一种带控制节点的最小生成树聚类方法

汪 闽 周成虎 裴 韬 韩志军 秦承志 蔡 强
(中国科学院资源与环境信息系统国家重点实验室，北京 100101)

摘 要 综合考虑对象间相对距离和高等级对象对低等级对象的集聚效应这两种聚类影响因素，提出了一种带控制节点的最小生成树聚类方法. 该方法用聚类对象间距离为权构建一棵最小生成树，将树中高等级节点作为分割最小树时选取被打断边的控制因素，使本次分割而成的两子树都包含控制节点，且被打断的边是在此条件下的最长边，最终使每棵子树包含且仅包含一个控制节点. 检验自构建数据和地震数据的聚类结果证明，该方法在某些情况下能够较好地揭示数据分布的真实规律.

关键词 聚类 控制节点 最小生成树
中图法分类号：TP391 **文献标识码**：A **文章编号**：1006-8961(2002)08-0765-06

A MST Based Clustering Method with Controlling Vertexes

WANG Min, ZHOU Cheng-hu, PEI Tao, HAN Zhi-jun, QIN Cheng-zhi, CAI Qiang
(*State Key Laboratory of Resources and Environment Information System, Chinese Academy of Sciences, Beijing* 100101)

Abstract Taking into consideration the two clustering factors, the mutual distance between clustering objects and the centralizing effects of the higher level objects on the lower, a new clustering method based on minimum cost span tree with control vertexes is proposed. The MST is built based on the power of the clustering objects' mutual distance, and the selecting standard of the splitted edges is controlled by the higher level vertexes. Each splitted edge should be the longest edge under the condition that the two descendant trees must include at least one controlling vertex, and each descendant tree would include one and only one controlling vertex by the end of the algorithm. It has been verified by clustering the data built by ourselves and the earthquake data that this method, with simple input and little intervention, can discover better the true law of data distribution in some cases. To fulfill the needs of data mining, the selecting standard of the controlling vertexes, the 'inconsistent edges' and the efficiency of the algorithm should be improved.

Keywords Clustering, Controlling vertex, Minimum cost spanning tree

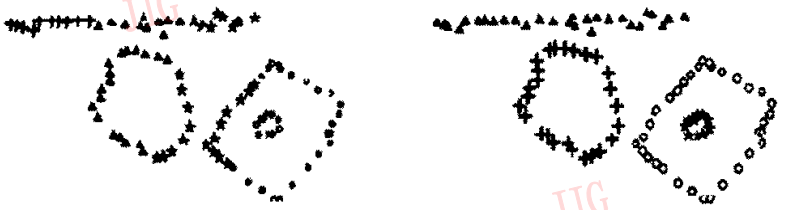
0 引 言

聚类分析是针对某个问题将对象分组，以发现有意义模式的过程[1]. 目前主要的聚类方法可分为如下几类[2]：分割聚类法(如 PAM/CLARA, CLARAN 等)、层次聚类方法(如 CURE, DENCLUE, BIRCH 等)、基于密度的方法(如 DBSCAN, OPTICS, CLIQUE 等)、基于格网的方法(如 STING、GRIDCLUS、WaveCluster 等)，基于模型的方法(如 SOM, ROCK 等)，但几乎所有聚类方法都是将聚类对象作为同级对象处理的，基本没有考虑对象之间的等级差异，即使如 ISODATA, CLARANS 这种有初始种子点的聚类方法也是如此. 而在现实世界中，同类事物或事件，等级高的往往对等级低的有一种"控制"作用，表现为高对低的集聚效应，如大城市和其周边中小城镇的空间集聚分布格局就是一个十分明显的实例. 在许多聚类问

基金项目：中国科学院知识创新项目(CX10G-D00-06、KZCX1-Y-02)
收稿日期：2001-06-22；改回日期：2001-12-07

题中漠视这种等级差异信息,有可能产生歪曲的聚类结果,或难以挖掘出真实的丛聚模式.

