文章编号: 1005－0523(2006)05－0079－04

基于 CLIQUE 的聚类算法研究

付 淇, 李正凡

(华东交通大学 信息工程学院, 江西 南昌 330013)

摘要: 聚类分析是数据挖掘研究领域中一个非常活跃的研究课题. 本文重点分析了高维度数据的自动子空间聚类算法(CLIQUE)及其各类改进算法, 并对其未来发展方向进行了简要展望.

关 键 词: 数据挖掘; 聚类; CLIQUE; 子空间
中图分类号: TP311 文献标识码: A

1 引言

聚类分析是数据挖掘中的核心技术, 聚类在电子商务、图像处理、模式识别、文本分类等领域有广泛的应用. 所谓聚类[1], 就是将一个数据单位的集合分割成几个称为簇或类别的子集, 每个类中的数据都有相似性, 它的划分依据就是"物以类聚". 一个好的聚类算法应具有处理不同类型属性的能力, 并能识别任意形状的聚类, 自动清除孤立点(孤立点指没有包含在任何聚类中的空间对象). 它对数据的输入顺序不敏感, 随输入数据的大小线性地扩展, 当数据维数增加时具有良好的可伸缩性, 并且聚类结果具有可解释性和可用性.

20 世纪 90 年代中期, 聚类研究主要集中在两个方面: 一是对早期算法的改进, 二是创新新的算法. 然而处理大量的高维数据一直是聚类所困挠的问题. Rakesh Agrawal 等在 1998 年提出的对高维数据的自动子空间聚类算法(CLIQUE)[2] 有效的解决了这个难题, 但其算法本身的局限性使聚类数据仍然面临新的问题, 一些新的技术, 如取样技术、浓缩技术、边界调整技术、树型空间索引技术、小波变换技术和细化的技术逐渐应用到该算法中. 一些性能较优的改进算法也相继提出, 如 MAFIA, ENCLUS, CON－CLIQUE, CGDCP 等, 它们开辟了 CLIQUE 聚类的新路.

本文不同于其它类似文章仅限于对聚类算法作总体性的逐个论述, 它按照 CLIQUE 聚类算法的发展脉搏的主线, 以其算法的局限性为依据, 比较全面地对各类改进算法进行分析比较, 并给出了比较的结果, 使用户对各类算法的优缺点更清楚, 使用更明确, 并对其未来的发展方向进行了展望.

2 高维度数据的自动子空间聚类算法(CLIQUE)

2.1 CLIQUE 算法概述

CLIQUE(Clustering In QUEst)[2] 是 IBM 的 Almaden 研究中心数据挖掘课题的研究成果. CLIQUE 在高维数据的子空间中识别稠密的聚类, 对聚类的说明是以最容易理解的 DNF 形式给出的, 所产生的理想结果与所给的输入数据无关, 并且对数据的分布没有任何的数学假设. 自动的高维空间子空间聚类问题可以描述如下: 给定一个数据对象(记录)的集合 D, 参数 ξ 和 τ, 在 k 维空间的所有子空间中进行聚类, 并对每个类给出 DNF 形式的最精简描述.

CLIQUE 算法自动标识高维空间的子空间, 使得在该子空间中数据能够很好地聚类. 所有搜索限制在原始空间的子空间中, 而不是引入新的维度, 这有利于产生可解释的聚类结果.

CLIQUE 算法采用了基于网格和密度的方法. 首

收稿日期: 2006－07－29

作者简介: 付 淇(1978－)女, 江西临川人, 华东交通大学硕士研究生, 主要研究方向为数据挖掘.

先对每个属性进行 ξ 等分，整个数据空间就被划分为一个超长方体集合，对每个单元进行数据点计数，大于某个阈值 τ 的单元称为稠密单元，然后对稠密单元进行连接就构成类. 不同于其它方法，它可以自动地识别嵌入在数据子空间中的类.

定理1 （反单调性质） 如果点集 D 在 k 维空间中是一个类，那么 D 在任何（$k-1$）维空间中的投影也构成一个类.

