一种改进的 DBSCAN 算法

赵　文[1]，　夏桂书[2]，　苟智坚[1]，　闫振兴[3]

(1. 成都信息工程学院 网络工程学院，四川 成都 610225；　2. 中国民用航空飞行学院 航空工程学院，四川 广汉 618307；
3. 北京理工大学 信息与电子学院，北京 100081)

摘要：聚类技术是数据挖掘中的一项重要技术，它能够根据数据自身的特点将集中的数据划分为簇. DBSCAN 是一种经典的基于密度的聚类算法，能发现任意数量和形状的簇，但需设置 Eps 和 MinPts 参数，且聚类效果对参数敏感. 提出一种改进的 DBSCAN 算法，该算法采用自适应的 Eps 参数使得 DBSCAN 算法能对具有不同密度的簇的数据集进行聚类. 仿真实验结果验证了所提算法的有效性.

关键词：聚类；密度；自适应；DBSCAN 算法

中图分类号：TP301　　**文献标志码**：A　　**文章编号**：1001－8395(2013)02－0312－05

doi：10.3969/j.issn.1001－8395.2013.02.032

聚类是数据挖掘中的一项重要技术，它能将大量数据根据其相互间的相似程度划分成若干个类或簇，对于簇内数据间的相似程度相对较大，而不同簇的数据间相似程度相对较小[1]. 聚类分析的结果反映了数据的分布情况及相互关系，同时也为数据的进一步处理提供了基础. 聚类分析作为一种重要的数据处理方法，如今在很多领域中如生物、信息等得到了广泛的应用.

目前，聚类算法分为 5 大类：1) 划分聚类算法，如 K－MEANS 算法[2]、K－MEDOIDS 算法等. 这类算法需设定簇的数量 K，然后随机选取 K 个中心，根据和中心的相似程度，将数据划分为 K 个类，然后，重新确定中心，再划分，如此重复，直到划分结果不再发生变化. 2) 层次聚类算法，该类算法分为“凝聚”和“分裂”2 种方法. “凝聚”的方法是将数据集中的每一个数据都单独作为一个类，然后将最近的 2 个类合并为一个类，如此重复，直到满足一定条件为止；“分裂”的方法是将数据集中的所有数据作为一类，然后将其分解为不同类，直到满足条件，其代表算法有：CURE 算法[3]、BIRCH 算法等. 3) 基于模型的聚类算法，通过假定一个模型，寻找出符合模型的数据集，从而实现聚类. 主要是基于统计的模型和基于神经网络的模型. 4) 基于网格的聚类算法，该算法是将数据空间划分为单个的网格单元，然后针对每个网格单元分别进行聚类，其代表算法有：CLIQUE 算法[4]、WAVE－CLUSTER 算法等. 5) 基于密度的聚类算法，该算法是根据数据的稠密程度进行聚类，通过寻找定义为密度相连的对象的最大集合来形成簇，其代表算法有：DBSCAN 算法[5]、DENCLUE 算法等.

在上述聚类算法中，基于密度的聚类算法不需要预先确定所要划分的类数，并且能在含有噪声的空间数据集中发现任意形状的簇，因此，适合于对未知数据进行聚类. DBSCAN(density－based spatial clustering of applications with noise)算法是一种经典的基于密度的聚类算法[6－12]. 它将空间中的数据抽象为数据点，通过计算点的区域密度来进行聚类，因此它需要邻域阈值(Eps)和点数阈值(MinPts)2 个参数，然后根据参数将具有一定密度的区域划分为簇，其聚类效果对设定的参数值敏感. 对于 Eps 和 MinPts 这 2 个参数的设定，目前已有诸多文献对其进行了研究[13－18]. 针对密度均匀的数据的聚类，文献[13－14]假定 MinPts 参数，分别采用遗传算法和距离排序来估计 Eps. 夏鲁宁等[15]则通过分析数据的统计特性来自适应确定 Eps 和 MinPts 这 2 个参数. 针对具有不同密度的数据的聚类，文献[16]根据基于网格与基于密度的聚类算法间的等效规则来计算不同密度的密度阈值. 周水庚等[16]

收稿日期：2012－05－07

基金项目：四川省教育厅自然科学重点基金(11ZA114)资助项目

作者简介：赵　文(1972—)，男，副教授，主要从事络信息安全、网络舆情监测与分析的研究

提出基于数据分区的 PDBSCAN 算法. 文献 [8] 则提出基于网格分区来确定 Eps 的方法. 这些方法通过确定适合不同密度的簇的阈值参数来使得 DBSCAN 算法能对具有不同密度的数据实现聚类.

