约束聚类算法研究

郭建军, 梁敬东, 牛又奇

(南京农业大学 信息科学技术学院, 江苏 南京 210095)

[摘要] 约束聚类是聚类研究中的热点之一. 文章就此探讨了在聚类过程中引入领域知识进行 "约束 "的方法. 介绍了约束聚类的定义, 并按约束的应用将约束条件归并为全局约束、实例约束、其它约束等, 然后概括了相应约束条件下的算法, 最后介绍了约束对于聚类带来的益处和问题.

[关键词] 聚类, 约束聚类, 全局约束, 实例约束

[中图分类号] TP 301 [文献标识码] A [文章编号] 1672-1292(2008)04-0128-04

Research on Algorithms of the Constrained Clustering

Guo Jianjun, Liang Jingdong, Niu Youqi

(College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract Constrained Clustering is one of the hotspots in clustering researches. The methods of importing background information to constrain clustering is discussed. The concept of constrained clustering, classify the constrained conditions into global constraints, instance constraints and others constraints according to the application of constraints are presented. Then the algorithms under different constrained conditions are summarized. Finally, the benefits and problems from constraints are discussed.

Key words clustering; constrained clustering; global constraints; instance constraints

聚类是按照对象的某些属性, 将对象分成相似的对象类的过程. 聚类的目标是使得类内的对象尽可能地相似, 类间的对象尽可能地相异. 与分类不同, 聚类通常没有先验知识或背景知识作为指导, 是一个基于对象相似性的自动探测的、无监督学习过程. 由于用户常常具有清晰的应用需求, 所以在许多实际应用中, 有效的解决方案更倾向于将用户偏好或约束加入聚类过程中, 从而对知识发现的结果产生重要的影响, 帮助我们寻找到用户感兴趣的或更加符合用户需求的知识模式.

1 约束聚类的定义

约束聚类是指特定的领域知识以"约束"的形式表达, 并嵌入到聚类过程中的方法. 约束聚类[1] 的定义为: 给定一个具有 n个对象的数据集合 D, 距离函数为 $df: D \times D \rightarrow R$, 一个正整数 k, 一组约束条件 C, 将数据集合 D 划分为 k个不连接的部分 $(Cl_1, Cl_2, \ldots, Cl_k)$, 使得目标函数 $\mathrm{DISP} = \sum_{i=1}^{k} \mathrm{disp}(Cl_i, rep_i)$ 最小, 并且每个类都满足约束条件 C, 记做 $Cl_i \models C$. 其中, $\mathrm{disp}(Cl_i, rep_i)$ 定义为 $\sum_{p \in Cl_i} df(p, rep_i)$, 是类 Cl_i 中的每个对象与代表点的距离总和; rep_i 是类 Cl_i 中的代表点.

2 约束类型及算法

约束是一种背景知识也称为领域知识, 是关于挖掘领域的已知信息. 对于约束的分类, 有多种划分方法. Tung[1] 等按约束的性质和应用, 把约束分为单个对象约束、障碍物约束、聚类参数约束和单个簇约束. Wagstaff[2] 将聚类问题中可用的背景知识按约束作用的范围分为 4种: 全局级约束、聚类级约束、特征级约束和实例级约束. Han和 Kamber[3] 按约束性质将约束分为对距离或相似度函数的约束、对各个簇的性

收稿日期: 2008-06-18

通讯联系人: 梁敬东, 副教授, 研究方向: 数据挖掘和地理信息系统. E-mail: lgd@njau.edu.cn

质用户指定的约束及基于“部分”监督的半监督聚类. 我们采用 Wagstaff的分法, 将约束归并为全局约束、实例约束和其它约束.

2.1 全局约束

全局约束是对全体数据有效的约束, 它的形式常用的有近邻信息和障碍信息. 近邻信息用于二维图像以及三维空间数据对象等的分析, 障碍信息主要用于空间数据的挖掘.