定理 1 的一种等价形式可以表述为：如果点集 D 在一个（$k-1$）维空间中的投影不是一个类，那么 D 在 k 维空间中也不构成一个类. 上面的叙述表明从“不是”的角度看，该性质是单调的，称这种性质为反单调性质. 在高维空间子空间中进行聚类时，可以利用该性质进行“剪枝”，其用法类似于发现关联规则的 Apriori 算法. 一个 cluster 是指连接的密集单元的最大集合.

2.2 CLIQUE 的优点和局限性

CLIQUE 具有网格类算法效率高的优点，对数据输入顺序不敏感，无需假设任何规范的数据分布. 它随输入数据的大小线性地扩展，当数据维数增加时具有良好的可伸缩性，对于大型数据库中的高维数据的聚类非常有效. 但需要用户输入数据聚类空间等间隔距离和密度阈值参数. 但是 CLIQUE 不能自动去除孤立点，并且由于方法大大简化，它也存在着很多的局限性，主要是以下几个方面：

1）CLIQUE 算法采用固定划分网格的方法，这很容易破坏密集区域的边缘，降低最终结果的准确性.

2）CLIQUE 算法不能自动去除数据集中的孤立点，需要增加额外的计算步骤去除孤立点，这就增加了计算复杂性.

3）CLIQUE 算法利用最小描述长度技术来进行剪枝，以减少候选密集单元的数目. 但是，利用这种技术可能会剪掉一些密集单元，对最终的聚类结果质量造成影响.

4）CLIQUE 算法的很多步骤都采用近似算法，聚类结果的精确性可能因此降低.

3 改进方案

近几年的一些新的研究都着眼于对以前算法的进一步改进和寻找新的聚类途径. 但是没有一种算法能满足所有的标准. 因此，对数据聚类的进一步改进和创新算法仍然任重道远. 针对近年来 CLIQUE 的一些改进研究，如改进压缩技术的 MAFIA 算法、基于熵的比较的 ENCLUS 算法、处理高维数据的 OptiGrid 算法和 ORCLUS 算法、带约束条件的数据聚类 CON－CLIQUE 算法、以及凝聚点的快速聚类算法 CGDCP 等等，对于更有效聚类数据开辟了新路. 下面对这些算法作简要的分析：

3.1 MAFIA 算法

Goil Sanjay 等提出的 MAFIA[3] 是对 CLIQUE 的更新，其主要的改进在于消除了用于子空间检测的剪枝技术，增补了一个合适大小的间隔，并根据维内的分布对维进行划分. 在第一次扫描时，一个维的最小箱数被计算并且置于一个圆柱内. 如果箱与箱之间邻近且有相同的圆柱值则合并. 箱与箱之间的边界并非像 CLIQUE 那么严格，因此，得到的簇的形状得到较大的改进. 文[3] 证实，MAFIA 的运行速度比 CLIQUE 要快 44 倍，并能很好处理大型和高维数据. MAFIA 计算复杂度为 $O(c^{k'})$ 其中 c 是常量，k' 为数据集中簇的子空间的维数.

3.2 ENCLUS 算法

ENCLUS（Entropy－based cLustering）算法[4] 也是在 CLIQUE 算法基础上发展的，但使用了不同的子空间选择准则. 该准则是基于熵的比较，如果由属性 $A_1, \ldots, A_q$ 组成的子空间的熵 $H(A_1, \ldots, A_q) < \omega$（门限值），则该子空间适合分类. 取消这种 MDL 修剪算法裁剪较小子空间方法的，还有朱倩提出的，通过减少样本点数量的方法来达到减少稠密单元的数量.

3.3 OptiGrid 算法和 ORCLUS 算法

CLIQUE 限制在只对原始空间的子空间的搜索，使得结果更简单及更容易理解，最初的数据维对用户一般都有实际的意义. 这种自动子空间聚类方法的实用性和高效性，带来了子空间聚类方法的空前发展. 其中最有代表的就是 OptiGrid 算法、ORCLUS 和 PROCLUS 算法.

