引用格式：颜金彪,郑文武,段晓旗,等.改进的最小生成树自适应空间点聚类算法[J].地球信息科学学报,2018,20(7):887-894. [Yan J B, Zheng W W, Dan X Q, et al. Improved adaptive spatial points clustering algorithm based on the Minimum Spanning Tree[J]. Journal of Geo-information Science, 2018,20(7):887-894.] **DOI:**10.12082/dqxxkx.2018.180081

改进的最小生成树自适应空间点聚类算法

颜金彪[1,2],郑文武[1,2],段晓旗[1,2*],邓运员[1,2],郭元军[1,2],胡　最[1,2]

1. 传统村镇文化数字化保护与创意利用技术国家地方联合工程实验室, 衡阳 421002；
2. 湖南省古村古镇文化遗产数字化传承协同创新中心, 衡阳 421002

Improved Adaptive Spatial Points Clustering Algorithm Based on Minimum Spanning Tree

YAN Jinbiao[1,2], ZHENG Wenwu[1,2], DUAN Xiaoqi[1,2*], DENG Yunyuan[1,2], GUO Yuanjun[1,2], HU Zui[1,2]

1. National-Local Joint Engineering Laboratory on Digital Preservation and Innovative Technologies for the Culture of Traditional Villages and Towns, Hengyang 421002, China; 2. Cooperative Innovation Center for Digitalization of Cultural Heritage in Traditional Villages and Towns, Hengyang 421002, China

Abstract: In this paper, we proposed an improved adaptive spatial point clustering algorithm based on minimum spanning tree (MSTAA in abbreviation) to solve the problems existed in the traditional clustering algorithms. The first problem of these classical clustering algorithms is that the noise edges are determined using the global invariant. Another one is that the initial clustering parameters such as edge weight tolerance, edge variation factor, the number of clusters and initial clustering centers are determined by the users empirically. Furthermore, these algorithms can't find the noise edges at the local level. Based on these problems mentioned above, the algorithm put forward in this article aims to overcome the influence of subjective factors by defining two clipping factors. These trimming factors do not need to be determined by the users and can be automatically obtained according to the statistical features of the side length in the minimum spanning tree. The detailed realization process is as follows. In the first place, the pruning operation on the minimum spanning tree from the global level is carried out, which can eliminate the noises in the global environment. After the first round of tailoring, the initial minimum spanning tree becomes sub-tree collections. In the second place, in order to eliminate the noises at the local level, the algorithm performs the second round of pruning operation by setting the adaptive local cutting factor in the light of the side length statistics of each sub-tree. After the above two rounds of cutting, the MSTAA algorithm will get the final clustering result. In order to validate the effectiveness of the algorithm, both a simulated data and a practical application are adopted. By comparing with 4 classical clustering algorithms (k-means, DBSCAN, SEMST, HEMST), we find that the improved algorithm presented in this paper could find clusters of arbitrary shape and density in the environment where no one provides experience parameters. At the same time, not only does the MSTAA algorithm eliminate the obvious global noise points, but also it can distinguish noise points at the local environment so as to ensure a high similarity degree of cluster point set. All of the features of the MSTAA algorithm mentioned above make it possible to automatically mine hidden information behind spatial point data.

收稿日期：2018-01-30；**修回日期**：2018-04-15.

基金项目：国家自然科学基金项目(41471118、41771150、41771188)；衡阳师范学院青年项目(16A01、17A02)。 [**Foundation items:** National Natural Science Foundation of China, No.41471118, 41771150, 41771188; Youth Project of Hengyang Normal university, No.16A01,17A02.]

作者简介：颜金彪(1987-)，男，湖南衡阳人，硕士生，主要从事时空数据挖掘研究。E-mail: 715829216@qq.com

***通讯作者**：段晓旗(1990-)，男，山东泰安人，硕士，主要从事地图综合研究。E-mail: 201997125@qq.com

Key words: Minimum Spanning Tree; global clipping; local clipping; adaptive; clustering
***Corresponding author:** DUAN Xiaoqi, E-mail: 201997125@qq.com

