

数据挖掘中的聚类算法综述*

贺 玲, 吴玲达, 蔡益朝

(国防科学技术大学 信息系统与管理学院, 湖南 长沙 410073)

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。全面总结了数据挖掘中聚类算法的研究现状, 分析比较了它们的性能差异和各自存在的优点及问题, 并结合多媒体领域的应用需求指出了其今后的发展趋势。

关键词: 数据挖掘; 聚类; 聚类算法

中图法分类号: TP391

文献标识码: A

文章编号: 1001-3695(2007)01-0010-04

Survey of Clustering Algorithms in Data Mining

HE Ling, WU Ling-da, CAI Yi-chao

(College of Information System & Management, National University of Defense Technology, Changsha Hunan 410073, China)

Abstract: Clustering is an important technique in Data Mining (DM) for the discovery of data distribution and latent data pattern. This paper provides a detailed survey of current clustering algorithms in DM at first, then it makes a comparison among them, illustrates the merits existing in them, and identifies the problems to be solved and the new directions in the future according to the application requirements in multimedia domain.

Key words: Data Mining; Clustering; Clustering Algorithm

1 引言

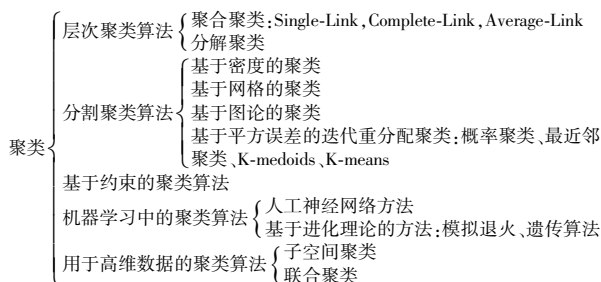
随着信息技术和计算机技术的迅猛发展, 人们面临着越来越多的文本、图像、视频以及音频数据, 为帮助用户从这些大量数据中分析出其间所蕴涵的有价值的知识, 数据挖掘 (Data Mining, DM) 技术应运而生。所谓数据挖掘, 就是从大量无序的数据中发现隐含的、有效的、有价值的、可理解的模式, 进而发现有用的知识, 并得出时间的趋向和关联, 为用户提供问题求解层次的决策支持能力。与此同时, 聚类作为数据挖掘的主要方法之一, 也越来越引起人们的关注。

本文比较了数据挖掘中现有聚类算法的性能, 分析了它们各自的优缺点并指出了其今后的发展趋势。

2 DM 中现有的聚类算法

聚类是一种常见的数据分析工具, 其目的是把大量数据点的集合分成若干类, 使得每个类中的数据之间最大程度地相似, 而不同类中的数据最大程度地不同。在多媒体信息检索及数据挖掘的过程中, 聚类处理对于建立高效的数据库索引、实现快速准确的信息检索具有重要的理论和现实意义。

本文以聚类算法所采用的基本思想为依据将它们分为五类, 即层次聚类算法、分割聚类算法、基于约束的聚类算法、机器学习中的聚类算法以及用于高维数据的聚类算法, 如图 1 所示。



2.1 层次聚类算法

层次聚类算法通过将数据组织成若干组并形成一个相应的树状图来进行聚类, 它又可以分为两类, 即自底向上的聚合层次聚类和自顶向下的分解层次聚类。聚合聚类的策略是先将每个对象各自作为一个原子聚类, 然后对这些原子聚类逐层进行聚合, 直至满足一定的终止条件; 后者则与前者相反, 它先将所有的对象都看成一个聚类, 然后将其不断分解直至满足终止条件。

对于聚合聚类算法来讲, 根据度量两个子类的相似度时所依据的距离不同, 又可将其分为基于 Single-Link, Complete-Link 和 Average-Link 的聚合聚类。Single-Link 在这三者中应用最为广泛, 它根据两个聚类中相隔最近的两个点之间的距离来评价这两个类之间的相似程度, 而后两者则分别依据两类中数据点之间的最远距离和平均距离来进行相似度评价。