然而,由于空间结构的复杂性,低等级对象对高等级对象的聚集,有时并不表现为围绕高等级对象的圆或近圆分布,如地震序列(在发震机制上具有某种内在联系,或有共同发震构造的一组地震的总称[3])中,中小地震群体在空间分布上,有时一方面很靠近序列的一个或几个强震,另一方面又沿一定的地质断裂密集分布.此外,直观理解,空间上最接近的两点,往往在性质上最接近,反映在聚类问题中,就可理解为两点之间越接近,就越应该聚成一类.基于最小生成树(Minimum Cost Spanning Tree,MST)的聚类方法是这种聚类思想的体现,因此其聚类结果在空间上的分布模式就有别于以最小均方误差为标准的聚类方法(如 ISODATA),可以表现为较随意的形状(如图 1).



(a) ISODATA 聚类算法 (b) 最小生成树法

图 1 ISODATA 和最小生成树聚类结果比较

带控制节点的 MST 聚类方法是在以上两种聚类思想的指导下而提出的.它一方面初步考虑了聚类对象之间的等级差异,认为高等级对象对低等级对象有控制作用,会对事物(事件)聚类模式产生很大影响,另一方面则力图将空间上接近的对象尽可能地归并到一类中去,因此可以将其看成是两种聚类影响因素的折衷考虑.

1 基于 MST 的聚类方法

最小生成树 MST 的概念如下[4]:对于图 G,一个包含全部 n 个顶点的 $n-1$ 条边的树,称为 G 的生成树,若给 G 的各个边规定一权值,则得到原图的加权图.权值和为最小的树为最小生成树,也叫最小张树.

MST 的常见构造算法有普里姆(Prim)算法、克鲁斯卡尔(Kruskal)算法等[5].前者的时间复杂度为 $O(n^2)$,后者的时间复杂度为 $O(e\log e)$(e 为图 G 边的数目).对聚类问题而言,节点数一般较多,而且每对节点之间的边都需参与生成计算,利用上述方法构造 MST 是相当耗时的,为此人们提出了一些改进方法,如在构建过程中,引入并行计算[6].

基于 MST 的聚类就是用一定标准对树求割集.Zahn 介绍了 MST 和求割有关的几个性质[7],指出 MST 的任一条边都是某个点集到其余节点的最小跨度.出于人类感官系统习惯于将点间距离较小的对象聚为一类的考虑,他对如何利用 MST 聚类不同形态的样本进行了详细研究,提出了诸如对类接触样本寻找最小深度节点的分割方法、对正态分布样本寻找最小密度节点的分割方法等等[4,7].此外尚有 Fehlauer 等结合 MST 和点密度函数(point density function,PDF)的聚类方法[8],Koontz 等[9]、*Mizoguchi* 等[10]基于有向树的聚类算法等等.

这里介绍 Zahn 提出的一种直观的 MST 聚类方法[1,7](以下称之为 Zahn 法):对于类间相互分离的样本点构成的 MST,那种比其两边相邻的边明显要长(权重明显大)的边应该打断.Zahn 称这种边为"不一致(inconsistent)"的边.这样,打断 MST 中最"不一致"的一条边,将 MST 分割成两棵子树,接着再将子树中的最"不一致"边打断,成为 3 棵子树,依次进行下去,打断 n 次,就有 $n+1$ 个子树,也就是 $n+1$ 个类.最"不一致"边打断之后,分割的两棵子树依旧是最小生成树,且这种分割方式产生的两棵子树的权总和,一般来说相对于打断其他边的权总和是最小的,它符合"将最接近的对象聚成一类"的思想.应该指出,Zahn 法"不一致"边的概念是比较广的,不同样本分布情况下,选择标准不同,如上面所提到的对由类接触样本构造的 MST 做聚类分割时,"不一致"边就是所谓最小深度节点所依附的边.在带控制节点的 MST 方法中,"不一致"边为当前待分割树中的最长边,这符合聚类思想,实践也证明了其结果是比较合理的.