摘要：针对传统的最小生成树聚类算法存在使用全局不变阈值确定噪声边，聚类需要用户根据经验确定初始化聚类参数，如“边权值倍数容差”，“边长变化因子”等，聚类不能发现局部噪声的问题，本文提出了一种改进的最小生成树自适应空间点聚类算法。该算法在无需用户输入参数的前提下，克服主观因素的影响，根据最小生成树边长的数理统计特征定义裁剪因子。算法首先从宏观层面对最小生成树进行首轮删枝操作，消除全局环境下的噪声边，进而根据各子树的边长统计情况，自适应设定局部裁剪因子，进行第二轮删枝操作，消除局部环境下的噪声边。最后，采用1个模拟数据和1个实际应用验证算法的有效性，结果表明本文提出的改进算法在无需人为提供经验参数的环境下能够发现任意形状、不同密度的簇，能够准确的识别出空间点中的噪声数据，从而能够实现空间点数据背后隐藏信息的自动挖掘。
关键词：最小生成树；全局裁剪；局部裁剪；自适应；聚类

1 引言

作为空间数据挖掘、知识发现的重要研究内容，聚类分析已经应用在国计民生的各个方面，如交通事故热点分析[1]、生物种群分布[2]、医疗卫生[3]、犯罪热点分析[4]、地震监测分析[5]等。现有的空间聚类算法主要可以分为5类[6]：① 基于划分的算法，如K-means[7]，AP（Affinity Propagation）[8-9]；② 基于层次的算法，如CURE[10]；③ 基于密度的算法，如DBSCAN[11]；④ 基于图论的算法，如MST；⑤ 基于格网的算法，如STING[12]。在众多的聚类算法当中，基于最小生成树的聚类算法已经被广泛的研究。

汪闽等[13]提出了一种带控制节点的最小生成树聚类算法，通过添加控制点与设定优先级的方法来优化聚类结果，但是需要用户设定控制点的个数及位置；Grygorash等[14]在使用全局不变阈值确定最小生成树的噪声边，但当簇密度相差较大时，该算法聚类结果不理想；王小乐等[15]采用图论和层次聚类算法的思想，从最小生成树中的一条边开始，通过一个控制参数，逐步合并该边连接的两个部分，该算法能够发现任意形状，不同密度的聚类，但是需要用户根据经验确定“边权值倍数容差”；邓敏等[6]提出了MSTLSC算法，该算法针对不同的空间局部分布密度，分别生成一系列子树，但需要用户设定“边长变化因子”。

上述基于最小生成树的聚类算法大都需要一定数量的初始化参数，徐晨铠等[16]提出了改进的最小生成树自适应分层聚类算法，该算法根据最近邻关系，自动为每个簇设定独立的裁剪阈值，使之适应分布密度相差较大的情况，并能够自动确定聚类数目，但是该算法不能保证所选链路是最小开销的链路[17]；贾瑞玉等[18]提出了一种基于最小生成树的层次K-means算法，该算法改进了K-means初始聚类中心选择和聚类数目确定的问题，但是该算法在确定聚类数目的时候，采用了与经验阈值比较的方法，影响了聚类的准确率，同时该算法不能发现噪声点；朱利等[19]基于最小生成树提出一种自适应确定聚类数目的算法，但实际是利用边阈值代替聚类个数的人为输入。

针对上述问题，本文提出了一种改进的最小生成树自适应聚类算法（Adaptive Clustering algorithm based on Minimum Spanning Tree，MSTAA）。一方面，该算法要解决聚类需要输入初始化参数的问题，如初始聚类数目、聚类中心、“边权值倍数容差”、“边长变化因子”；另一方面，要求该算法能够发现任意形状，在包含大量噪声的空间点中识别出不同密度的簇。

2 MSTAA算法原理与基本概念

根据视觉认知的Gestalt规则[20-21]，人们在认识事物的过程中总是先从事物的整体进行认知，而后才从事物的部分进行认知。基于这种朴素的认知思想，本文首先从宏观层面，根据最小生成树边长的统计特征，定义宏观裁剪因子对最小生成树进行首轮删枝操作，继而从局部出发，定义局部裁剪因子，在首轮删枝操作后的基础上进行第二轮删枝操作，从而获得最终的聚类结果。为便于描述，下面给出一些必要的定义：

定义1：设$G=\{V,E\}$表示最小生成树，V代表顶点，E代表边。当存在图$subG=\{subV,subE\}$，其中$subV\subseteq V,subE\subseteq E$且$subG$保持连通，则认为$subG$为$G$的1棵子树。