对于二维图像来说, 近邻信息是: 与当前像素相邻的像素更有可能与当前像素归入同一类. 基于这种假设, Theiler和 Gisler提出的 contig-k-means算法[4]综合考虑了方差和近邻两方面的影响, 并通过 $\lambda \in [0, 1]$来调节两方面的影响, 即目标函数 $E = \lambda D + (1 - \lambda) V^*$, 其中 D 是与近邻信息有关的函数, V^* 由方差决定.

障碍信息是指空间数据中, 河流、湖泊等障碍物对挖掘的约束. Tung, Hou和 Han最早界定了在障碍物约束下的聚类问题, 并且提出了 COD-CLARANS[5]算法. 该算法的核心思想: 在考虑障碍物约束的条件下, 通过 1个预处理步骤构造一个可视图, 根据图来计算存在障碍物的情况下 2个对象之间的距离, 这个距离不是直线距离, 而是可能因为障碍物而变成的折线距离即障碍距离. 然后用障碍距离来计算任意两样本点的最近距离, 并将采样技术和 PAM 相结合, 通过迭代的方法来完成在障碍物约束下的聚类问题.

AUTOCLUST+[6]是基于算法 AUTOCLUST, 利用 Delaunay三角网来对数据点进行划分, 同时将约束物以多边形来表示. 先构建 Delaunay三角网, 然后在不考虑任何障碍物的情况下, 计算所有点的边长的标准误差均值, 以获得全局信息, 再将所有与障碍相交的边删除, 最后基于前面的步骤, 将 AUTOCLUST 算法用于这个平面图, 达到分割聚类的目的. DBCluC[7]算法以基于密度的算法 DBSCAN 为基础, 根据对象间的可见性和连通性来体现障碍约束和便利体约束. 用多边形表示各种形状、大小的障碍物, 并用一个预处理即多边形约简算法来降低搜索障碍多边形的复杂度. 此算法能发现任意形状的簇, 并且对噪声和输入顺序不敏感.

DBCOD[8]算法基于 DBCLuC算法, 在预处理时, 采用多边行障碍线法将障碍物多边形合并化简, 然后采用可视图方法得到最短障碍距离, 由于采用障碍物与最小边界矩形 (MBR)的交叠操作, 有效的减少了多边形边的数量. DBRS+[9]算法在基于 DBRS算法, 考虑了障碍物和便利体, 提出“Chop and Conquer”的方法来处理障碍对象, 无需任何预处理, 约束条件已经在聚类过程中处理完成. GKSC-OC[10]算法结合遗传算法和 K-Medoids算法, 首先生成初始群体, 然后计算每个个体的适应度, 并选择优胜个体得到新的群体, 再进行交叉和变异操作, 最后用 K-Medoids算法选取最优个体. 该算法加快了收敛速度和提高了全局搜索能力.

2.2 实例约束

实例约束是对个别实例或一些实例间的关联所作的约束, 它可以是对单个实例独立的约束, 也可以是对实例间的关联的约束.

Wagstaff引入的实例之间的关联[11], 限定某些数据对象应该分入一类或某些数据对象不应分入一类. 为了描述数据对象间的这两种限制, Wagstaff引入了 Must-link和 Cannot-link两种约束, 并分析了它们的对称性及有限的传递性. 一般来说, 假定所给的限制不存在互相矛盾的情况, 即对于 2个给定的数据对象 P 和 Q: Must-link(P, Q) = true和 Cannot-link(P, Q) = true不会同时成立. 已有的基于对象间这两种约束的算法在利用约束实例对之前, 均会依据以上性质, 对所给的限制加以扩充. 这样, 算法在执行时实际所加的约束对数一般会多于最初提供的. Ian Davidson等引入了 δ约束和 ε约束[12], δ约束规定了 1个类中的实例与其它的类中的每个实例的最小距离为 δ; ε约束规定 1个类中的每个实例都必须有另一个实例, 并且两者的距离最大为 ε.