CURE, ROCK 和 CHAMELEON 算法是聚合聚类中最具代表性的三个方法。

Guha 等人在 1998 年提出了 CURE 算法^[1]。该方法不用单个中心或对象来代表一个聚类, 而是选择数据空间中固定数目的、具有代表性的一些点共同来代表相应的类, 这样就可以

识别具有复杂形状和不同大小的聚类,从而能很好地过滤孤立点。ROCK 算法^[2]是对 CURE 的改进,除了具有 CURE 算法的一些优良特性之外,它还适用于类别属性的数据。CHAMELEON 算法^[3]是 Karypis 等人于 1999 年提出来的,它在聚合聚类的过程中利用了动态建模的技术。

2.2 分割聚类算法

分割聚类算法是另外一种重要的聚类方法。它先将数据点集分为 k 个划分,然后从这 k 个初始划分开始,通过重复的控制策略使某个准则最优化以达到最终的结果。这类方法又可分为基于密度的聚类、基于网格的聚类、基于图论的聚类和基于平方误差的迭代重分配聚类。

2.2.1 基于密度的聚类

基于密度的聚类算法从数据对象的分布密度出发,将密度足够大的相邻区域连接起来,从而可以发现具有任意形状的聚类,并能有效处理异常数据。它主要用于对空间数据的聚类。

DBSCAN^[4]是一个典型的基于密度的聚类方法,它将聚类定义为一组密度连接的点集,然后通过不断生长足够高密度的区域来进行聚类。DENCLUE^[5]则根据数据点在属性空间中的密度来进行聚类。从本质上讲,DENCLUE 是基于密度的聚类算法与基于网格的预处理的结合,它受目标数据的维度影响较小。此外,Ankerst 等人提出的 OPTICS, Xu 等人提出的 DB-CLASD 和马帅等人提出的 CURD 算法也采用了基于密度的聚类思想,它们均针对数据在空间中呈现的不同密度分布对 DB-SCAN 作了相应的改进。

2.2.2 基于网格的聚类

基于网格的聚类从对数据空间划分的角度出发,利用属性空间的多维网格数据结构,将空间划分为有限数目的单元以构成一个可以进行聚类分析的网格结构。该方法的主要特点是处理时间与数据对象的数目无关,但与每维空间所划分的单元数相关;而且,基于其间接的处理步骤(数据→网格数据→空间划分→数据划分),该方法还与数据的输入顺序无关。与基于密度的聚类只能处理数值属性的数据所不同的是,基于网格的聚类可以处理任意类型的数据,但以降低聚类的质量和准确性为代价。

STING^[6]是一个基于网格多分辨率的聚类方法,它将空间划分为方形单元,不同层次的方形单元对应不同层次的分辨率。STING+^[7]则对其进行了改进以用于处理动态进化的空间数据。CLIQUE^[8]也是一个基于网格的聚类算法,它结合了网格聚类与密度聚类的思想,对于处理大规模高维数据具有较好的效果。除这些算法以外,以信号处理思想为基础的 Wave-Cluster^[9]算法也属基于网格聚类的范畴。

2.2.3 基于图论的聚类

基于图论的方法是把聚类转换为一个组合优化问题,并利用图论和相关的启发式算法来解决该问题。其做法一般是先构造数据集的最小生成树(Minimal Spanning Tree, MST),然后逐步删除 MST 中具有最大长度的那些边,从而形成更多的聚类。基于超图的划分和基于光谱的图划分方法^[10]是这类算法的两个主要应用形式。该方法的一个优点在于它不需要进行一些相似度的计算,就能把聚类问题映射为图论中的一个组合优化问题。

2.2.4 基于平方误差的迭代重分配聚类

基于平方误差的重分配聚类方法的主要思想是逐步对聚类结果进行优化、不断将目标数据集向各个聚类中心进行重新分配以获得最优解(判断是否是最优解的目标函数通常通过平方误差计算法得到)。此类方法又可进一步分为概率聚类算法、考虑了最近邻影响的最近邻聚类算法以及 K-medoids 算法和 K-means 算法。