定义2：对于最小生成树G，求取其平均边长

$mstMeanE$，如式(1)所示。

$$mstMeanE=\frac{\sum_{i=1}^{N-1}edge_i}{N-1} \tag{1}$$

式中：$edge_i$为最小生成树G中第i条边长(本文采用欧氏距离，下同)；N为最小生成树G中顶点数(下同)。

定义3：对于最小生成树G，求取边长的中误差$mstVarE$，如式(2)所示：

$$mstVarE=\sqrt{\frac{\sum_{i=1}^{N-1}(edge_i-mstMeanE)^2}{N-2}} \tag{2}$$

定义4：对于子树$subG_q$，定义子树$subG_q$的平均边长为$sonMstMeanE_q$，如式(3)所示。

$$sonMstMeanE_q=\frac{\sum_{j=1}^{m-1}edge_j}{m-1} \tag{3}$$

式中：m为子树$subG_q$中顶点数(下同)；$edge_j$表示子树$subG_q$中边长(下同)；q表示子树的编号(下同)。

定义5：对于子树$subG_q$，求取子树$subG_q$的边长中误差$sonMstVarE_q$，如式(4)所示。

$$sonMstVarE_q=\sqrt{\frac{\sum_{j=1}^{m-1}(edge_j-\text{son}\,MstMeanE_q)^2}{m-2}},m\geqslant 3 \tag{4}$$

3　算法的设计与分析

3.1　算法流程

MSTAA算法使用逐步标记和分而治之的策略实现空间点的聚类，从宏观到局部两个层次逐步打断最小生成树中的长边，整个流程如图1所示。

3.2　算法实现

3.2.1　数据预处理

该过程的主要任务是由空间点集得到最小生成树G，主要过程如下所述：① 数据清理。由于空间点可能存在压盖的情况，影响后续Delaunay三角网的生成，因此将压盖位置的空间点保留其中任意一个；② 将清理后的空间点生成Delaunay三角网；③ 根据步骤②生成的Delaunay三角网，利用prime算法生成最小生成树G。

3.2.2　全局裁剪

根据数据预处理得到的最小生成树G，计算出平均边长及边长中误差，从而计算出全局裁剪值$mstCutV_i$，将其作为全局裁剪阈值，如式(5)所示。

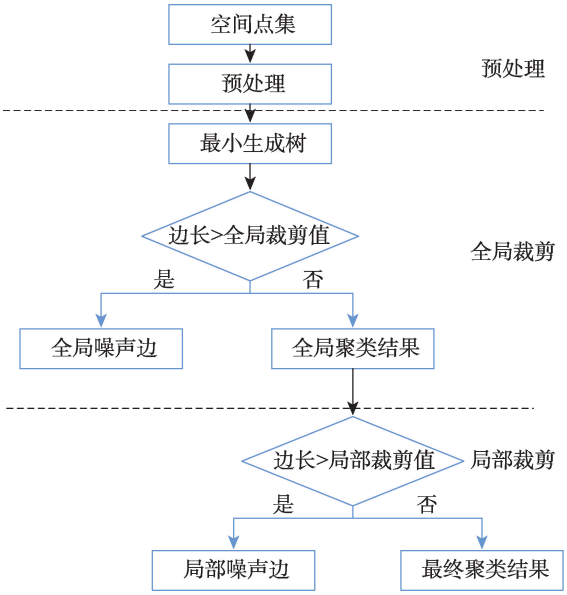


图1　MSTAA算法流程图

Fig. 1　The flow chart of MSTAA algorithm

$$mstCutV_i=mstMeanE+\alpha\times\frac{mstMeanE}{Edge_i} \tag{5}$$

式中：α称为调和因子，值愈低对长边愈加敏锐，反之愈加迟钝，具体计算如式(6)所示，这保证了$mstCutV_i$不至于太大，也不至于太小，经后续实例验证，比较符合实际。

$$\alpha=\frac{mstMeanE}{2\times mstVarE} \tag{6}$$

从式(5)-(6)可以看出，$mstCutV_i$具备自适应特征，对于长边，则阈值愈低，那么该边更易断链，反之对于短边，阈值将变大，断链的可能性降低。

算法具体步骤如下：

(1)计算最小生成树G的平均边长$mstMeanE$与中误差$mstVarE$；

(2)选取最小生成树G中一条边$edge_i$；

(3)计算边$edge_i$对应的全局裁剪阈值$mstCutV_i$；

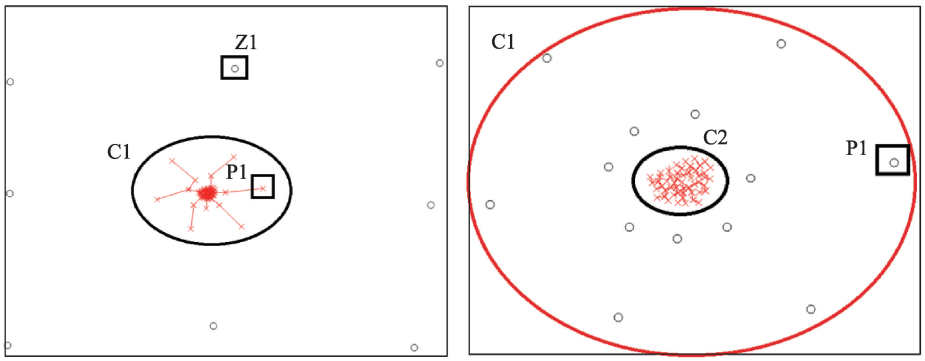
(4)if $edge_i>mstCutV_i$，则将$edge_i$断开，标记$edge_i$；

