文章编号:1671-6833(2006)02-0091-03

基于网格的参数自动化聚类算法

邱保志,张西芝

(郑州大学信息工程学院,河南 郑州 450052)

摘 要:提出了一种基于网格的参数自动化聚类算法PAG,主要目的是解决传统的网格聚类算法对参数敏感的问题.算法采用参数自动化技术来处理参数,即算法开始运行时所需的参数直接由参数自动化技术中的公式计算得出,不需要用户输入任何参数.通过对大量数据集的实验表明,该算法可扩展性好,能处理任意形状和大小的聚类,能够很好地识别出孤立点或噪声,并且有很好的精度.

关键词:网格聚类;参数自动化;孤立点

中图分类号:TP 311 **文献标码**:A

0 引言

聚类是数据挖掘中的一种重要技术,它的目标是将数据集分成若干个子集,同一个子集中的对象是相似的,不同子集中的对象不相似.在几何方面,聚类是在整个数据集中确定由稀疏区域分开的密集区域.目前基于相似性已经有很多聚类算法,这些聚类算法大体上可分为基于划分的聚类算法、基于密度的聚类算法、基于层次的聚类算法和基于网格的聚类算法等.其中基于网格的聚类算法由于只考虑网格单元而不是考虑每个点,所以它的计算效率比较高.基于网格的聚类算法认为:当网格划分的比较细时,每个网格内的点可看作是相似的.但是这些算法都需要用户输入一些参数,并且聚类结果对参数是十分敏感的.

现有的大多数聚类算法如:DBSCAN[1]、Chameleon[2]、Cure[3]等,它们都是致力于如何发现任意形状和大小的类,但在聚类的过程中需要用户输入一些参数,并且很难有效的处理好聚类结果对参数的依赖性.在文献[4]中给出了DBSCAN算法由于参数选取的不同导致了聚类结果的巨大差异,同时表明DBSCAN算法聚类结果的好坏在很大程度上取决于参数的选择.因此,笔者提出了基于网格的参数自动化技术来解决目前聚类算法对参数敏感的问题.

1 网格聚类的概念

给定一个 d 维数据集合,其属性都是有界的.将 d 维数据空间的每一维 K 等分,即把数据空间分割成 K^d 个网格单元.对任一个网格单元,将其每一维再平均两等分,从而形成 2^d 个网格子单元.一个网格单元的邻居是指与其有共同边界的或有共同点的那些网格单元.若一个网格单元中数据点个数大于给定的阈值MinPts,则认为其是高密度单元;否则视其为低密度单元.对一个低密度的网格单元,若其所有邻居单元都是低密度单元,则这个低密度单元中的点为孤立点.聚类就是相邻高密度单元相连的最大集合.

对孤立点/噪声,在聚类过程中应该将其丢弃,如果一个低密度单元的邻居存在高密度单元,那么该单元中的点可能是聚类的边界点,也可能是噪声.为此,我们要将该单元划分为子单元,将靠近高密度单元的子单元中的点合并到高密度单元中,如果子单元相邻的单元都不是高密度单元,子单元中的点认为是噪声数据而被舍弃.这样做的好处是可以提高聚类的质量,也可以避免由于密度的差异而造成聚类结果含有空洞的现象.

2 参数自动化处理

对于网格聚类,传统的算法要求用户输入网格划分值 K 以及密度阈值MinPts,而在PAG算法中,我们使用参数自动化技术来处理参数 K 和MinPts,以解决聚类结果对参数的依赖性问题.

2.1 参数 K 的处理

令 N 为数据集 D 中数据点的个数,则参数 K

收稿日期:2005-11-18;**修订日期**:2006-02-14

基金项目:河南省科技攻关资助项目(324220066);郑州大学青年骨干教师基金资助项目

作者简介:邱保志(1964-),男,河南驻马店人,郑州大学副教授,博士,主要从事数据库、数据挖掘方面的研究.

万方数据

的处理如下:

$$K=\sqrt{N} \tag{1}$$

2.2 密度阈值处理

密度阈值的处理总体上是依据于网格中点数的平均值,在计算过程中会涉及3个平均值,具体计算过程如下:

根据 K 值划分网格,将数据集 D 映射到网格单元中,计算出网格单元中点数的最大值 Max 和非空的网格数 G_n.令,$H=\sqrt{\text{Max}}$,$A_n=A_{n-1}-H$,其中 $H>1$,$n>1$ 且 $n\leqslant H$,$A_1=\text{Max}$;$B_m=(A_m+A_{m+1})/2$,其中 $m\geqslant 1$,且 $m\leqslant H-1$;$C=N/G_n$,则密度阈值 MinPts 的取值如下:

$$\text{MinPts}=[\sum_{m=1}^{H-1}B_m/(H-1)]C/[\sum_{n=1}^{H}A_n/H] \tag{2}$$

以图1为例,该图中的数据集来自文献[5],数据集点数是12 917,根据数据点数可以计算 $K=113$.根据 K 值划分网格,其中网格中点数的最大值 Max=16,$G_n=1\ 130$,则 $H=4$,$C=N/G_n=11$.由以上数据我们可以计算 $A_1=16$,$A_2=A_1-H=12$,$A_3=A_2-H=8$,$A_4=A_3-H=4$;$B_1=(A_1+A_2)/2=14$,$B_2=(A_2+A_3)/2=10$,$B_3=(A_3+A_4)/2=6$.最后由以上数据我们可以根据密度阈值处理公式计算出 MinPts=11.