(1) 概率聚类算法的重要代表是 Mitchell 等人于 1997 年提出的期望最大化算法(Expectation Maximization, EM)^[11]。它除了能处理异构数据之外,还具有另外几个重要的特性:①能处理具有复杂结构的记录;②能够连续处理成批的数据;③具有在线处理能力;④产生的聚类结果易于解释。

(2) 最近邻距离的计算在聚类过程中起着基础性的作用,这也正是导致产生最近邻聚类算法的直接因素。共享最近邻算法(Shared Nearest Neighbor, SNN)^[12]就是该类算法的典型代表之一,它把基于密度的方法与 ROCK 算法的思想结合起来,通过只保留数据点的 K 个最近邻居从而简化了相似矩阵,并且也保留了与每个数据点相连的最近邻居的个数,但是其时间复杂度也提高到了 $O(N^2)$ (N 为数据点个数)。

(3) K-medoids 方法用类中的某个点来代表该聚类,这种方法能有效处理异常数据。它的两个最早版本是 PAM 和 CLARA 算法^[13],此后又有 CLARANS^[14]及其一系列的扩展算法。这类方法具有两个优点:它能处理任意类型的属性;它对异常数据不敏感。

(4) K-means 算法是目前为止应用最为广泛的一种聚类方法,其每个类别均用该类中所有数据的平均值(或加权平均)来表示,这个平均值即被称作聚类中心。该方法虽然不能用于类别属性的数据,但对于数值属性的数据,它能很好地体现聚类在几何和统计学上的意义。但是,原始 K-means 算法也存在如下缺陷:①聚类结果的好坏依赖于对初始聚类中心的选择;②容易陷入局部最优解;③对 K 值的选择没有准则可依循;④对异常数据较为敏感;⑤只能处理数值属性的数据;⑥聚类结果可能不平衡。

为克服原始 K-means 算法存在的不足,研究者从各自不同的角度提出了一系列 K-means 的变体,如 Bradley 和 Fayyad 等人从降低聚类结果对初始聚类中心的依赖程度入手对它作了改进,同时也使该算法能适用于大规模的数据集^[15];Dhillon 等人则通过调整迭代过程中重新计算聚类中心的方法使其性能得到了提高^[16];Zhang 等人利用权值对数据点进行软分配以调整其迭代优化过程^[17];Pelleg 等人提出了一个新的 X-means 算法来加速其迭代过程^[18];Sarafis 则将遗传算法应用于 K-means 的目标函数构建中,并提出了一个新的聚类算法^[19];为了得到平衡的聚类结果,文献[20]利用图论的划分思想对 K-means 作了改进;文献[21]则将原始算法中的目标函数对应于一个各向同性的高斯混合模型;Berkhin 等人^[22]将 K-means 的应用扩展到了分布式聚类。

2.3 基于约束的聚类算法

真实世界中的聚类问题往往是具备多种约束条件的,然而由于在处理过程中不能准确表达相应的约束条件、不能很好地利用约束知识进行推理以及不能有效利用动态的约束条件,使

得这一方法无法得到广泛的推广和应用。这里的约束可以是对个体对象的约束,也可以是对聚类参数的约束,它们均来自相关领域的经验知识。该方法的一个重要应用在于对存在障碍数据的二维空间数据进行聚类。COD (Clustering with Obstructed Distance)^[23]就是处理这类问题的典型算法,其主要思想是用两点之间的障碍距离取代了一般的欧氏距离来计算其间的最小距离。更多关于这一聚类算法的总结可参考文献[24]。

2.4 机器学习中的聚类算法

机器学习中的聚类算法是指与机器学习相关、采用了某些机器学习理论的聚类方法,它主要包括人工神经网络方法以及基于进化理论的方法。

自组织映射 (Self-Organizing Map, SOM)^[25]是利用人工神经网络进行聚类的较早尝试,它也是向量量化方法的典型代表之一。该方法具有两个主要特点:①它是一种递增的方法,即所有的数据点是逐一进行处理的;②它能将聚类中心点映射到一个二维的平面上,从而实现可视化。此外,文献[26]中提出的一种基于投影自适应谐振理论的人工神经网络聚类也具有很好的性能。