本文针对具有不同密度的数据的聚类这一问题,提出一种新的改进的 DBSCAN 算法,该算法在聚类过程中针对不同密度的簇,通过对一维平均距离数据进行 DBSCAN 聚类来估计适用于不同密度的 Eps 参数,然后采用这些自适应变化的 Eps 参数进行聚类,使得 DBSCAN 算法能对具有不同密度的簇的数据集进行聚类. 实验结果验证了所提算法的有效性.

1　DBSCAN 算法

基于密度的聚类算法 DBSCAN 通过计算数据集中每个数据点的区域密度来进行聚类,需设置 Eps 和 MinPts 这 2 个参数. 下面给出 DBSCAN 算法中的一些基本定义和引理.

定义 1　空间中任意一点 P 的 Eps 邻域: 以该点 P 为圆心、以 Eps 为半径的球形区域.

定义 2　空间中任意一点 P 的密度: 点 P 的 Eps 邻域内包含的点的数目.

定义 3　直接密度可达: 给定 Eps 和 MinPts,若点 Q 在点 P 的 Eps 邻域内,且点 P 的密度大于 MinPts,则点 Q 从点 P 直接密度可达.

定义 4　密度可达: 给定 Eps 和 MinPts,若存在一个点链 P_1、P_2、…、P_n,且有 P_{i+1} 从 P_i 直接密度可达,则点 P_n 从点 P_1 密度可达.

定义 5　密度相连: 给定 Eps 和 MinPts,若存在点 O,使得点 P 和点 Q 都从 O 密度可达,则点 P 和点 Q 密度相连.

定义 6　核心点和边界点: 给定 Eps,其密度不低于 MinPts 的点,称为核心点; 不是核心点,但是从核心点密度可达的点,称为边界点.

定义 7　簇和噪声: 基于密度可达性的最大密度相连对象的集合称为簇,数据集 D 中不属于任何簇的点称为噪声点.

引理 1　若 p 是核心点,且 O 是从 p 密度可达的点集,则 O 是一个簇.

引理 2　假定 C 是一个簇,p 是 C 中的任意一个核心点,则 C 等价于从 p 密度可达的点集.

由以上定义和引理可知,一个簇就是密度相连的点的最大集合,且可以由其中任意一个核心点唯一确定.

基于上述事实,DBSCAN 的算法思想是: 从数据集 D 中的任意选择一个点 p 开始,查找 D 中所有关于 Eps 和 MinPts 的从 p 密度可达的点. 如果 p 为核心点,则其 Eps 邻域内的所有点和 p 同属于一个簇,将这些点作为下一轮的考察对象(即候选点),通过不断查找从候选点的密度可达的点来扩展它们所在的簇,直至找到一个完整的簇; 如果 p 不是核心点,即没有对象从 p 密度可达,则 p 被暂时标注为噪声点. 然后,算法对 D 中未被处理的点重复上述过程,进行其他簇的扩展. 最后 D 中不属于任何簇的点即为噪声点.

DBSCAN 算法通过这种迭代查找的方式查找所有直接密度可达的对象,其时间复杂度为 $o(n^2)$,n 为数据集 D 中所有数据的数量. 它能够发现任意数量和任意形状的簇,并能自动识别出噪声,因此,对未知数据集的聚类能力强. 但是该算法也存在着一些缺点:

1) 对输入参数敏感. 对于未知数据集,很难确定合适的能够满足条件的参数 Eps、MinPts,若选取不当,将严重影响聚类的质量;

2) 在该算法中,变量 Eps、MinPts 是唯一不变的,因此,当数据密度分布不均匀时,无法确定出一组能够满足 2 个簇的参数,会导致聚类结果不合理.

由于经典的 DBSCAN 算法中参数 Eps 和 MinPts 在聚类过程中是不变的,使得该算法难以适应密度不均匀的数据集. 下面针对这一问题,提出改进的 DBSCAN 算法来实现对具有不同密度的簇的数据集进行聚类.

