数据挖掘聚类算法研究

喻彪[1],骆雯[2],赖朝安[1]
(1华南理工大学现代制造信息系统研究中心,广州 510640;
2华南理工大学机械与汽车工程学院,广州 510640)

摘要:聚类是数据挖掘中用来发现数据分布和隐含模式的一项重要技术。全面总结了大部分常用聚类算法的主要特点,对一些经典聚类算法进行比较,并提出了相关结论,最后对几种新型的聚类算法进行基本概括。

关键词:聚类;数据挖掘;聚类算法
中图分类号:O242 **文献标识码**:A **文章编号**:1671—3133(2009)03—0141—05

Research of clustering algorithms based on data mining

YU Biao[1], LUO Wen[2], LAI Chao-an[1]
(1 Research Center of Contemporary Manufacturing Information System, South China University of Technology, Guangzhou 510640, CHN; 2 School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou 510640, CHN)

Abstract: Clustering is an important technology in Data Mining (DM) for the discovery of data distribution and latent data pattern. Provides a detailed survey of primary trait of most general clustering algorithm at first, then makes a comparison among some clustering algorithm and get some conclusion, and generalize some new clustering algorithm basically at last

Key words: clustering; Data Mining (DM); clustering algorithm

0 引言

聚类(Clustering)分析是数据挖掘技术的重要组成部分,它能从潜在的数据中发现新的、有意义的数据分布模式,已经广泛应用于模式识别、数据分析、图像识别及其他许多方面。聚类[1]是在事先不规定分组规则的情况下,将数据按照其自身特征划分成不同的群组。其重要特征是"物以类聚",即要求在不同群组的数据之间差距越大、越明显越好,而每个群组内部的数据之间要尽量相似,差距越小越好。

聚类是一个具有挑战性的研究领域,目前对聚类算法的研究非常多。基本上所有的聚类算法都具有其各自的特点,只适用于某些特定领域,目前还没有能适用于各种领域的聚类算法。如较常用的 K-MEANS算法主要以方法简单、执行效率高见长,但只能识别大小近似的球形类;DBSCAN算法能很好地过滤噪声数据,但其时间复杂度却为 $O(n^2)$,效率不高。聚类算法大体可分为五类:划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法。

本文主要分析聚类算法的几大类别及常用算法特点比较,并对聚类的新发展进行归纳。

1 数据挖掘聚类算法分类

目前存在着大量的聚类算法,而算法的选取主要取决于所研究数据的类型、聚类的目的和应用等方面,要针对某一具体问题选用一种合适的聚类算法。主要聚类算法的分类见表 1。

表 1 主要聚类算法分类

类 别	包括的主要算法
划分(分裂)方法	K-MEANS算法(K-平均)、K-MEDOIDS算法(K-中心点)、CLARANS算法(基于选择的算法)
层次方法	BIRCH算法(平衡迭代规约和聚类)、CURE算法(代表点聚类)、CHAMELEON算法(动态模型)
基于密度的方法	DBSCAN算法(基于高密度连接区域)、DENCLUE算法(密度分布函数)、OPTICS算法(对象排序识别)
基于网格的方法	STING算法(统计信息网络)、CLIQUE算法(聚类高维空间)、WAVE-CLUSTER算法(小波变换)
基于模型的方法	统计学方法、神经网络方法

1.1 划分方法(partitioning method)

给定一个有 N 个元组或者记录的数据集,该方法

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

将构造 K个分组 (要构造的 K个分组划分即最后聚类的结果簇数),每个分组代表一个聚类,要求 $K<N$。K个分组满足下列条件:1)每一个分组至少包含一个数据记录;2)每一个数据记录仅属于一个分组。对应给定的 K,该算法首先给出一个初始的分组方法,以后通过反复替代来改变分组,使得每一次改进之后的分组方案都比前一次要好。好坏的标准是:同一分组中的记录越近越好,不同分组记录越远越好。该类算法的典型代表有 K-MEANS算法、K-MEDOIDS算法等。这些聚类算法对在中小规模的数据库中发现球状簇很适用,对大规模的数据集进行聚类,划分方法还需要进一步扩展。

最著名、最常用的划分方法是 K-MEANS、K-MEDOIDS以及它们的改进和变种。K-MEANS算法是目前应用最为广泛的一种算法,其每个类别均用该类别中所有数据的平均值或加权平均值来表示,这个平均值或加权平均值即为聚类中心。该算法尝试找出使平均误差函数值最小的 K个划分。当结果簇是密集的、而簇与簇之间区别明显时,它的效果较好。当处理大数据集时,该算法相对可伸缩和高效率,其复杂度为 $O(nkt)$,n表示所有对象的数目,k是簇的数目,t为迭代次数。但是,原始 K-MEANS算法也存在如下缺陷:1)聚类结果的好坏一定程度上依赖于对初始聚类中心的选择;2)通常得到的结果只是局部最优解;3)对 K值的选择没有特别明确的准则可依循;4)对"噪声"和离群数据较为敏感;5)只能处理数值属性的数据,不适用于有分类属性的数据;6)该算法不适合于发现非凸面形状或者大小差别很大的簇。