(5)if $edge_i\leqslant mstCutV_i$，则$edge_i$保持不变，标记$edge_i$，继续执行；

(6)重复步骤(2)-(5)，直至最小生成树G中每条边均被标记；

(7)经过步骤(1)-(6)，最小生成树G将被剖分成多棵子树。

通过移除宏观噪声点，如图2(a)中的Z1点等，得到图2(a)中的C1簇。经过全局裁剪之后，可以发现C1簇内部仍然存在不一致点，如图2(a)中P1点等。由此可以发现，仅通过移除全局噪声边，仍



(a) 移除宏观的结果　　(b) 移除局部噪声边的结果

图2 剔除最小生成树中的噪声边

Fig. 2 Remove the noises in the minimum spanning tree

不能精确识别空间点簇，需要在此基础进一步裁剪，消除局部噪声边对聚类结果的影响。

3.2.3 局部裁剪

经过全局裁剪之后，原始的最小生成树G变成子树集合*sonMst*。算法首先从子树集合中找到最长边，同时搜索出最长边所在的子树$sonMst_q$，计算子树$sonMst_q$中边长的均值与中误差，从而得到针对子树$sonMst_q$对应的局部裁剪阈值$sonMstCutV_q$，具体如式(7)所示。

$$sonMstCutV_q = sonMstMeanE_q + \beta \times sonMstVarE_q \quad (7)$$

式中：$sonMstMeanE_q$与$sonMstVarE_q$分别对应子树$sonMst_q$的边长平均值和中误差。

β称为聚类敏感因子，值愈低对噪声愈加敏锐，越大对噪声反应愈加迟钝，β取值一般介于(2-3)之间，本文实验中取值为3。尽管MSTAA对于聚类敏感因子β的选取凭借的是经验，但是实际是借用了统计学上的显著性概念，因此具有一定的数学意义。当子树$sonMst_q$的边长值满足正态分布，聚类敏感因子取值为3时，如果子树集合中最长边大于其所属子树$sonMst_q$的局部裁剪阈值时，将以近99.7%的概率认为该最长边属于局部噪声边。

由上可见，$sonMstCutV_q$同样具备自适应特征，将其作为最长边是否断链的判断标准。局部裁剪的具体实现如下步骤所述：

(1)从子树集合*sonMst*中找出最长边max，得到最长边max所属的子树$sonMst_q$，然后用式(3)-(4)计算出子树$sonMst_q$的平均边长与中误差；

(2)用式(7)计算出$sonMstCutV_q$值；

(3)if max>$sonMstCutV_q$，则将最长边所在的链断开，在子树集合*sonMst*中标记最长边的长度为0，同时将该棵子树在断链的位置进一步剖分为2棵子树，并将剖分结果更新到子树集合*sonMst*当中；

(4)if max≤$sonMstCutV_q$，即最长边短于局部裁剪阈值，则可认为该棵子树中已不存在局部噪声边，该棵子树不需要进一步的剖分，将其作为最终聚类结果的一个子簇C_q，同时在子树集合*sonMst*中标记$sonMst_q$子树中所有边长值为0；

(5)重复执行步骤(1)-(4)，直至子树集合*sonMst*中所有边长全标记为0；

(6)输出最终聚类结果。

通过移除局部噪声点，如图2(b)中的P1点等，C1簇精简为图2(b)中的C2簇。可以明显发现C2簇点密度高，将其作为最终聚类结果的一个子簇较为合理。

3.3 算法的时间复杂度分析

本文中MSTAA算法在没有建立索引的前提下分析其时间复杂度。设空间点的数目为n，算法的时间主要集中在文中3.2节中的过程。其中，数据预处理中Delaunay三角网的建立采用分治算法，极端情况下需要花费$o(n\lg n)$[22]，最小生成树采用prime算法需要消耗$o(n^2)$[23]，全局裁剪需要花费$o(n-1)$，局部裁剪在极端情况下(即n个点呈现均匀分布)需要花费$o(n^2\cdot(n-1)/2)$，因此在最为极端的情况下，MSTAA算法的时间复杂度为$o(n\lg n+n^2+n-1+n^2\cdot(n-1)/2)\approx o(n^3)$。