OptiGrid[5] （Optimal Grid－Clustering）是处理高维数据的特定算法. 由于维数很高，通常意义上的概念（如均匀分布）就比较模糊了. 算法采用多维网格进行分裂迭代的数据分割，利用超平面对不同密度范围的数据进行分类. 基于网格的算法在处理高维数据时受限于两个条件：把数据集划分为低密度区和尽可能鉴别簇. OptiGrid 就是解决这两个问题的回归算法. 每次回归试图把数据集划分为子集. 每个子集含有一个簇也同样进行回归，当没有分割位面发现时回归结束. 假定数据集含有 n 个数据点，

每个数据点是 d 维的, 算法的计算复杂度在 O(nd) 和 $O(dn\log n)$之间.

PROCLUS(PROjected CLUstering)算法[6] 和 ORCLUS(Oriented projected CLUster generation)算法[7] 是将数据库分成多个子集, 将高维空间分成多个子空间, 形成子集－子空间对, 子集在子空间中的映射形成紧凑的映射类. 两种算法的差异在于子空间的划分不同. 由于 ORCLUS 比 PROCLUS 提出的时间较晚, 且它更稳定、精确些, 这里就着重介绍 ORCLUS 的特点. ORCLUS 使用新的扩展的簇特征向量(ECF)的概念, 来搜索隐藏的子空间, 并通过映射到最类似的结果, 在簇内取消最稀疏的子空间.

3.4　CON－CLIQUE 算法

CON－CLIQUE(CONstrained CLIQUE)[8] 带约束条件的聚类, 是指将特定的领域知识以"约束"的形式表达, 并嵌入到聚类过程中的方法, 旨在使其能够处理实例对约束条件. 算法的基本思路是将约束条件同 CLIQUE 算法的反单调性质结合起来, 共同用于对候选聚类进行"剪枝"操作, 减少 CLIQUE 算法搜索过程中的"盲目性", 提高其效率和聚类质量, 这对于聚类能更好地应用到现实生活中是很必要的工作.

3.5　CGDCP 算法

凝聚点的快速聚类算法 CGDCP[9], 该算法首先通过标准网格单元扫描整个数据空间, 计算出所有的凝聚点. 在生成凝聚点之后, 对数据空间重新划分网格单元, 按照牛顿爬山法原则对凝聚点集进行聚类处理. 将连通的局部密度最优网格单元合并, 由这些局部密度最优网格单元所覆盖的全部网格单元形成一个新子类. 凝聚点能够准确的反映输入数据空间的几何特征, 减少噪声数据对而后进行的聚类过程的干扰. CGDCP 算法的时间复杂性可近似为 $O((1+g)N_p+2+d)N_g)$, 为线性时间复杂性. 数据空间中包含的数据点的数目是 N_p, 形成凝聚点之后重新划分网格的数目为 N_g, 形成的全部子类数目为 N_c. CGDCP 算法具有高效性, 适合于大规模数据挖掘. 其聚类效率明显优于传统爬山法、CLIQUE 算法和 DBSCAN 算法.

3.6　其他技术的应用

针对 CLIQUE 网格硬划分的不足, 陈梅兰和王洪艳相继提出了细化技术[10][11], 其基本思想是移动非密集单元. 当非密集单元的中心和重心不在同一个位置, 其重心偏向于邻近的密集单元, 我们可以以该单元的重心作为新的中心点重新画一个单元, 使得其中的数据点分布尽可能均匀. 从本质上来讲, 新的单元相当于原来单元向密集单元移动, 如果此时找到相连通的密集子空间, 就可以获得比较精确的 cluster.

自适应划分网格[12] 也是用得很活跃的改进, 采用的技术有树型空间索引、小波变换、边界调整技术等.[13] 文提出的改进算法利用小波变换来生成自适应网格的方法, 来去除原始数据的孤立点, 还可以减少候选密集单元的数目, 从而提高算法的准确性, 加速了算法的执行效率, 提高聚类结果的精确性, 并将改进算法采用 Master－Slave 模式并行化实现以增强聚类维数升高时算法的可伸缩性, 同时也有效地降低了算法的计算复杂度. 改进算法采用小波变换形成 $k-1$ 维空间中的聚类, 这些聚类由一些自适应网格组合而成. 小波聚类强调点密集的区域, 而忽略在密集区域外的较弱信息. 这样, 数据的聚类自动显示出来, 并清理了周围的孤立点. 聚类的边界也就精确确定了, 而不需要附加边界修正算法. 算法在生成自适应网格之前, 主要针对预分配网格进行计算, 计算复杂度是 $O(M_P^K)$. 生成自适应网格后的主要计算是以自适应网格为基础, 计算复杂度是 $O(C_P^K)$. 总的空间复杂度为 $O(N)$. 这里维数为 K, 从结点数为 P, M_p 是各 slave 结点生成自适应网格时每一维上预分配的网格数目最大值的平均值, C_p 是各 slave 结点的每一维中自适应网格数目最大值的平均值.

