,2019(22):134-135.

[2] 赵悦利,刘遵月. 基于博物馆文化的文创产品研究策略分析研究---以南京博物院为例[J]. 艺术评鉴,2017(5):167-169.

[3] 刘俊哲,王倩,刘彦. 江苏地区旅游文创产品调查与消费者偏好因子分析[J]. 家具,2018(6):85-88.

[4] 赵子青,李毅. 基于 Citespace 的近五年医养结合研究热点探析[J]. 智能计算机与应用,2020(1):290-293.

[5] Garfied E. Scientography : Mapping the Tracks of Science [J]. Current Contents: Social& Behavioral Science,1997,7(45):5-10.

[6] 郭敏,李含伟. 毕业大学生住房保障的知识图谱分析 [J]. 智能计算机与应用,2020(1):149-152.

[7] 陈梦飞,王丽岩. 基于科学知识图谱的中国可穿戴设备研究可视化分析[J]. 智能计算机与应用,2017(1):53-56,59.

[8] 李若辉. 基于“互联网+”的设计专业课程教学改革途径研究[J]. 设计,2018(13):94-95.