4 算法实例分析

本文设计了2个实验，其中包含1个模拟数据集[21]及1个真实的空间数据集，图3采用不同颜色

来区分不同的点簇。为了验证MSTAA算法的有效性，将本文算法与K-means、DBSCAN、传统的最小生成树聚类算法进行对比分析。

4.1 MSTAA 与 K- means、DBSCAN、SEMST、HEMST 的对比研究

如图3(a)的数据集，理想聚类结果分为6个簇，该数据包含不同密度、不同形状的簇，且点簇的周边存在不同层次下的噪声点，聚类结果采用“o”表示噪声点，“×”表示聚类点。

K-means需要初始化聚类数目k及初始化聚类中心，通过不断优化平方误差准则，获得对数据集的k个划分。本实验中将k设定为6，同时人工获取6个理想的点作为其初始聚类中心，聚类效果如图3(b)。

DBSCAN聚类算法基本思想在于采用一定邻域范围内包含的空间实体的最小数目定义空间密度的概念，并通过不断生长的高密度区进行空间聚类操作，本实验中通过人工探索，将DBSCAN算法的邻域半径eps设置为6，密度阈值minpts设置为4，聚类效果如图3(c)所示。

SEMST[24]算法首先将最小生成树的边长按照从大到小排列，根据用户设定的初始化聚类数目k，从中剔除k−1条长度排在前列的边。本文中将k值设定为6，聚类效果如图3(d)所示。

HEMST[14]算法属于层次聚类算法的一种，算法首先计算最小生成树的平均边长与中误差，得到裁剪阈值ω，如式(8)所示。

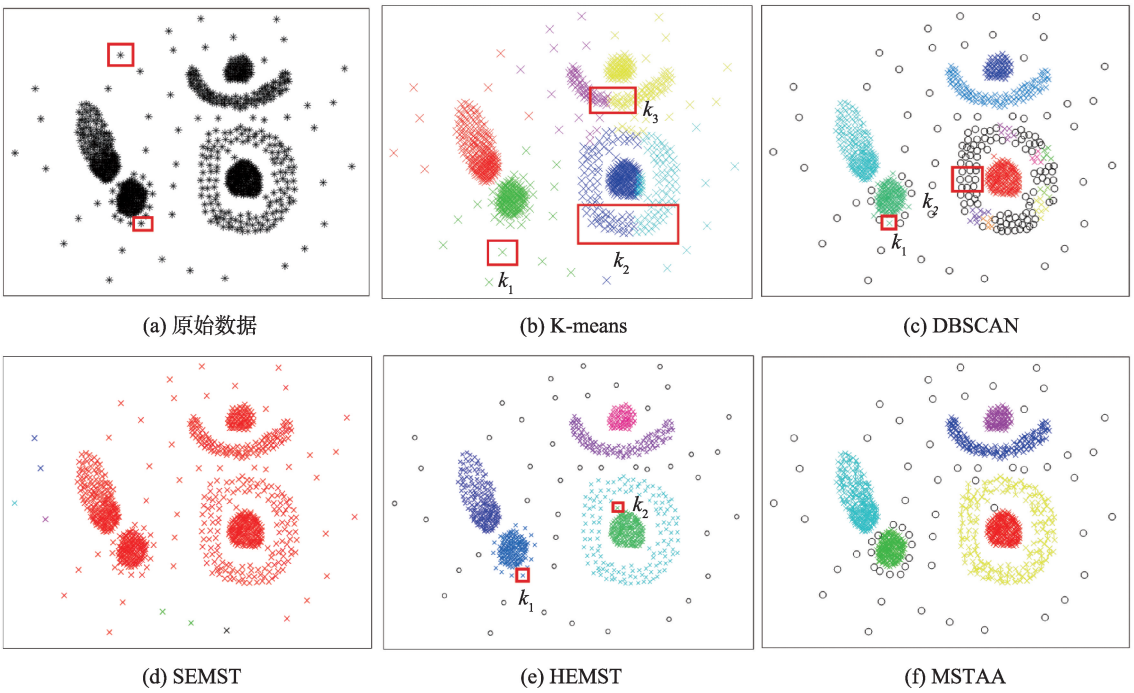
$$\omega=\bar{\omega}+\sigma \tag{8}$$

式中：$\bar{\omega}$为最小生成树的平均边长，σ为边长中误差。当最小生成树中边长大于ω时，将该边打断。如果此时生成的簇数量大于用户设定的簇数k，算法将距离较近的点簇合并，直至最终簇数等于设定值；如果此时簇数少于用户设定的数量k，算法将最小生成树中长边依次打断，直至簇的数量满足用户设定值，算法停止。本实验中将HEMST算法的初始化聚类数目也设定为6，聚类结果如图3(e)所示。MSTAA算法聚类结果如图3(f)所示。