K-MEANS算法有很多变种,主要有 K模方法,综合 K-MEANS和 K模方法的 K原型方法,以及期望最大 (Expectation Maximization, EM)算法等。它们都是扩展了 K-MEANS算法的某些方面而发展起来的,在处理数据时也能得到较好的结果。

K-MEDOIDS是对原始 K-MEANS算法的改进,在初始聚类中心的选取上,K-MEDOIDS不采用簇中对象的平均值作为参照点,而选用簇中位置最中心的对象即中心点作参照点。K-MEDOIDS算法能处理任意类型属性的数据,且能够消除对异常数据的敏感性。但是由于 K-MEDOIDS算法试图找出最好的中心点,它要对所有的数据对象进行分析,因此它的执行代价比 K-MEANS算法要高,同时仍要求指定结果簇的数目 K。

K-MEDOIDS对于小的数据集非常有效,但对于大的数据集没有良好的可伸缩性,为了处理较大的数据集合,可以采用基于选择的方法 (Clustering Large Application, CLARA)的算法,它的主要思想是不考虑整个数据集合,选择实际数据的一小部分作为数据样本,然后对选出的样本使用 K-MEDOIDS方法。

1.2 层次方法 (hierarchical method)

层次方法是对给定的数据对象集合进行层次的分解,层次分解的方向包括自底向上和自顶向下两种,因此层次的聚类方法可以进一步分为凝聚的 (agglomerative)和分裂的 (divisive)层次聚类。现在算法研究主要集中在凝聚层次聚类方面,分裂方面的较少,而凝聚层次聚类的关键点是寻找"距离"最近的两个样本结合。层次聚类算法比较简单,其主要缺陷在于,一旦一个步骤完成,它就不能再被撤销,因此不能更正错误的决定。改进层次方法聚类质量的一个很有前途的方向,是把层次聚类和其他聚类方法结合起来,形成多阶段的聚类。最典型的层次聚类算法主要是 BIRCH算法和 CURE算法。

综合的层次聚类 (Balanced Iterative Reducing and Clustering using Hierarchies, BIRCH)方法,由 Zhang Tian[2]等人提出。它引入了两个概念:聚类特征和聚类特征树 (CF树),它们用于概括聚类描述。这些结构辅助聚类方法在大型数据库中取得高的速度和可伸缩性。CF树是带有两个参数 B和 T的平衡树,其中 B指定分支因子的最大数目,T指定存储在叶节点上的子簇的最大直径。BIRCH算法的核心是采用一个三元组的 CF树,汇总一个簇的相关信息,从而使一个簇的表示可以用对应的聚类特征,而不必用具体的一组点表示,通过构造满足分支因子和簇直径限制的聚类特征树来求聚类。

聚类特征树可以动态地构造,因此不要求所有的数据读入内存,而可以在外存中逐个地读入数据项,新的数据项总是插入到树中与该数据距离最近的叶子中,如果插入后使得该叶子的直径大于类直径 T,则把该叶子节点分裂。其他叶子节点也需要检查是否超过分支因子来判断断裂与否,直至该数据插入到叶子中,并且满足不超过该类直径,而每个非叶子节点的子女个数不大于分支因子。可以通过改变类直径大小,修改特征树大小来控制其内存容量。

BIRCH算法适合于大型数据库,通过一次扫描就可以进行较好的聚类,且具有较好的伸缩性。但由于该算法在聚类的过程中借助了半径或直径概念来控制聚类的边界,因此它不适应于非球状簇的情况;同时在处理过程中需要提供参数聚类个数 K和类直径 T,这两个参数在聚类时很难确定,尤其是对高维数据对象。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

绝大多数聚类算法在处理球形和相似大小的聚类,或者在存在离群数据时变得比较脆弱。代表点聚类(Clustering Using Representatives, CURE)算法解决了偏好球形和相似大小的问题,同时在处理离群数据上也更加健壮。

CURE采用了一种新型的层次聚类算法,该算法选择了基于质心和代表对象方法之间的中间策略。它不用单个质心或对象代表一个簇,而是选择数据空间中固定了数目的具有代表性的点。要得到一个簇的代表点,首先选择簇中的分散对象,然后根据特定的收缩因子向簇中心"收缩"它们。在算法进行时,有最近距离代表点的两个簇被合并。