Zahn 法以对象间相对距离为聚类标准，在一般情况下可以取得相当不错的聚类效果（如图 1），但是，在实际应用中也发现其明显缺点，例如对图 2 的点集，分成两类且大致从 1 处分割应是最为合理的，它符合人们的直觉思维，而且分割后两类点数相差不大. 右上角虽仅仅有 3 个点，但由于它们和其最邻近点距离最远，Zahn 法将在 2 处分割点集，因此造成聚类结果个数相差太大，不甚合理，这说明 Zahn 法容易受类似“噪声”情况的影响. 显然，如果图中对象是围绕某两个对象（控制节点）集聚分布的（这在现实世界中是常见的），那么用这两个对象来控制边的选择，使控制节点均分到两类中去，就有可能避免这种情况的发生.

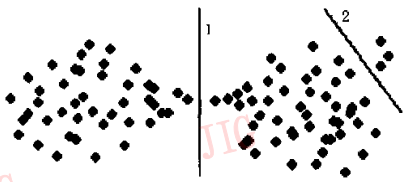


图 2　Zahn 法聚类结果

2　带控制节点的 MST 聚类方法

带控制节点的 MST 聚类方法的主要步骤如下：

（1）控制节点的选取　将待聚类的对象按等级由大到小排序，选取级别高的做控制节点. 但如何划定控制和被控制的分界？其标准是，如果类的个数需预先确定（如 m 个），那么就选取不少于 m 个最高等级对象作候选控制节点. 候选点不少于 m 个的原因在于：在实际应用中发现，有时有些候选控制节点相互之间距离过近，造成分割结果不甚理想，为此这类候选点应有取舍. 可采取设置门槛距离的方法来完成这一任务：如果两候选点之间的距离小于门槛，则将低等级的舍去，保留等级较大的一个. 该门槛值可以根据候选点分布及需要的类别个数，由计算机自动计算，也可由用户指定，这样最终保证控制节点个数等于所需类别个数；如果类别个数预先不确定，那么可采用人机交互方式，从高到低依次挑选，使控制节点尽可能均匀分布在整个试验区，然后类似于前，合并相互之间距离过近的候选点.

（2）构造一棵最小生成树.

（3）尝试打断　对于一棵最小生成树，尝试打断其最“不一致”的边（这里选择标准是当前树中的最长边），如果它将原树合理地分割为两棵子树（都至少包含一个控制节点），那么就打断它，然后递归打断子树中的最长边；否则就应选取其次长边进行打断，这个过程持续进行下去，直到每棵子树都包含且仅包含一个控制节点为止.

以下给出带控制节点的最小生成树聚类算法的主要步骤：

输入：n 个待聚类对象和 m 个控制节点（$m>1$）

输出：m 棵子树（m 个类）

（1）清空待选边集 E，当前试分割树集 TRY.

（2）由 n 个对象选择合适的权标准构造一棵最小生成树 T，加入到 TRY.

（3）如果 TRY 为空，转第 7 步，否则从 TRY 中取走第 1 棵树，对其边按降序排序，加入到 E.

（4）从 E 中取走最长边 e_{max}.

（5）打断 e_{max}，将 T 分为两棵子树，设为 T_1，T_2.

（6）判断 T_1，T_2 的包含控制节点情况：

如果 T_1，T_2 两子树中有一个不包含控制节点，则重新补上 e_{max} 边，转第 4 步；

如果 T_1，T_2 中有一个包含且仅包含一个控制节点（设为 T_1），则说明本打断是合理的，清空 E，将 T_2 加入 TRY，转第 3 步；

如果 T_1，T_2 都包含多于一个的控制节点，则说明本打断是合理的，清空 E，将 T_1，T_2 加入 TRY，转第 3 步.

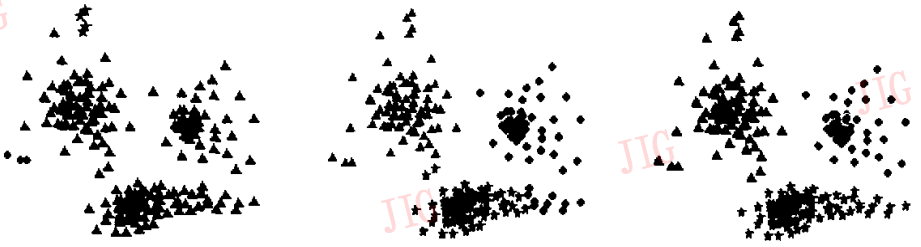
（7）输出子树集.