通过分析图3可知：

(1)经典的K-means聚类算法不能发现噪声点(图3(b)中k_1等)，只能够发现凸形的簇，难以发现条带状的簇(图3(b)中k_2、k_3)。这是由于该算法核心在于通过一定的划分准则将整个数据集分成k个划分，同时将空间距离作为评判点集之间空间相似度的唯一指标，因此该算法不能发现噪声点且难以发现条带形状的簇。

(2)经典的DBSCAN算法在整个聚类过程中保持恒定的邻域半径及密度阈值，尽管DBSCAN算法能够识别出噪声点，但对于空间点分布密度存在变



(a) 原始数据　(b) K-means　(c) DBSCAN

(d) SEMST　(e) HEMST　(f) MSTAA

图3　算法聚类结果

Fig. 3　Clustering results

化的情况，算法容易发生误判，如此例中将部分密度稍低的空间点识别为噪声（如图3(c)k_2等），同时将局部噪声点归并到聚类结果中（如图3(c)k_1等）；

（3）从图3(d)中可以看出，SEMST算法得到的6类结果与理想聚类结果相差甚大，完全不能满足需求。这是由于该算法仅从最小生成树中依次裁剪了k−1条最长边，针对最小生成树中噪声边的数量远大于k−1的情况（图3(a)），将导致该算法完全失效。

（4）HEMST算法能够识别出噪声点，能够发现不同密度、不同形状的簇，但由于该算法整个聚类的过程中保持相同的裁剪阈值（均值+中误差），而局部环境下的噪声边长易低于全局裁剪阈值，因此难以发现局部环境下的噪声边，如图3(e)中的k_1、k_2区域。

（5）MSTAA算法首先采用全局裁剪将明显的噪声点全部予以剔除，之后根据子树边长的数理统计特征自适应设定针对当前子树的局部裁剪阈值，从而能够剔除局部环境下的噪声点。从图3(f)中可以发现，得到的聚类结果与理想聚类结果吻合得较好。

4.2 实际应用

为了验证MSTAA算法的实用性，将其应用于地震活跃带的识别研究。本例中数据介于21°~35 °N，94°~107 °E之间，研究区域内由北往南分布着天水兰州、河西走廊、康定甘孜、安宁河谷、西藏察隅、滇西、滇东、腾冲−廊沧8个地震带。实验选取2008−2017年间震级M≥5的地震，共计111个震中点，如图4所示，数据均来源于中国地震信息网。

MSTAA算法的聚类结果如图4所示，采用“o”表示聚类结果，“×”表示孤立的地震震中，数据共分成23个簇，其中包括6个线形簇（L1−L6），4个圆形簇（C1−C4），13个小簇（S1−S13），同时算法识别出17个孤立震中，表明该算法在无需用户提供初始聚类数目、初始聚类中心等参数的情况下能够发现任意形状（线形、圆形）、不同密度（如C1、C2）的簇，且能准确识别出噪声点。

地震震中通常沿着活跃断层发生[25]，观测的震中位置一般沿着这些断层聚集起来。从图5可以看出：① 河西走廊地震带周边由北往南共计分布着8个簇，表明该地震带在过去十年之间非常活跃。其中7个簇为西南至东北方向，与河西走廊地震带的方向基本保持一致；② 滇西、腾冲廊沧地震带周边各分布着3个簇，表明其于过去十年之内较为活