每个簇有不止一个的代表点使得 CURE可以适用于非球形的几何形状。簇的收缩或凝聚也可以使 CURE在离群数据存在的情况下,得到高质量的聚类结果。同时对于大型数据库, CURE也具有良好的伸缩性,不会影响聚类质量。但是该算法要求用户给出诸如样本大小、希望聚类的数目以及收缩因子等参数,这将给分析带来一定难度和无法确定的主观性,同时这些参数的设置对聚类质量有着非常显著的影响。

1.3 基于密度的方法 (density-based method)

绝大多数划分方法是基于对象之间的距离进行聚类,这样的方法只能发现球状的或凸形的簇。为了发现任意形状的聚类结果,提出了基于密度的聚类方法,这类方法将簇看作是数据空间中被低密度区域分割开的高密度对象区域,其主要思想是只要邻近区域的密度超过某个阈值,就继续聚类,数据稀疏区域中的数据点被认为是噪声数据。这样的方法可以用来过滤"噪声"孤立点数据,发现任意形状的聚类。

基于高密度连接区域 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN)的聚类算法把具有足够高密度的区域划分为簇,并可以在带有"噪声"的空间数据库中发现任意形状的聚类,它定义簇为基于密度的点的最大集合。这里将涉及到一些新的概念定义[3],如 ξ领域、核心对象、直接密度可达、密度可达以及密度相连等。DBSCAN进行聚类的基本思想是 DBSCAN首先通过检查数据库中每个点的 ξ领域来寻找聚类。如果一个点 p的 ξ领域包含多余 MinPts(MinPts表示该对象 ξ领域内包含对象的最小数目)个点,则建立一个以 p为核心对象的新簇。然后, DBSCAN反复地寻找从这些核心对象直接密度可达的对象,这个过程可能涉及到某些密度可达集的合并。当没有新的点可以被添加到任何簇时即意味着该过程结束。

该算法的最大优点是可以发现任意形状的簇,其主要缺点是对用户定义的参数非常敏感,不同的 ξ和 MinPts对聚类的结果影响非常大,且带来很大的主观性,同时该算法不具有良好的伸缩性,算法效率也不是很高。

1.4 基于网格的方法 (grid-based method)

基于网格的聚类方法采用了一个多分辨率的网格数据结构。它将空间量化为有限数目的单元,聚类操作都在这些单元形成的网格上进行。该方法的主要优点是处理速度快,且处理时间独立于数据对象的数目,只与量化空间中每一维的单元数目有关。其主要代表方法包括统计信息网络 (STING)、小波转换方法 (WaveCluster)和聚类高维空间 (CLIQUE)等。

统计信息网络[4] (Statistical Information Grid, STING)算法是一种基于网格的多分辨率聚类技术,它将空间区域划分为矩形单元。针对不同级别的分辨率,通常存在多个级别的矩形单元,这些单元形成了一个层次结构:高层的每个单元被分为多个低一层的单元。关于每个网格单元属性的平均值、最大值和最小值等统计信息被预先计算和存储,便于进行查询处理。

该算法的主要优点在于网格结构有利于并行处理和增量更新,并且效率很高。但是由于 STING采用了一个多分辨率的方法来进行聚类分析,其聚类的质量取决于网格结构最底层的粒度。若粒度较细,则该处理过程花费的代价会显著增加;若粒度较粗,又不能得到较好的聚类结果。而且,该算法在构建父单元时没有考虑子单元与邻近单元间的关系,因此得到的结果簇形状或水平或垂直,没有对角的边界,降低了结果簇的质量和精确性。

1.5 基于模型的方法 (model-based method)

基于模型的聚类方法试图优化给定的数据和某些数学模型之间的适应性。它为每一个簇假定一个模型,寻找数据对给定模型的最佳拟合。一个基于模型的算法可能通过构建反映数据点分布的密度函数来定位聚类。它同时也基于标准的统计数字自动决定聚类的数目,考虑"噪声"数据或孤立点,从而产生健壮的聚类方法。基于模型的方法主要有两类:统计学方法和神经网络方法。由于当前基于模型的算法主要应用于仿生学方面,故在此不做详细讨论。

2 常用聚类算法性能比较

通过对各种常用聚类算法的分析,对各算法的性

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能从以下七个方面进行比较:适合的数据类型、算法的效率、发现的聚类形状、能否处理大数据集、是否受初始聚类中心影响、对异常数据的敏感性和对输入数据顺序的敏感性。算法性能比较见表 2。