3 PAG网格聚类算法

首先根据数据点总数 N,利用式(1)计算出 K 值,再根据 K 值将数据空间划分为网格,并将数据集映射到网格单元中,同时计算出网格中点数的最大值 Max 和非空的网格数 G_n,最后按照密度阈值的处理方法即式(2)计算出密度阈值 MinPts.根据密度阈值判断每个网格单元是否为高密单元,然后连接相邻的高密单元形成聚类.PAG算法如下:

输入:N,Min

输出:clusters,noises/outliers

步骤1:根据数据点数 N 计算网格划分值 K.

步骤2:根据 K 值划分 d 维数据空间.

步骤3:将数据集 D 映射到网格单元中,并计算出网格单元中点数的最大值 Max 和非空网格单元数 G_n.

步骤4:计算密度阈值 MinPts,并标识每个网格单元.

步骤5:连接相邻的高密单元形成聚类.

步骤1是根据数据点总数 N 计算出 K 的值.步骤2是根据 K 的值将 d 维数据空间的每一维划分为 K 个大小相等的区间,形成 K^d 个网格单元.对每一个网格单元,将其每一维两等分,形成 2^d 个网格子单元,为下一步做准备.步骤3是通过对数据集 D 的一遍扫描,将 D 中的点映射到数据空间中,并计算网格中点的最大值 Max 和非空的网格数 G_n.步骤4是根据密度阈值的处理方法计算出密度阈值 MinPts,并检查每一个网格单元,如果一个网格单元空间中点的个数大于等于 MinPts,则该单元被标记为高密度单元,否则被标记为低密度单元.步骤5是将网格单元中相连的高密度单元合并,从而形成聚类.当聚类结果生成时,要检查聚类结果中每个聚类中的点数,若点数小于用户给定的阈值 Min,则将其作为小聚类从聚类结果中删除.

4 实验结果与分析

在本实验中所使用的计算机具有256 MB内存,奔腾Ⅳ CPU 2.40 GHz,使用的操作系统是windows XP专业版,算法是用VC进行编程设计的.

PAG算法的时间复杂度为 $O(N+K^d+L*2^d)$,其中 N 是数据点的个数,d 是数据空间的维数,L 是边界单元的个数,K 是每一维上划分区间的个数.算法的时间复杂度与每一维上划分的区间数和边界单元的个数成正比,是数据点个数的线性函数,所以适合于对大数据集进行聚类.

图1中的数据集是文献[5]提供的数据集,该图是算法PAG聚类的结果,从该图中我们可以看到:类和孤立点或噪声能够被有效地分离开来.图2的数据集来自文献[4]的数据集,该图是算法PAG聚类的结果,在该图中孤立点或噪声也能被有效地识别出来.图3的数据集来自文献[4],该图也是算法PAG聚类的结果,在该图中我们可以看到:聚类结果达到了一个较好的精度.

图1 算法PAG的聚类结果图

Fig.1 Clustering result of PAG

万方数据

图 2　算法 PAG 的聚类结果图

Fig.2　Clustering result of PAG

图 3　算法 PAG 的聚类结果图

Fig.3　Clustering result of PAG

5　结论

我们提出一种基于网格的参数自动化聚类算法 PAG，该算法只要求对数据集进行一遍扫描，采用了参数自动化技术，不需要用户输入任何参数，可扩展性好．该算法适用于含孤立点或噪声较少的数据集，并能处理任意形状和大小的聚类．

参考文献：

[1] ESTER M, KRIEGEL H P, SANDER J, et al. A density - based algorithm for discovering clusters in large spatial databases with noise[A]. Proceeding of 2nd Int Conf On Knowledge Discovery and Data Mining[C], Portland: AAAI Press, 1996. 226 ~ 231.

[2] KARYPIS G, HAN E H, KUMAR V. Chameleon: A hierarchical clustering algorithm using dynamic modeling [J]. IEEE Computer, 1999, 32(8): 68 ~ 75.

[3] GUHA S, RASTOGI R, SHIM K. CURE: An Efficient Clustering Algorithm for Large Databases[C]. New York: ACM Press, 1998. 73 ~ 84.

[4] ERTOZ L, STEINBACH M, KUMAR V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data[C]. Canada: SIAM Press, 2003, 2003. 1 ~ 12.

[5] HSU C M, CHEN M S. Subspace Clustering of High Dimensional Spatial Data with Noises [C]. Germany: Springer, 2004. 31 ~ 40.

Grid - based Clustering Algorithm with the Parameter Automatization

QIU Bao - zhi, ZHANG Xi - zhi

(School of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract: This paper presents a grid - based clustering algorithm by the parameter automatization(PAG). The purpose of PAG is to solve the problem that the traditional grid clustering algorithm relies on the parameter of algorithm. PAG does not need the user to input any parameter and it handles the parameter by the technique of parameter automatization. Scanning the dataset only once, the PAG can discover clusters of arbitrary shapes. The experiment results show that it can discover outliers or noises effectively and get good cluster quality.

Key words: grid clustering; parameter automatization; outlier

万方数据