2　改进的 DBSCAN 算法

2.1　自适应选择 Eps 参数　对于不均匀数据分布,各个数据与周围数据的相似程度不同,因此,针对每个点,将距离该点最近的多个点的距离平均值作为该点处的稠密程度的评判标准. 即对任意一点 P,根据距离矩阵,选取与 P 点最近的 k 个点,计算距离的平均值. 此时,每个点都能够得出一个 k 最近点平均距离.

然后对所有点的一维 k 最近点平均距离数据进行 DBSCAN 聚类. 再对聚类结果中每类 i 找到其最大平均距离的点. 最后将该点与它的第 k 个最近

点的距离作为该类的邻域阈值 Eps_i,并将其保存以备聚类.

这种发现 Eps 的方法主要考虑到对于不同密度的数据集,应根据各个数据的稠密程度,分别选取合适的阈值进行聚类. 由于聚类中所采用的参数 Eps 只能够确定聚类结果中同一类数据中的密度差别,所以,参数选取所引起的误差不会对聚类结果产生很大影响.

2.2 基于变参数的 DBSCAN 聚类

1) 将 2.1 中得出的邻域阈值 Eps_i 按照由小到大的顺序排列,准备聚类;

2) 选取最小的邻域阈值,MinPts 可以不变,对数据进行 DBSCAN 聚类;

3) 然后使用下一个邻域阈值和 MinPts 作为参数,对标注为噪声的数据再次进行 DBSCAN 聚类;

4) 不断循环,直到所有邻域阈值均使用完毕,聚类结束.

在多次聚类过程中,邻域阈值要由小到大进行聚类. 使用较小阈值进行聚类时,距离较大的类中的数据由于不满足该阈值而不被处理到,所以较小的距离阈值只能处理密度较大的点,不会对小密度数据产生影响.

2.3 算法实现流程 改进 DBSCAN 算法的流程涉及到 2 个结构体.

1) 点结构体:

```
struct dot//点信息
{
CPoint cpt; //点位置
int clu; //点所处的类
bool iscore; //是否为核心点
};
```

2) 距离结构体:

```
struct dist_all//距离信息
{
double dist; //距离
bool iscore; //是否核心
int clu; //距离类
};
```

改进 DBSCAN 算法的具体流程如图 1 所示.

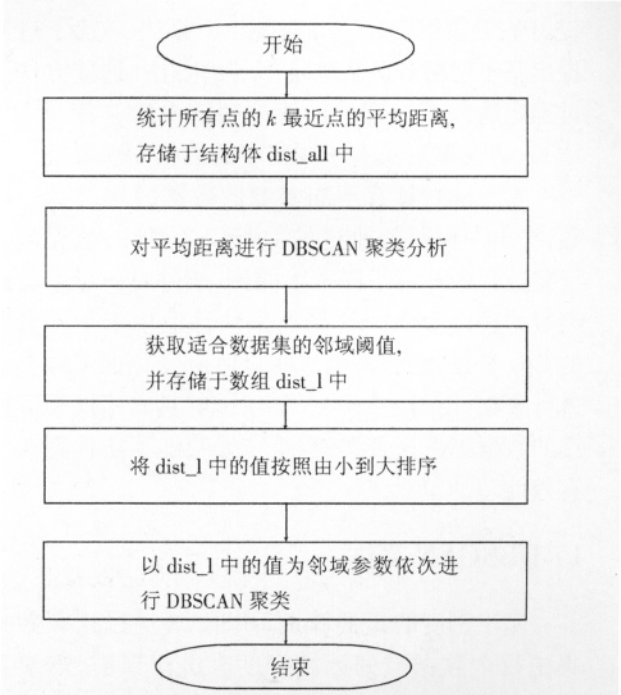

图 1 改进 DBSCAN 算法的实现流程

Fig. 1 The implementation procedure of the improved DBSCAN algorithm

3 仿真结果分析

改进 DBSCAN 算法用 C 语言编程实现,在 VC6.0 环境下调试运行. 在实验中使用了 3 个二维数据样本集,如图 2(a)、图 3(a) 和图 4(a) 所示. 第一个数据集为带有噪声的 2 个不同密度的簇; 第 2 个数据集为带有噪声的 3 个簇,其中 1 个密度较大数据较少,1 个密度较大数据较多,1 个密度较小; 第 3 个数据集为带有噪声的密度均不同的 4 个簇.