表 2　常用聚类算法性能比较

聚类算法	适合数据类型	算法效率	发现的聚类形状	能否处理大数据集	是否受初始聚类中心影响	对异常数据敏感性	对输入数据顺序敏感性
K-MEANS	数值型	较高	凸形或球形	能	是	非常敏感	不敏感
K-MEDOIDS	数值型	一般	凸形或球形	否	否	不敏感	不敏感
BIRCH	数值型	高	凸形或球形	能	否	不敏感	不太敏感
CURE	数值型	较高	任意形状	能	否	不敏感	不太敏感
DBSCAN	数值型	一般	任意形状	能	是	敏感	敏感
STING	数值型	高	任意形状	能	否	一般	不敏感

由表 2可得到以下结论:1)大部分常用聚类算法只适合处理数值型数据;2)若考虑算法效率、初始聚类中心影响性和对异常数据敏感性,其中 BIRCH算法、CURE算法以及 STING算法能得到较好的结果;3) CURE算法、DBSCAN算法以及 STING算法能发现任意形状的聚类。

由于每种方法都有着其各自的优缺点,没有哪种方法可以适用于各种领域的问题,在应用中应针对具体要解决的问题选择适当的算法,也可以对某类算法进行分析研究,进行改进,使其性能得到更进一步的提高,能更好地解决相关问题。

3　聚类算法的新发展

目前经典的聚类算法在很多领域都已经得到非常成功的应用。但由于各聚类算法本身的缺陷和不足,再加上实际问题的复杂性和数据属性等问题,使得大部分的聚类算法都只能很好地解决某一类问题。近年来,随着人工智能、机器学习、模式识别和数据挖掘等领域中,传统方法的不断发展以及各种新方法和新技术的出现,聚类算法得到了长足的发展,分类简介如下。

3.1　基于群的聚类方法

基于群的聚类方法模拟了生物界中蚁群、鱼群和鸟群在觅食或逃避敌人时的行为。该方法一般分为两类:一类是蚁群算法或蚁群优先 (Ant Colony Optimization, ACO);另一类是粒子群优化[5] (Particle Swarm Optimization, PSO)算法。

基于蚁群算法的聚类方法从原理上可以分为四种:1)根据蚂蚁觅食原理,利用信息素实现聚类;2)利用蚂蚁自我聚集行为来聚类;3)基于蚂蚁堆的形成原理实现数据聚类;4)运用蚁巢分类模型,利用蚂蚁化学识别系统进行聚类。蚁群聚类算法具有灵活性、健壮性、分布性及自组织性等特征,适合本质上是分布、动态和交错的问题求解,能解决无人监督的聚类问题,前景广阔。

PSO是进化计算的一个新分支,它主要模拟鱼群或鸟群的行为。在预测精度和运行速度方面,PSO优势明显。

目前,对 ACO和 PSO在数据挖掘中的应用研究仍处于早期阶段,要将这些方法实际用到大规模数据挖掘的聚类分析中还需要做大量的研究工作。

3.2　基于粒度的聚类算法

粒度计算[6]即信息的粒化处理,是关于信息处理的一种新的概念和计算范式,覆盖了粒度方面的方法、理论和技术等几乎所有领域,是人工智能领域的研究热点之一。它模仿人类的思考方式,即人们能从极不相同的粒度上观察分析同一问题,而且能够非常方便地从一个粒度世界跳至另一粒度世界,使其在知识发现领域有着非常广泛的应用。而聚类主要是通过观察式学习,将研究的数据对象按要求分成多个类或簇的过程。聚类和粒度具有天然的相通性

如何将粒度计算与聚类分析结合起来目前仍处于起步阶段,尚未形成一个真正系统的完整理论框架。基于粒度的聚类算法作为一个新的研究方向目前还不是很成熟,尤其是对计算语义的研究还非常少。但是随着粒度计算理论的不断完善发展,相信今后它必将在数据挖掘聚类算法中得到越来越广泛的应用。

3.3　基于模糊的聚类算法

由于模糊聚类分析具有描述类属中间性的优点,能客观反映现实世界,已逐步成为当今聚类分析的主流。著名学者 Ruspini率先提出模糊划分的概念,之后人们相继提出了多种模糊聚类分析方法[7]。比较典型的有基于相似性关系和模糊关系的方法、基于模糊等价关系的传递闭包方法、基于模糊凸轮的最大树方法、基于数据集的凸分解、动态规划和难以辨识关

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

系等方法。但上述方法均不适用于大数据的情况,不能满足实时性较高的场合。基于目标函数的模糊聚类方法把聚类归结成一个带约束的非线性规划,通过优化求解获得数据集的模糊划分和聚类,该算法已成为新的研究热点。