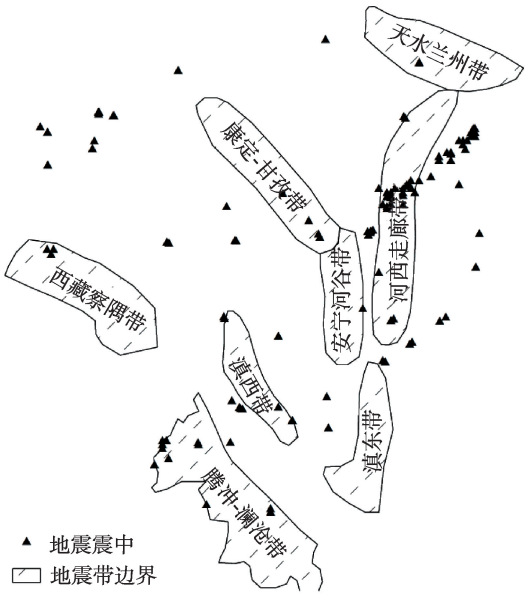


图4 地震震中位置

Fig. 4 The epicenter of the earthquake

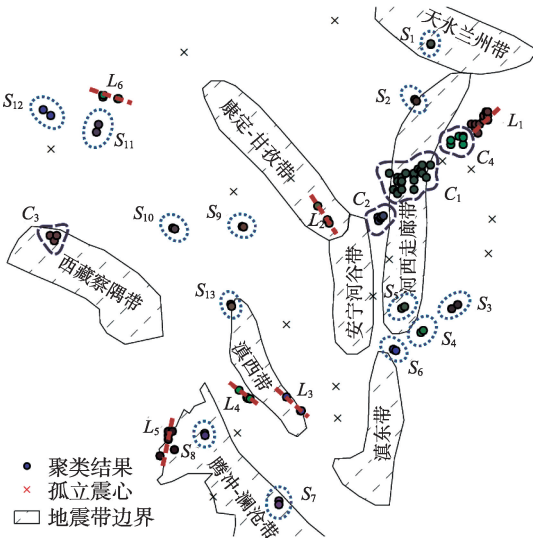


图5 MSTAA对地震数据的聚类结果

Fig. 5 Spatial clustering results on seismic data by MSTAA

跃。其中L3、L4线形簇的方向为西北至东南，与滇西地震带的方向较为一致；③ 天水兰州、康定甘孜、西藏察隅、滇东地震带周边各分布着1个簇，意味该地震带在2008−2017年为一般活跃。其中L2簇与康定甘孜地震带的方向保持一致，均为西北至东南方向，天水兰州与西藏察隅地震带近十年发生的地震震中聚类结果不具有明显的方向性；④ 安宁河谷地震带在近10年没有发现带有聚集特征的地震震中簇发生，表明该地震带在近10年处于不活跃状态；⑤ 在康定甘孜与西藏察隅地震带之间分布着

1个线形簇与4个小簇，其中线形簇L6与S9、S10、S12的方向均为西北至东南方向，与康定甘孜、西藏察隅地震带的方向吻合得较好，具体原因有待进一步的调查研究。

5 结论

本文提出了一种改进的最小生成树自适应空间点聚类算法，从整体到局部2个层次实现空间点的聚类，以期能够在无人工干预的影响下得到准确的聚类结果，通过与4种经典的聚类算法对比和实际应用后发现：① MSTAA算法具备良好的自适应特征，用户不需指定初始聚类数目、初始聚类中心、搜索邻域范围大小、“边长变化因子”，“边权值倍数容差”，降低用户使用聚类算法的知识储备要求；② 算法不仅能剔除宏观层面的噪声点，同时可以识别子树的局部噪声点，从而确保簇内点集的高度相似性；③ 在局部裁剪时，如果子树的边长近似于正态分布，那么MSTAA对于聚类敏感因子β的选取实际借用了统计学上的显著性概念，具有一定的数学意义[26]，但反之则聚类结果的准确性将受到影响；④ 算法能够进行不同密度、任意形状及任意聚类数目的聚类。

研究空间点群的聚类方法，不仅为知识发现、模式识别提供技术支持，还可以为数据挖掘的发展提供重要参考。本文主要是以点目标的空间位置进行研究，未涉及到其它属性信息，在后续的研究中，将考虑融合空间位置与其它非空间的属性数据，以便于发现更深层次的信息；同时，建立空间索引，提高算法的执行效率，降低算法的时间复杂度；扩大实验范围，将改进后的算法应用于中国传统村落的分类研究当中。

参考文献(References):

[1] Eckley D C, Curtin K M. Evaluating the spatiotemporal clustering of traffic incidents[J]. Computers, Environment and Urban Systems, 2013,37:70-81.

[2] Dale M R T. Spatial pattern analysis in plant ecology[M]. Cambridge: Cambridge university press, 2000.

[3] Miller H J, Han J Geographic. Data mining and knowledge discovery, second edition[M]. New York: CRC Press, 2009.

[4] Grubesic T H, Mack E A. Spatio-temporal interaction of urban crime[J]. Journal of Quantitative Criminology, 2008,24(3):285-306.

[5] Wang M, Wang A, Li A. Mining spatial-temporal clusters from geo-databases[J]. Advanced Data Mining and Applications, 2006:263-270.