下面使用改进的 DBSCAN 算法对样本数据集进行聚类. 在聚类中所设参数为: 求取平均距离时取 $k=5$ 个点; 对一维平均距离数据进行 DBSCAN 聚类时所用 $Eps=10$,$MinPts=3$; 将平均距离聚类所得的各类中最大平均距离点的 5 个距离的最大值为聚类时所需的邻域阈值. 3 个数据集的自适应邻域阀值分别为: 1) 数据集 1 中 $Eps_1=22.022\ 716$,$Eps_2=79.075\ 913$; 2) 数据集 2 中 $Eps_1=37.576\ 588$,$Eps_2=79.397\ 733$,$Eps_3=94.175\ 368$; 3) 数据集 3 中 $Eps_1=40.199\ 502$,$Eps_2=83.743\ 656$. 而 MinPts 参数在 3 个数据集中均固定设置为 3. 改进的 DBSCAN 算法的聚类结果如图 2(b)、图 3(b) 和图 4(b) 所示,其中不同的簇用不同颜色和形状来标注. 可以看出改进的 DBSCAN 算法对不同密度的簇实现了有效的聚类.

为说明改进算法的优势,同时给出了 DBSCAN 算法的聚类结果,以数据集 1 为例,由图 5 和图 6 可以看出当 DBSCAN 算法的邻域阀值变化时,其聚

图 2　改进 DBSCAN 算法对数据集 1 的聚类结果

Fig. 2　Clustering result of data set 1 using the improved DBSCAN algorithm

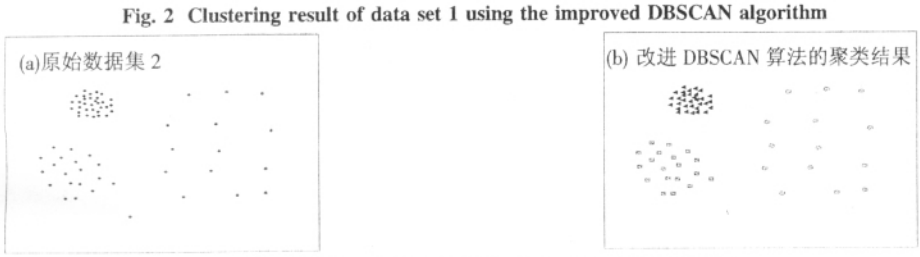

图 3　改进 DBSCAN 算法对数据集 2 的聚类结果

Fig. 3　Clustering result of data set 2 using the improved DBSCAN algorithm

图 4　改进 DBSCAN 算法对数据集 3 的聚类结果

Fig. 4　Clustering result of data set 3 using the improved DBSCAN algorithm

图 5　原 DBSCAN 算法对数据集 1 的聚类结果

Fig. 5　Clustering result of data set 1 using the DBSCAN algorithm

图 6　原 DBSCAN 算法对数据集 1 的聚类结果

Fig. 6　Clustering result of data set 1 using the DBSCAN algorithm

类结果会发生很大变换,说明 DBSCAN 算法对参数敏感,当参数取得不合适时,其聚类结果将不准确. 当数据的密度不均匀时,DBSCAN 算法无法确定出一组参数能够保证不将高密度的簇融进低密度的簇的同时,还能保证不将低密度中的元素作为噪声点处理.

由上述结果可见,改进算法保持了原有的性能,能够发现任意形状和任意数量的簇. 最重要的是,从 3 个数据集的聚类结果可以看出,对簇间的密度差别没有过于严格的限制,即使在同一数据集中存在着 2 个或多个密度差别较大的簇,依然能够成功的得出较为合适的聚类结果. 所以,改进的

DBSCAN 算法能够自动识别各种密度的簇.

4 结语

传统的 DBSCAN 算法由于使用全局统一的邻域阈值参数,不能有效处理具有不同密度的数据集. 本文提出一种改进的 DBSCAN 算法来实现簇间密度差较大的数据集的聚类. 该算法通过采用不同的邻域阀值来逐次实现不同密度的簇的聚类,实验结果验证了该算法的有效性.