去除构造 MST 部分，算法的时间复杂度可用如下方法计算其上限：算法主要分为排序和尝试打断部分. 其中，排序部分，其算法复杂度因选择的排序方法而异，设 n 个对象排序，需 $s(n)$ 次基本运算，很显然，排序部分运算次数不会超过 $(m-1)s(n)$；尝试打断部分，考虑到第 k 次（$k\geq 1$）合理打断时被打断的树的边数至少为 $n-k$，最差情况下，尝试次数不会超过 $n-k$ 次，而 m 个控制节点，此种合理打断要进行 $m-1$ 次，因此尝试打断次数总和，在最差情况下不超过 $n-1+n-2+\cdots+n-(m-1)=(m-1)(n-m/2)$，在不考虑优化情况下，每次尝试打断都要对 n 个对象进行一次扫描，以判断控制节点分布情况，因此尝试打断部分基本运算次数不超过 $n(m-1)(n-m/2)$，故算法最差情况下的总体时间复杂度可简化表述为 $O(ms(n)+mn^2)$.

3 应用实例

为说明本方法的有效性和优点,构造如图 3 的数据:假设有 3 个控制节点(如较高等级的城镇),它们之间存在难以逾越的障碍(图中斜线多边形部分,如水域、高山等),那么假设经过若干年发展在其周围出现大致如图 3 右部的居民点聚集情况应是合理的. 在不加入有关障碍先验知识的情况下,用Zahn法、ISODATA 法和基于控制节点的 MST 方法对此数据的聚类结果作了对比,结果如图 4 所示.

图 3 一种可能的集聚模式



(a) Zahn 法 (b) ISODATA 法 (c) 带控制节点的 MST 法

图 4 不同方法聚类结果对比

从图 4 中可以看出,Zahn 法由于受上、左部分点的"干扰",对图 3 数据的聚类效果最差,完全不能反映数据直观丛聚模式;ISODATA 方法对其他两类聚类效果尚可,但将下角条状类别割裂;而基于控制节点的 MST 方法,则对 3 个类别各自有很好地聚集.

用真实世界中的地震数据对算法的有效性做进一步验证. 选用地震数据的原因在于:地震是地球内部弹性介质的破裂和能量释放的过程[11]. 通常一次较大的地震发生之后,周围会有许多大大小小的余震发生,主震震级越大,其余震震中分布范围越大[11]. 此外,一次大地震的发生,释放掉一个相当大的地区内长期积累起来的应力,而在这个空间范围内若再次积累足以发生另一次大地震的应力,则需要相当长的时间. 换句话说,就是一次地震发生后,短时间内附近再次发生另一次大地震的可能性很小,即这个区域在短时间内是"免疫"了的[12]. 大地震之间的相互"排斥"和中小地震的伴随发生,造成大地震对其周围中小地震的一种类似控制模式下的空间集聚效果,这是选用带控制节点的 MST 方法对其丛聚特性进行研究的原因.

地震数据来源于自行编录的中国及邻区地震数据库,共收集地震条目510 255条[13]. 从中抽取华北地区(32°~42°N,109°~122°E)1 500年来大于 4.7 级的地震共 518 条,做了聚类分析. 类别个数没有预先确定,而把控制节点选取标准定为地震震级大于 6.8 级以上的地震,共 16 个,设置门槛距离为 110km,最终控制节点为 13 个. 为了有助于对比研究,对同一数据又利用 Zahn 法进行了聚类分析,将其也分为 13 类. 图 5、图 6 是两种方法聚类结果对比,图中虚线背景是华北地区主要的构造断裂带,个体较大的符号是选取的控制节点,个体较小而样式相同的,则是属于该控制节点"控制"下的类别.

可以明显看出,对于此数据集来说,Zahn 法的聚类结果很不尽如人意:某些类别的点数过于稀疏(如图左下角);而图中符号为五角星的点,则一共有 427 个 ,占总数的 80%多. 显然这样的聚类结果是难以反映真实的丛聚模式的.