[6] 邓敏,刘启亮,李光强,等.一种基于似最小生成树的空间聚类算法[J].武汉大学学报·信息科学版,2010,35(11):1360-1364. [Dend M, Liu Q L, Li G Q. A spatial clustering algorithm based on minimum spanning tree-like[J]. Geomatics and Information Science of Wuhan University, 2010,35(11):1360-1364.]

[7] Macqueen J. Some methods for classification and analysis of multivariate observations[C]. Proceeding of Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[8] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007,315(5814):972-976.

[9] 孙磊磊. AP聚类算法研究及其在电子病历挖掘中的应用[D].大连:大连理工大学,2017. [Sun L L. Study on affinity propagation clustering algorithm and its application in mining electronic medical records[D]. Dalian: Dalian University of Technology, 2017.]

[10] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases[J]. International System, 2001,26(1)35-38.

[11] Ester M. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceeding of International Conference on Knowledg Discovery and Data Mining, Portland, 1996:226-231.

[12] Wang W, Yang J, Muntz R R. STING: A statistical information grid approach to spatial data mining[C]. International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. 1997:186-195.

[13] 汪闽,周成虎,裴韬,等.一种带控制节点的最小生成树聚类方法[J].中国图象图形学报,2002,7(8):765-770. [Wang Min, Zhou C H, Pei T, et al. A MST based clustering method with controlling vertexes[J]. Journal of Image and Graphics, 2002,7(8):765-770.]

[14] Grygorash O, Zhou Y, Jorgensen Z. Minimum Spanning Tree based clustering algorithms[C]. IEEE International Conference on TOOLS with Artificial Intelligence, IEEE, 2006:73-81.

[15] 王小乐,刘青宝,陆昌辉,等.一种最小生成树聚类算法[J].小型微型计算机系统,2009,30(5):877-882. [Wang X L, Liu Q B, Lu C H. Minimum Spanning Tree clustering algorithm[J]. Journal of Chinese Computer Systems, 2009, 30(5):877-882.]

[16] 徐晨凯,高茂庭.改进的最小生成树自适应分层聚类算法[J].计算机工程与应用,2014,50(22):149-153. [Xu C K, Gao M T. Improved adaptive hierarchical clustering algo-

rithm based on minimum spanning tree[J]. Computer Engineering and Applications, 2014,50(22):149-153.]

[17] 邱雪松,蔺艳斐,邵苏杰,等.一种面向智能电网数据采集的传感器聚合布局构造算法[J].电子与信息学报,2015,37(10):2411-2417. [Qiu X S, Lin Y F, Shao S J, et al. Sensor aggregation distribution construction algorithm for smart grid data collection system[J]. Journal of Electronics & Information Technology, 2015,37(10):2411:2417.]

[18] 贾瑞玉,李振.基于最小生成树的层次K-means聚类算法[J].微电子学与计算机, 2016,33(3):86-88. [Jia R Y, Li Z. The level of K-means clustering algorithm based on the Minimum Spanning Tree[J]. Microelectronics and Computer, 2016,33(3):86-88.]

[19] 朱利,邱媛媛,于帅,等.一种基于快速k-近邻的最小生成树离群检测方法[J].计算机学报,2017,40(12):2856-2870. [Zhu L, Qiu Y Y, Yu S, et al. A fast kNN-Based MST outlier detection method[J]. Chinese Journal of Computers, 2017,40(12):2856-2870.]

[20] Humphrey G. The psychology of the gestalt[J]. Journal of Educational Psychology, 1924,15(7):401-412.

[21] Deng M, Liu Q, Cheng T, et al. An adaptive spatial clustering algorithm based on delaunay triangulation[J]. Computers Environment & Urban Systems, 2011,35(4):320-332.

[22] 余杰,吕品,郑昌文. Delaunay三角网构建方法比较研究[J].中国图象图形学报,2010,15(8):1158-1167. [Yu J, Lu P, Zheng C W. A comparative research on methods of Delaunay triangulation[J]. Journal of Image and Graphics, 2010,15(8):1158-1167.]

[23] Graham R L, Hell P. On the history of the minimum Sspanning tree problem[J]. Annals of the History of Computing, 1985,7(1):43-57.

[24] Asano T, Bhattacharya B, Keil M, et al. Clustering algorithms based on minimum and maximum spanning trees [C]. Symposium on Computational Geometry, DataBase systems and Logic Programming, 1988:252-257.

[25] Xu X, Ester M, Kriegel H P, et al. A Distribution-Based clustering algorithm for mining in large spatial databases [C]. International Conference on Data Engineering, 1998. Proceedings, IEEE, 1998:324-331.

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