参考文献

[1] Chem M S, Han J H, Yu P S. Data mining: An overview from a database perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866 – 883.

[2] Kaufan L, Rpusseeuw P J. Finding Group in Data: An Introduction to Cluster Analysis [M]. New York: John Wiley & Sons, 1990.

[3] Guha S, Rastogi R, Shi M K. CURE: An Efficient Clustering Algorithm for Large Databases [C]//Proc 1998 ACMSIGMOD Inter Conf Manage Data. New York: ACM Press, 1998: 73 – 84.

[4] Agrawal R, Gehrke J, Gunopolos D, et al. Automatic subspace clustering of high dimensional data for data mining application [C]//Proc ACM SIGMOD Inter Conf Very Large Data Base. Roma: Morgan Kaufmann Publishers, 2001: 331 – 340.

[5] Ester M, Kriegel H P, Sander J, et al. A density based algorithm for discovering clusters in large spatial database with noise [C]//Proc 2nd Inter Conf Know Discove Data Mining. Portland: AAAI Press, 1996: 226 – 231.

[6] 荣秋生,颜君彪,郭国强. 基于 DBSCAN 聚类算法的研究与实现 [J]. 计算机应用,2004,24(4):45 – 46.

[7] 冯少荣,肖文俊. 基于密度的 DBSCAN 聚类算法的研究及应用 [J]. 计算机工程与应用,2007,43(20):216 – 221.

[8] 李莉平,沈俊媛. 基于数据挖掘的 DBSCAN 算法及其应用 [J]. 科技创业月刊,2009(8):134 – 135.

[9] 何中胜,刘宗田,庄燕滨. 基于数据分区的并行 DBSCAN 算法 [J]. 小型微型计算机系统,2006,27(1):115 – 116.

[10] 李杰,贾瑞玉,张璐璐. 一个改进的基于 DBSCAN 的空间聚类算法研究 [J]. 计算机技术与发展,2007,17(1):114 – 116.

[11] 王桂芝,玉广亮. 改进的快速 DBSCAN 算法 [J]. 计算机应用,2009,29(9):2505 – 2508.

[12] 冯少荣,肖文俊. 一种提高 DBSCAN 聚类算法质量的新方法 [J]. 西安电子科技大学学报:自然科学版,2008,35(3):523 – 529.

[13] Lin C Y, Chang C C, Lin C C. Fundamental Informaticae, 2005, 68(4): 315 – 331.

[14] Yue S H, Li P, Guo J D, et al. J Zhejiang University Science, 2005, A6(1): 71 – 78.

[15] 夏鲁宁,荆继武. SA – DBSCAN:一种自适应基于密度聚类算法 [J]. 中国科学院研究生院学报,2009,26(4):530 – 538.

[16] 谭颖,胡瑞飞,殷国富. 多密度阈值的 DBSCAN 改进算法 [J]. 计算机应用,2008,28(3):745 – 748.

[17] 周水庚,周傲英,曹晶. 基于数据分区的 DBSCAN 算法 [J]. 计算机研究与发展,2000,37(10):1153 – 1159.

[18] 庞洋,徐巧凤. 基于网格分区确定 DBSCAN 参数的方法 [J]. 计算机与现代化,2010(5):16 – 18.

An Improved DBSCAN Algorithm

ZHAO Wen[1], XIA Guishu[2], GOU Zhijian[1], YAN Zhenxing[3]

(1. *Department of Network Engineering, Chengdu College of Information Technology, Chengdu* 610225, *Sichuan*;
2. *Aviation Engineering Institute, Civil Aviation Flight College of China, Guanghan* 618307, *Sichuan*;
3. *Department of Electronic Engineering, Beijing Institute of Technology, Beijing* 100081)

Abstract: Clustering is an important technique in data mining, which can classify data according to the characteristic of data. DBSCAN is a classical density-based clustering algorithm, which can automatically determine the number of clusters and deal with clusters of arbitrary shapes, however it needs to specify two parameters of Eps and MinPts before clustering and the clustering results are very sensitive to the two parameters. In this paper an improved DBSCAN algorithm is proposed, which can specify Eps adaptively to deal with data sets with different density clusters. Experimental results demonstrate effectiveness of the improved algorithm.

Key words: clustering; density; adaptive; DBSCAN algorithm (编辑 李德华)