在基于进化理论的聚类方法中,模拟退火的应用较为广泛,SINICC 算法^[27]就是其中之一。在模拟退火中经常使用到微扰因子,其作用等同于把一个点从当前的聚类重新分配到一个随机选择的新类别中,这与 K-means 中采用的机制有些类似。遗传算法也可以用于聚类处理,它主要通过选择、交叉和变异这三种遗传算子的运算以不断优化可选方案从而得到最终的聚类结果。

利用进化理论进行聚类的缺陷在于它依赖于一些经验参数的选取,并且具有较高的计算复杂度。为了克服上述不足之处,有研究者尝试组合利用多种策略,如将遗传算法与 K-means 结合起来,并且使用变长基因编码,这样不仅能提高 K-means 算法的效率,还能运行多个 K-means 算法以确定合适的 K 值^[28]。

2.5 用于高维数据的聚类算法

高维数据聚类是目前多媒体数据挖掘领域面临的重大挑战之一。对高维数据聚类的困难主要来源于以下两个因素:①高维属性空间中那些无关属性的出现使得数据失去了聚类趋势;②高维使数据之间的区分界限变得模糊。除了降维这一最直接的方法之外,对高维数据的聚类处理还包括子空间聚类以及联合聚类技术等。

CACTUS^[29]采用了子空间聚类的思想,它基于对原始空间在二维平面上一个投影处理。CLIQUE 也是用于数值属性数据的一个简单的子空间聚类方法,它不仅同时结合了基于密度和基于网格的聚类思想,还借鉴了 Apriori 算法,并利用 MDL (Minimum Description Length) 原理选择合适的子空间。

联合聚类对数据点和它们的属性同时进行聚类。以文本为例,文献[30]中提出了文本联合聚类中一种基于双向划分图及其最小分割的代数学方法,并揭示了联合聚类与图论划分之间的关系。

3 现有聚类算法的性能比较

从上面的分析介绍不难看出,这些现有的聚类算法在不同

的应用领域中均表现出了不同的性能,也就是说,很少有一种算法能同时适用于若干个不同的应用背景。

总体来说,分割聚类算法的应用最为广泛,其收敛速度快,且能够扩展以用于大规模的数据集;缺点在于它倾向于识别凸形分布、大小相近、密度相近的聚类,而不能发现形状比较复杂的聚类,并且初始聚类中心的选择和噪声数据会对聚类结果产生较大的影响。层次聚类方法不仅适用于任意属性和任意形状的数据集,还可以灵活控制不同层次的聚类粒度,因此具有较强的聚类能力,但它大大延长了算法的执行时间;此外,对层次聚类算法中已经形成的聚类结构不能进行回溯处理。基于约束的聚类通常只用于处理某些特定应用领域中的特定需求。机器学习中的神经网络和模拟退火等方法虽然能利用相应的启发式算法获得较高质量的聚类结果,但其计算复杂度往往较高,同时其聚类结果的好坏也依赖于对某些经验参数的选取。在针对高维数据的子空间聚类和联合聚类算法中,虽然通过在聚类过程中选维、逐维聚类和降维从一定程度上减少了高维度带来的影响,但它们均不可避免地带来了原始数据信息的损失和相应的聚类准确性的降低,因此,寻求这类算法在聚类质量和算法时间复杂度之间的折中也是一个重要的问题。

本文选取聚类算法所处理的目标数据的属性(数值型 N/类别型 C)、算法的时间复杂度、能否处理大规模数据集、能否处理异常数据(噪声数据)、能否处理高维数据、能否发现复杂的聚类形状、是否受初始聚类中心影响以及是否受数据输入顺序影响这八个参数,总结比较了一些有代表性的算法的性能,如表 1 所示。