人类最初对于客观事物的认识往往带有模糊性,通常都是通过一些模糊词语来进行交流通信,通过这些模糊信息的综合,人类就能识别客观事物,并获得一些较精准的结论。模糊聚类算法实质上就是利用了人认识事物的规律,使计算机逐步接近人类智能。模糊聚类分析仍将在今后的聚类算法研究中占有重要的一席之地。

3.4 基于综合其他领域的聚类方法

大多数常用的聚类算法是基于距离的分割聚类算法,其聚类结果往往不尽如人意。同时,很多聚类算法需事先确定聚类数目,而且聚类结果受初值影响严重。借鉴物理学量子理论的量子聚类算法可以很好地解决以上问题,并能得到不错的结果。

最近几年,人们开始关注另一种有效的聚类算法——谱聚类算法。该类方法建立在谱图理论基础之上,利用数据相似矩阵的特征向量进行聚类。谱聚类算法是一种基于两点间相似关系的方法,这使得该方法适用于非测度空间。谱聚类算法不仅思想简单,易于实现,不易陷入局部最优解,而且具有识别非凸分布的能力,非常适用于许多实际应用问题,目前该方法已应用于语音识别、图像分割和文本挖掘等领域。但目前谱聚类算法仍然处在发展初期阶段,还有许多问题值得深入研究。

3.5 多种聚类算法的融合

一般聚类算法都有其各自的特点,只能在某些特定问题或领域取得良好的效果,而实际问题往往非常复杂,使用单一的聚类算法不能解决问题。因此对多种聚类算法的融合应运而生,并取得一定的成果。

研究者对多种聚类算法的融合进行了分类:1)基于传统聚类方法的融合,主要包括 CLIQUE、CUBN、CURD以及 RDVS等;2)模糊理论与其他聚类方法融合的方法;3)遗传算法与机器学习融合的方法;4)传统聚类方法与其他学科理论融合的方法,如量子算法、谱聚类算法等。

4 结语

聚类算法作为数据挖掘的重要组成部分,广泛应用于社会的各个领域。目前已经成熟的聚类算法非常多,在实际应用中应该根据具体问题具体分析,选用能解决问题的最佳算法。本文分析了常见的各类聚类算法,并重点论述和比较了各类经典聚类算法的性能和优缺点,同时,也对目前几种新型的聚类方法进行了初步的介绍。对聚类算法的研究仍然是数据挖掘领域的热点之一。

参考文献:

[1] 林宇,等. 数据仓库原理与实践 [M]. 北京:人民邮电出版社, 2005.
[2] 苏新宁. 数据仓库与数据挖掘 [M]. 北京:清华大学出版社, 2006.
[3] 夏火松. 数据仓库与数据挖掘技术 [M]. 北京:科学出版社, 2004.
[4] Wang W, Muntz R, STING: A Statistical Information Grid Approach to Spatial Data Mining[C]. Athens Proceedings of the 23rd Conference on VLDB, 1997, 186 - 195.
[5] 张丽娟,李舟军. 分类方法的新发展:研究综述 [J]. 计算机科学, 2006, 33(10): 11 - 15.
[6] 朱强. 粒度计算在聚类分析中的应用 [D]. 合肥:安徽大学, 2007.
[7] 李明华,等. 数据挖掘中聚类算法的新发展 [D]. 苏州:苏州大学计算机科学与技术学院, 2008.
[8] 徐知行,刘向勇,周晓勤. 基于蚁群算法的选择装配 [J]. 现代制造工程, 2007(9).
[9] 武美先,张学良,杨明亮,等. 带活力因子的粒子群优化算法 [J]. 现代制造工程, 2007(8).
[10] 张建华,江贺,张宪超. 蚁群聚类算法综述 [D]. 阜阳:大连理工大学软件学院, 2006.
[11] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, 2000.
[12] 王鑫,等. 数据挖掘中聚类方法比较研究 [D]. 济南:山东师范大学管理学院, 2006.
[13] 欧阳喜得,黄地龙. 基于模糊聚类的数据挖掘方法与应用 [D]. 成都:成都理工大学信息工程学院, 2008.
[14] 焦李成. 智能数据挖掘与知识发现 [M]. 西安:西安电子科技大学出版社, 2006.

作者简介:喻彪,硕士研究生,研究方向:制造业信息化,数据挖掘。
赖朝安,博士,讲师,主要研究方向:现代制造信息系统。
作者通讯地址:华南理工大学学生宿舍西十八 611室 (广州 510640)
E-mail: byu_yub@163.com
收稿日期:2008-08-26

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net