4　结束语

尽管目前已经提出的聚类算法有近百种, 但不同的算法却有其各自的特点, 在实际应用中应该根据具体问题具体分析, 选择使用或设计最佳聚类方法. 且算法向着处理更高维数据、更大型数据库的方向发展, 算法之间的融合更加紧密. 本文总结了 CLIQUE 聚类算法的特点, 其创新之处在于对 CLIQUE 算法的聚类功能的优缺点及各种改进算法进行了重点、详细的阐述和对比. 这些新算法努力把静态的聚类推向动态的、适应性强的、带约束条件的及与生活联系紧密的聚类. 随着聚类分析对象数据集规模的急剧增大, 改进现有算法以获得满意的效率受到越来越多的重视, 对大规模、高维数据库的高效聚类分析依然是个有待研究的开放问题.

参考文献:

[1] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M]. Morgan Kaufmann Publishers. Inc. 2001.

[2] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Application[C]. In: Proceedings of the 1998 ACM—SIGMOD International Conference on Management of Data, Seattle, Washington, 1998—06.

[3] Goil Sanjay, Harasha Nagesh, Alok Choudhary. MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets [R]. Technical Report NumberCPDC—TR—9906—019, Center for Parallel and Distributed Computing, Northwestern University, 1999.

[4] Cheng C, Fu A, Zhang Y. Entropy—based subspace clustering for mining numerical data[C]. In: Proceeding of the 5th ACM SIGKDD, San Diego. CA, 1999: 84~93.

[5] Hinncburg A, Keim D. Optimal Grid—Clustering: Towards Breaking the Curse of Dimensionality in High—Dimensional Clustering[C]. Jn: Proceedings of the 25th Very Large Databases Conference, Edinburgh, Scotland, 1999.

[6] Aggarwal C C, Procopiuc C, Wolf J L, Yu P S, Park J S. Fast algorithms for projected clustering. In Proc. of the ACM SIGMOD Conference PhiladelPhia, PA, 1999: 61—72.

[7] Aggarwal C C, Yu P S. Finding generalized projected clusters in high dimension spaces. In Proc. ACM SIGMOD Int. Conf. 2000.

[8] 冯兴杰, 黄亚楼. 带约束条件的聚类算法研究[J]. 计算机工程与应用, 2005(7): 12~15.

[9] 陈 卓, 孟庆春, 魏振钢, 任丽婕, 窦金凤. 一种基于网格和密度凝聚点的快速聚类算法[J]. 哈尔滨工业大学学报, 2005, 27(12): 1654~1657.

[10] 陈梅兰. 基于网格和密度聚类算法研究[J]. 计算机与现代化, 2005, (2): 1~6.

[11] 王洪艳. 基于聚类的数据挖掘技术在 CRM 中的研究与应用[D]. 武汉大学, 2005.

[12] Nagesh H S, Goil S, Choudhary A N. A Scalable Parallel Subspace Clustering Algorithm for Massive Data Sets[C]. In: Int. Conf. on Parallel Processing, 2000.

[13] 冯 永, 吴开贵, 熊忠阳, 吴中福. 一种有效的并行高维聚类算法[J]. 计算机科学. 2005, 32(3): 216~218.

The Research of Clustering Algorithm Based on CLIQUE

FU Qi, LI Zheng-fan

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Clustering is an active topic in data mining. In this paper, the CLIQUE algorithm and some of its enhanced algorithms are emphatically analyzed. Finally, the paper forecasts the development direction of the data clustering.

Key words: data mining; clustering; CLIQUE; subspace