表 1 部分聚类算法性能总结与比较

聚类算法	目标数据属性	时间复杂度	能否处理大规模数据集	能否处理异常点	能否处理高维数据	能否发现复杂的聚类形状	是否受初始聚类中心的影响	是否受数据输入顺序的影响
CURE	N	$O(n_{\text{sample}}^2)$	能	能	否	能	否	否
DBSCAN	N	$O(n \log n)$	能	能	否	能	是	是
Wave-Cluster	N 或 C	$O(n)$	能	能	能	能	否	否
Hyper-graphic	N 或 C	$O(n)$	能	能	能	否	否	否
CLARANS	N 或 C	$O(n^2)$	能	能	否	能	否	否
K-means	N	$O(n)$	能	否	否	否	是	是
SNN	N 或 C	$O(n^2)$	否	能	能	能	否	否
GA	N	与适应度函数相关	能	能	否	能	是	是

表 1 中,算法的时间复杂度都是针对低维数据而言的,K-means 和 GA 也均为原始的标准算法; n 为目标数据的数目,对于 CURE 算法来讲,由于它的执行依赖于对样本集 (Sample) 的选择,所以其时间复杂度由样本集的数据数目来决定。

从表 1 中反映出来的一个最突出的问题在于,这些算法绝大多数不适用于高维数据,而那些少数可以用于高维数据的算法,其时间复杂度也往往会随着维度的升高而显著增高。

总之,虽然一些算法相对其他方法在某些方面的性能有了一定程度的提高,但它不可能在任何应用背景下均具有很好的结果,即几乎没有一个算法能同时在表 1 中所示的八个方面都具有优良的性能。因此对于它们的改进还有一个相当大的空间。

4 总结与展望

聚类算法的研究具有广泛的应用前景,其今后的发展也面

面临着越来越多的挑战。以其在多媒体领域中的应用为例,鉴于多媒体特征数据本身所具备的高维性、复杂性、动态性以及容易达到大规模的特性,对多媒体数据聚类算法的设计还应该更多地考虑以下几个方面的内容:

(1)融合不同的聚类思想形成新的聚类算法,从而综合利用不同聚类算法的优点。

(2)处理大规模数据和高维数据的能力,这是多媒体数据挖掘中聚类算法必须解决的关键问题。

(3)对聚类的结果进行准确评价,以判断是否达到最优解,这也自然要求聚类结果具有可解释性。

(4)选取合适的聚类类别数,这是一个重要的参数。它的确定应更多地依赖于相关的经验知识以及对目标数据集所进行的必要的预处理。

(5)对数据进行合理的预处理。该过程包括对高维数据以及对大规模数据建立索引等,它不仅是实现(4)的前提之一,也为获得更准确的聚类结果提供了一个重要的手段。

(6)在聚类过程中使用合适的相似计算公式及评价准则。合理的相似性评判准则对聚类结果的准确性起着不容忽视的作用。

(7)将领域知识引入聚类过程。领域知识的引入不仅有助于选择合适的模式表达机制、选择合适的聚类算法,还能使以上很多方面的问题都能得到合理的解决,从而提高相应的聚类算法的性能。

在多媒体数据聚类的应用中,对原始数据如图像等进行特征提取,并用这些特征数据代替原始数据进行聚类,均体现了领域知识的融合。

参考文献:

- [1] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases[C]. Seattle: Proceedings of the ACM SIGMOD Conference, 1998. 73-84.
- [2] Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes[C]. Sydney: Proceedings of the 15th ICDE, 1999. 512-521.
- [3] Karypis G, Han E-H, Kumar V. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling[J]. IEEE Computer, 1999, 32(8): 68-75.
- [4] Ester M, Kriegel H-P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. Portland: Proceedings of the 2nd ACM SIGKDD, 1996. 226-231.
- [5] Hinneburg A, Keim D. An Efficient Approach to Clustering Large Multimedia Databases with Noise[C]. New York: Proceedings of the 4th ACM SIGKDD, 1998. 58-65.
- [6] Wang W, Yang J, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining[C]. Athens: Proceedings of the 23rd Conference on VLDB, 1997. 186-195.
- [7] Wang W, Yang J, Muntz R R. STING+: An Approach to Active Spatial Data Mining[C]. Sydney: Proceedings of the 15th ICDE, 1999. 116-125.
- [8] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]. Seattle: Proceedings of the ACM SIGMOD Conference, 1998. 94-105.
- [9] Sheikholsami G, Chatterjee S, Zhang A. WaveCluster: A Multiresolution Clustering Approach for Very Large Spatial Databases[C]. New York: Proceedings of the 24th Conference on VLDB, 1998. 428-439.
- [10] Chris Ding. A Tutorial on Spectral Clustering[C]. ICML, 2004.
- [11] Mitchell T. Machine Learning[M]. New York: McGraw-Hill, 1997.
- [12] Ertöz L, Steinbach M, Kumar V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data[R]. Minneapolis: University of Minnesota, 2002.
- [13] Kaufman L, Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley and Sons, 1990.
- [14] Ng R, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining[C]. Santiago: Proceedings of the 20th Conference on VLDB, 1994. 144-155.
- [15] Bradley P, Fayyad U. Refining Initial Points for K-means Clustering[C]. Madison: Proceedings of the 15th ICML, 1998. 91-99.
- [16] Dhillon I, Guan Y, Kogan J. Refining Clusters in High Dimensional Data[C]. Arlington: The 2nd SIAM ICDM, Workshop on Clustering High Dimensional Data, 2002.
- [17] Zhang B. Generalized K-harmonic Means: Dynamic Weighting of Data in Unsupervised Learning[C]. Chicago: Proceedings of the 1st SIAM ICDM, 2001.
- [18] Pelleg D, Moore A. X-means: Extending K-means with Efficient Estimation of the Number of the Clusters[C]. Proceedings of the 17th ICML, 2000.
- [19] Sarafis I, Zalazala A M S, Trinder P W. A Genetic Rule-based Data Clustering Toolkit[C]. Honolulu: Congress on Evolutionary Computation (CEC), 2002.
- [20] Strehl A, Ghosh J. A Scalable Approach to Balanced, High-dimensional Clustering of Market Baskets[C]. Proceedings of the 17th International Conference on High Performance Computing, Bangalore: Springer LNCS, 2000. 525-536.
- [21] Banerjee A, Ghosh J. On Scaling up Balanced Clustering Algorithms[C]. Arlington: Proceedings of the 2nd SIAM ICDM, 2002.
- [22] Berkhin P, Becher J. Learning Simple Relations: Theory and Applications[C]. Arlington: Proceedings of the 2nd SIAM ICDM, 2002. 333-349.
- [23] Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles[C]. Heidelberg: Proceedings of the 17th ICDE, 2001. 359-367.
- [24] Han J, Kamber M, Tung A K H. Spatial Clustering Methods in Data Mining: A Survey[C]. Geographic Data Mining and Knowledge Discovery, 2001.
- [25] Kohonen T. Self-Organizing Maps[M]. Springer Series in Information Sciences, 2001. 30.
- [26] Yongqiang Cao, Jianhong Wu. Dynamics of Projective Adaptive Resonance Theory Model: The Foundation of PART Algorithm[J]. IEEE Transactions on Neural Network, 2004, 15(2): 245-260.
- [27] Brown D, Huntley C. A Practical Application of Simulated Annealing to Clustering[R]. University of Virginia, 1991.
- [28] Cristofor D, Simovici D A. An Information-theoretical Approach to Clustering Categorical Databases Using Genetic Algorithms[C]. Arlington: The 2nd SIAM ICDM, Workshop on Clustering High Dimensional Data, 2002.
- [29] Ganti V, Gehrke J, Ramakrishna R. CACTUS-Clustering Categorical Data Using Summaries[C]. San Diego: Proceedings of the 5th ACM SIGKDD, 1999. 73-83.
- [30] Dhillon I. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning[C]. San Francisco: Proceedings of the 7th ACM SIGKDD, 2001. 269-274.

作者简介:

贺玲(1976-),女,博士研究生,主要研究方向为多媒体信息系统、多媒体数据挖掘;吴玲达,女,教授,博导,主要研究方向为多媒体信息系统与虚拟现实技术;蔡益朝(1976-),主要研究方向为智能决策。