分析结果图 6,可以明显看出,其有以下几个优点:

(1) 各类间点数目分布较为合理;

(2) 几个带状、椭圆状等形状各异的地震密集区被各自很好地聚集,较好地体现了"空间上相互之间距离越近的对象越该归为一类"的聚类思想;

(3) 类别的分布和形状大多和主要断裂带分布、走向基本吻合,说明本方法能够较好地反映地震事件聚集分布的内在机理.

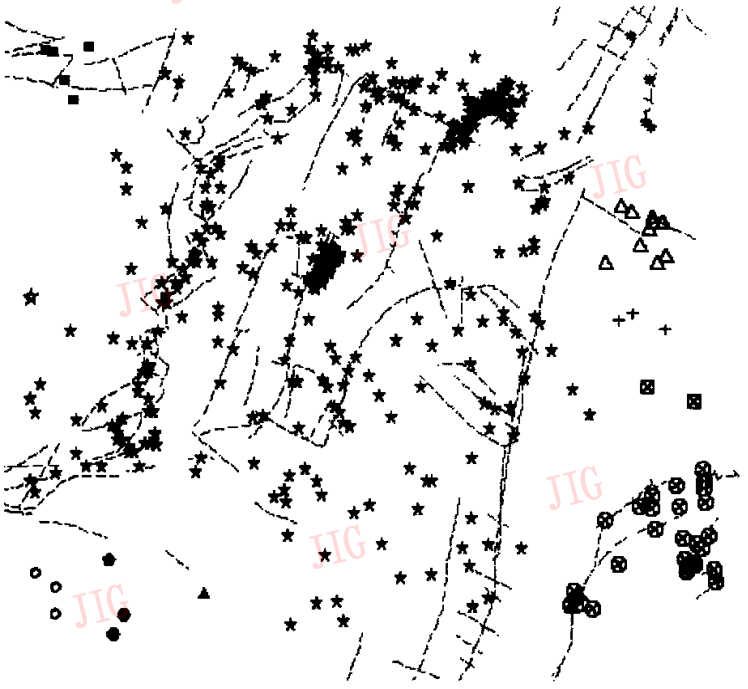


图 5　华北地震 Zahn 法的聚类结果

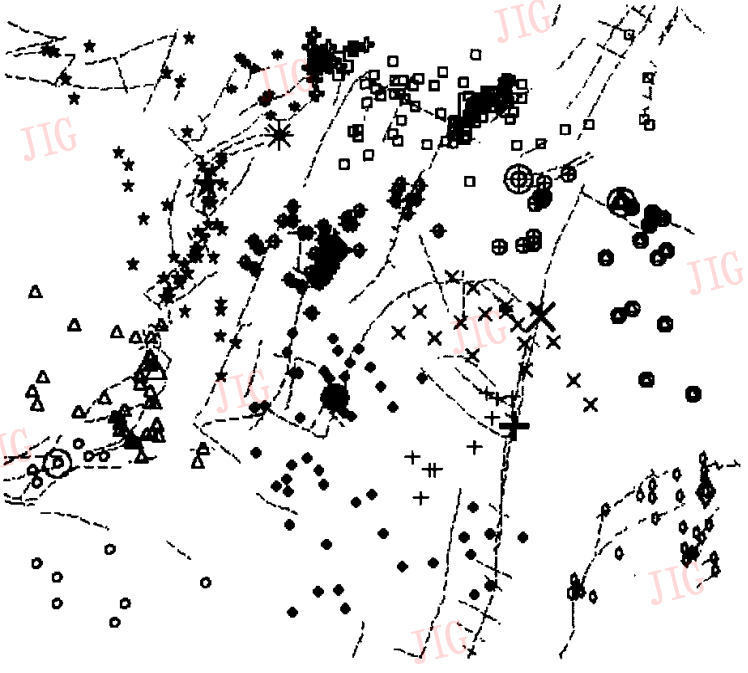


图 6　华北地震带控制节点的最小生成树方法的聚类结果

4 结 论

带控制节点的最小生成树聚类方法的主要特点是用高等级点控制最小生成树的分割,使各个控制节点的分布实现"分而治之"的效果.控制作用体现在:如果没有此控制,将依次打断"最不一致边"(这里把最不一致边定为当前生成树中的最长边),"次不一致边"等等;有了此控制,打断的就可能不是当前生成树中的最长边,而有可能是次长边,再次长边等等,总之它打断的是"能够将至少两个控制节点分割到两子树中去的最长边".这一方面在分割中加入了等级控制,另一方面体现了将相互最近的点尽可能地聚成一类的思想.实验证明,在某些情况下该方法能够较好地揭示数据分布的真实规律,且算法参数简单,较少人工干预,可操作性强.诚然,此方法可谓"初步"考虑了等级控制(两级).能否进一步丰富等级体系?如何在聚类中加入类似于此的先验知识,而又保持聚类"让数据本身说话"的特色,这是今后进一步努力的方向.此外,在控制节点的选取标准(如在某些情况下,可考虑选取局部密度极值点)、"不一致"边的选取标准、算法效率等问题上,尚需做进一步研究、改进,以满足挖掘海量空间数据复杂任务的需要.

参 考 文 献

1 Anil K Jain, Richard C Dubes. Algorithms for clustering data [M]. New Jersey: Prentice-Hall Inc, 1996: 55.

2 Anthony K H Tung, Jean Hou, Jiawei Han. Spatial clustering in the presence of obstacles [EB/OL]. URL: http://dbs.cs.sfu.ca, 2001-6-5.

3 吴开统,焦远碧,吕培苓等. 地震序列概论[M]. 北京:北京大学出版社,1990:2.

4 沈清,汤霖. 模式识别导论[M]. 长沙:国防科技大学出版社,1991:120~121.

5 严蔚敏,吴伟民. 数据结构[M]. 北京:清华大学出版社,1997:173~176.

6 王光荣,顾乃杰. 在消息传递并行机上的高效的最小生成树算法[J]. 软件学报,2000,11(7):889~898.