COP-COBWEB[11]算法基于 COBWEB算法, 在描述数据对象间的关系时, 引入了 Must-link和 Cannot-link两种约束, 对数据对象进行 Must-link或 Cannot-link约束检查, 对满足不同的约束, 分入不同的类别中. 算法分为 Must-link检查、添加、新建、合并、分解、更新等步骤. COP-KMEANS[13]算法在 K-means基础上加上 Must-link和 Cannot-link两种约束, 将数据对象归入 (或不归入) 同一组, 从而直接提高准确率. CCL[14]算法是对 Complete-link算法的改进, 在计算出数据间的距离后, 用 Must-link和 Cannot-link来调整距离矩阵, 将 Must-link约束的数据对象间的距离置 0, 然后再调整距离矩阵: 若 $D_{ab} + D_{ac} < D_{bc}$, 则令 $D_{bc} =$

$D_{ab}+D_{ac}$, 其中 a, b, c为数据对象, D 为两个数据对象间的距离. 最后将 Cannot-link约束的数据对象间的距离设为无穷大. 这样就基本满足了 Must-link和 Cannot-link约束. 由于 Must-link约束得到扩展, 算法准确率得到提高. CKS[15]算法引入了子集, 并将 Must-link和 Cannot-link约束应用于数据对象与集合间, 集合与集合间, 突破原有分隔模型的缺陷, 有效提高了准确率. CON-CLIQUE[16]算法在 CLIQUE算法基础上引入实例约束, 对 K 个一维空间聚类中类数目最小的维 D_{min}施加约束, 得到 D_{min}的条件稠密单元, 然后将其与其它 $K-1$个一维空间稠密单元生成候选稠密单元, 再与 CLIQUE算法的反单调性质相结合, 实现了对候选簇进行"剪枝"操作, 减少了算法搜索过程中的盲目性, 提高了效率和聚类质量.

2.3 其它约束

除了上述的约束类型外, 还有属性约束、数字约束[17]、利用标记数据产生约束、γ约束[18]等约束. 其中, γ约束采用三角不等式原理, 因为 $|D(a, b)-D(b, c)|\leqslant D(a, c)\leqslant D(a, b)+D(c, b)$, 如果 $|D(a, b)-D(b, c)|$超过 γ, 则 $D(a, c)$也必定超过 γ, 所以不必计算 $D(a, c)$, 并且 a, c不能放入同一个簇. 基于数字约束的 CDC[17]算法的主要思想就是计算点与各个类之间的代价函数 $Cost(Cl)$, 将使 Cost最小的数据对象放入该类中, 直到所有的点不再移动. 利用标记数据产生约束如 Constrained-K Means算法和 Seeded-KMeans算法[19]. 它们都使用已标记的数据进行初始, 然后用以标记数据产生的约束来指导聚类过程. 它们使用种子点初始化 K-Means算法, 不过不同于 K-Means的 K 随机均值, 而是用第 i个种子集合的均值初始化第 i个类; 不同的是 Seeded-KMeans算法种子点只用于初始化, 而不用于后续处理, 种子点的标记可能被改变, 而 Constrained-KMeans算法在聚类过程中种子数据的类标记不变, 只有非种子数据的标记被重新标记.

3 结语

约束是背景信息的一种体现, 它能够给聚类分析这一无监督分析过程带来指导. 对背景知识的有效利用, 可以使约束聚类算法获得更多的启发式信息, 从而减少搜索过程中的盲目性, 提高算法效率和聚类质量. 首先带有类标识的数据能够指导我们更好地设定距离或相似度的度量方式, 从而提高聚类精度; 其次, 聚类级约束能够帮助我们在聚类过程中获得具有特定性质的期望簇; 最后, 对分层聚类算法来说, 约束能够有效的减少运行时间, 提高聚类算法的效率. 不过约束的使用也带了不少问题, 主要包括两个方面: 是否存在满足所有约束条件的聚类可行解的可行性问题和高度不一致的约束集合引起的精度消减问题. 这两方面的问题仍在研究当中.

