网格聚类算法研究

李爱华[1]　尹斐斐[2]
(1.保定学院　河北　保定　071000;2.华北电力大学　河北　保定　071000)

【摘　要】聚类分析是数据挖掘中非常重要的方法，并且在很多领域发挥了巨大的作用。本文以研究网格聚类算法为目的，介绍了常见的基于网格的聚类算法，并比较分析了各类算法的基本思想和优缺点。

【关键词】网格聚类算法;STING 算法;WaveCluster 算法;CLIQUE 算法

0.引言

聚类就是将多个数据对象分成不同的类或者簇，每个类中的对象之间具有较高的相似度，而不同类的对象相似度低。聚类算法是数据挖掘中的重要算法，可以应用于机器学习、统计学、模式识别、图像处理、考古学、市场营销和生物学等多个领域。

聚类是数据挖掘的主要任务之一，目前常见的文献中主要有以下几类聚类算法：划分方法、层次方法、基于密度的算法、基于网格的算法及基于模型的算法等。一些聚类算法集成了多种聚类方法的思想，所以有时不能将某个给定的算法划分为属于某一类特定的聚类方法。各类算法各有自己的特点，应用于不同的领域并且发挥了很大的作用，实现了数据的有效聚类。

1.基于网格的聚类方法(grid-based method)

基于网格的方法采用了网格的数据结构，首先将数据空间划分成为有限个单元(cell)，这些单元就形成了网格结构，所有的处理都是以单个的单元为对象的。这种方法的主要优点就是处理速度很快，处理时间与目标数据库中记录的个数无关的，但是又依赖于数据空间的单元数目。代表算法有：STING[1]、WaveCluster、CLIQUE。

1.1 STING(Statistical Information Grid，统计信息网格)算法

STING 算法是一种基于网格的多分辨率聚类算法，其基本思想是：先将数据空间区域划分成矩形单元。对于不同级别的分辨率，通常存在着不同级别的矩形单元，这些单元形成一个层次结构，高层的每一个单元被划分为多个低一层的单元。每个网格单元属性的统计信息如均值等都被预先计算和存储起来，以方便下一步的查询操作。

高层单元的统计参数可以通过计算低层单元获得，这些参数包括：属性无关的参数 count(计数)；属性相关的参数 mean(平均值)，stdev(标准偏差)，min(最小值)，max(最大值)，以及该单元中属性值遵循的分布(distribution)类型，例如一致分布、正态分布等。当数据被装载进数据库时，底层单元的一些参数(如 min、max、stdev、mean)可以直接由数据进行计算。如果分布的类型已经确定，distribution 的值可以由用户指定，也可以通过假设检验来获得。高层单元的分布类型的确定可以基于它对应的低层单元多数的分布类型，通过阈值过滤过程的合取计算来得到。如果低层单元的分布彼此不同，阈值检验失败，那么此时高层单元的分布类型就为 none。

当得到上述的统计参数后，就可以根据统计参数来进行查询处理。统计参数的使用可以按照自顶向下的基于网格的方法来进行查询。大体过程如下：首先，在层次结构中，选定一层(通常选定含少量单元的层)作为查询答复过程的开始点。对选定的当前层次的每个单元，估算其概率范围或者计算置信度区间，该概率用以反映该单元与给定查询的相关程度。此时得到一些不相关的单元和相关单元，不相关单元在以后操作中不再考虑。相关单元用于下一层较低单元的处理。反复进行该处理过程，直到达到底层。最后，如果满足查询要求，则返回相关单元。否则，检索和处理落在相关单元中的数据，直到它们满足查询要求。

与其他聚类算法相比，STING 算法具有以下优点：(1)基于网格的计算是独立于查询的。这主要是因为存储在每个单元中的统计信息提供了单元中的数据不依赖于查询的汇总信息，所以网格的计算独立于查询。(2)STING 算法通过扫描数据库一次来计算单元的统计参数，时间复杂度是 O(n)，n 是对象的数目。在生成层次结构后，一个查询响应时间是 O(g)，这里 g 是最低层网格单元的数目，通常远远小于 n，这些使该算法的效率非常高。(3)网格结构利于并行处理和增量更新。

1.2 WaveCluster(利用小波变换聚类)算法

WaveCluster 的基本思想是：首先通过在数据空间上强加一个多维网格结构，这个结构用来汇总数据，然后采用小波变换变换原特征空间，在变换后的空间中找到密集区域，该算法是一种多分辨率的聚类算法。这种方法中每个网格单元汇总了一组映射到该单元点的信息，它提供给多分辨率小波变换使用以及随后的聚类分析，可以存放在内存中。

该算法的优点是：(1)速度快，并且可以是并行的。(2)小波变换具有多分辨率的特性，该特性有助于发现不同精度的聚类。(3)提供了无指导聚类，并且能够自动排除离群点。

1.3 CLIQUE(Clustering In Quest，维增长子空间聚类算法)算法

CLIQUE 算法是典型的高维空间的子空间聚类算法，综合了基于密度和网格的聚类算法，该算法的基本思想是：给定一个多维数据点的数据空间，数据点在数据空间中通常是分布不平衡的。该算法区分空间中稀疏的和“拥挤的”区域(空间或单元)，找出数据集合的全局分布模式。在 CLIQUE 算法中，把相连的密集单元的最大集合成为簇。如果一个单元中包含的数据点数超过了某个输入参数，则该单元是密集的。

CLIQUE 通过以下两个步骤进行多维聚类：

第一步，CLIQUE 将多维数据空间划分为互不相交的长方形单元，识别每一维中的密集单元。代表密集单元的子空间取交集形成了一个候选搜索空间。

第二步，CLIQUE 为每个簇生成最小的描述。对每个簇，它确定覆盖相连的密集单元的最大区域，然后再为每一个簇确定最小的覆盖[2]。

该算法的优点：(1)对数据高维有良好的伸缩性，对数据输入顺序不敏感，具有处理噪声的能力。(2)方法简化，但是聚类结果的精确可能降低。

1.4 改进的网格聚类算法

基于上述分析，各类算法有各自的优缺点，为了更好的完善聚类算法，国内外出现了很多改进的网格聚类算法，这类算法大多都和其他的聚类算法相结合，如：基于密度和网格的聚类算法：SCI 算法、DCLUST 算法、MAFIA 聚类算法等；基于数据流的网格密度算法(RTCS)；基于网格的层次聚类算法；自动化网格聚类算法(GCA)等算法。

2.结束语

本文对常见的聚类算法进行了阐述和分析，每一种网格聚类算法都有其自身的优缺点，如何将网格聚类算法与实际问题相结合，如何将网格聚类算法更加有效地应用于实践成为本文作者下一步将要研究的问题。科

【参考文献】

[1]W.Wang,J.Yang,R.Muntz.A statistical information grid approach to spatial data mining [C].In Proc.1997 Int.Conf.Very Large Databases, Athens,reecs ,Aug.1997: 186~195.

[2]韩家炜.数据挖掘—概念与技术[M].

[3]范明,孟小峰,译.北京:机械工业出版社,2001 数据挖掘概念与技术.