7 Charles T Zahn. Graph-theoretical methods for detecting and describing gestalt clusters [J]. IEEE Transactions on Computers, 1971, C-20(1): 68~86.

8 John Fehlauer, Bruce A Eisenstein. Structural editing by a point density function[J]. IEEE Transactions on Systems, Man and Cybernetics. 1978, smc-8(5): 362~370.

9 Koontz W L G, Narendra P M, Fukunaga K. A graph-theoretic approach to nonparametric cluster analysis [J]. IEEE Transactions on Computers. 1976, C-25(9): 936~944.

10 Rhchiro Mizoguchi, Masamichi Shimura. A nonparametric algorithm for detecting clusters using hierarchical structure[J]. IEEEE Transactions on Pattern Analysis and Machine Intelligence. 1980, Paml-2(4): 292~300.

11 傅征祥. 中国大陆地震活动性力学研究[M]. 北京:地震出版社,1997:5~7.

12 国家地震局. 中国地震烈度区划图(1990)概论[M]. 北京:地震出版社,1996:26~27.

13 裴韬. 中国及邻区大型地震数据库时空特征分析及其方法研究[博士后出站报告][R]. 北京:中科院地理所,2000:23.

汪 闽 1975 年生,中科院资源与环境信息系统国家重点实验室博士生.主要研究方向为空间数据挖掘与知识发现、空间数据聚类、分类算法研究等.

周成虎 1964 年生,研究员,博士生导师.研究兴趣包括地理信息分析与应用模型、遥感影像地学理解与分析、空间数据挖掘与知识发现等.

裴 韬 1972 年生,1998 年获中国地质大学(武汉)博士学位,现为中科院资源与环境信息系统国家重点实验室副研究员.主要从事空间数据挖掘及软件编制、地理信息系统应用等方面的研究工作.

韩志军 1970 年生,2000 年获中国地质大学地球信息与探测技术专业博士学位,现于资源与环境信息系统国家重点实验室进行博士后研究工作.主要研究方向地学信息系统与数据挖掘.

秦承志 1977 年生,现为中国科学院资源与环境信息系统国家重点实验室博士研究生.研究方向为空间数据挖掘与可视化分析.

秦 强 1978 年生,中科院地理与资源研究所 99 级在读硕士生.研究方向为地理信息系统、空间数据挖掘、多重分形研究.