[参考文献] (References)

[1] Tung A K H, Han J W, Lakshmanan L V S, et al. Constraint-based clustering in large databases[C]//Proceedings of the 8th International Conference on Database Theory. London: Springer-Verlag, 2001: 405-419.

[2] Wagstaff K. Intelligent clustering with instance-level constraints[D]. Ithaca: Cornell University, 2002.

[3] Han J W, Micheline K amber. Data Mining Concepts and Techniques[M]. 2nd ed. Beijing: China Machine Press, 2006: 444-446.

[4] Theiler J, Gisler G. A contiguity-enhanced K-means clustering algorithm for unsupervised multispectral image segmentation [C]//Proceedings of SPIE. Bellingham WA: SPIE, 1997, 3(159): 108-118.

[5] Tung A, Hou J, Han J W. Spatial clustering in the presence of obstacles[C]//Proceedings of the 17th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2001: 359-367.

[6] Estivill-Castro, V and Lee I J. AUTOCLUST+: automatic clustering of point-data sets in the presence of obstacles[C]//Proceedings of Int' lWorkshop on Temporal, Spatial and Spatio-Temporal Data Mining. Lyon, France. London: Springer-Verlag, 2000: 133-146.

[7] Zaiane O R, Lee C H. Clustering spatial data when facing physical constraints[C]// Proceedings of the IEEE International Conf on Data Mining. Maebashi City, Japan. Washington, DC: IEEE Computer Society, 2002: 737-740.

[8] 杨杨, 孙志伟, 赵政. 一种处理障碍约束的基于密度的空间聚类算法[J]. 计算机应用, 2007(7): 1688-1691.
Yang Yang, Sun Zhiwei, Zhao Zheng. Density-based spatial clustering method with obstacle constraints[J]. Computer Applications, 2007(7): 1688-1691. (in Chinese)

[9] Xin Wang, Camilo Rostoker, Howard J Hamilton. Density-based spatial clustering in the presence of obstacles and facilitators [C]// Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases. New York: Springer-Verlag, 2004: 446-458.

[10] Zhang Xueping, Wang Jiayao, Wu Fang, et al. A novel spatial clustering with obstacles constraints based on genetic algorithms and *K*-medoids[C]// The 6th International Conference on Intelligent Systems Design and Applications. Washington, DC: IEEE Computer Society, 2006: 605-610.

[11] Wagstaff K, Cardie C. Clustering with instance-level constraints[C]// Proceedings 17th Int'l Conf on Machine learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000: 1 103-1 110.

[12] Davidson I, Ravi S S. Clustering with constraints: feasibility issues and the *K*-means algorithm[C]// SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2005: 138-149.

[13] Wagstaff K, Cardie C. Constrained *K*-means clustering with background knowledge[C]// Proceedings of the 18th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 577-584.

[14] Dan Klein, Sepandar D K, Christopher D M. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering[C]// Proceedings of the 19th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2002: 307-314.

[15] 何振峰, 熊范纶. 结合限制的分隔模型及 *K*-Means算法[J]. 软件学报, 2005, 16(5): 799-809.
He Zhenfeng, Xiong Fanlun. A constrained partition model and *K*-means algorithm[J]. Journal of Software, 2005, 16(5): 799-809. (in Chinese)

[16] 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): 12-14, 169.
Feng Xingjie, Huang Yalou. Research on the algorithm of the constrained clustering[J]. Computer Engineering and Applications, 2005(7): 12-14, 169. (in Chinese)

[17] Dai Bin, Lin Chengru, Chen Mingsyan. Constrained data clustering by depth control and progressive constraint relaxation [J]. The VLDB Journal, 2007, 16(2): 201-217.

[18] Davidson I, Ravi S S. Agglomerative hierarchical clustering with constraints: theoretical and empirical results[C]// Proceedings PKDD 2005. Berlin: Springer, 2005: 59-70.

[19] Basu S, Banerjee A, Mooney R. Semi-supervised clustering by seeding[C]// Proceedings of the 19th Int'l Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2002: 19-26.

[责任编辑: 顾晓天]