

基于烈度衰减椭圆阈值空间散点聚类研究

——以 12322 灾情速报平台为例*

程显洲^{1,2}, 肖兰喜^{1,3}, 董翔³, 李铂³

(1. 中国地震局兰州地震研究所, 甘肃 兰州 730000; 2. 上海市地震局, 上海 250000;
3. 山东省地震局, 山东 济南 250014)

摘要: 采用了以震中为中心的烈度椭圆的类间椭圆阈值, 通过对 12322 灾情速报平台汶川 8.0 地震的灾情速报信息的空间经纬度数据进行了空间散点聚类分析, 根据此阈值来确定实际的聚类次数, 通过聚类来剔除不合理的数据, 使得到的最终散点信息真正显示出实际的震情。

关键词: 空间聚类; 椭圆阈值; 烈度衰减; 灾情速报

中图分类号: X43 **文献标志码:** A **文章编号:** 1000-811X(2013)04-0205-04

聚类分析是研究对事物进行分类的一种多元统计方法。聚类问题在科学研究、生产实践, 社会生活中到处存在。空间聚类作为聚类分析的一个研究方向, 是指将空间数据集中的对象分成由相似对象组成的类。许多聚类算法基于欧几里得或者曼哈顿距离度量来决定聚类。基于这样的距离度量的算法趋向于发现具有相近尺度和密度的球状簇。Rose^[1]等提出了基于模拟退火算法的类间阈值聚类, 模拟退火法收敛过慢, 而且对于数据的初始值依然很敏感, 效果不是很好。张慧哲等^[2]提出了一种改进的 FCM 算法, 通过人为的设置初始化中心和加入数据对个聚类中心之间的距离最聚类影响的系数, 效果比较明显, 但是人为的干预使得聚类系统的可靠性降低。于剑等^[3]对最佳聚类数进行分析得出最佳的聚类数上限为 $C_{\max} \leq \sqrt{n}$, 其中 n 为数据个数。在此基础上, 谢宇等^[4]利用类间间距阈值搜索聚类数的方法确定最佳聚类数, 得到很好的效果。孙才质等^[5]在 J C Bezdek 等人的基础上根据方差分析理论, 提出了应用混合 F 统计量来确定最佳聚类数并应用模糊划分熵来验证最佳分类数的正确。但是对于空间点的统计分析并不能完全概括数据点的空间分布结构。在实际应用中由于空间中散点聚类不规则性类间阈值不一定为规则的圆形, 比如基于断裂的地震烈度衰减有可能是呈椭圆形状的分布并且具有明显的方向性。而普通的圆形阈值的 c 均值聚类不能胜任此工作, 给我们的空间散点聚类提出了难题。本文通过包含烈度衰减信息的椭圆形的类间阈值确定聚类数的方法对地震速报短信实

验平台的 73 条空间散点经纬度信息进行分段模糊 c 均值聚类, 对灾情信息数据的奇异值进行剔除与修改, 通过对比实际烈度, 表明此方法有效性和可行性。

1 实验平台

为确保发生破坏性地震时, 各项应急工作高效、有序地进行, 最大限度地减少人员伤亡和财产损失, 国务院抗震救灾指挥部建立了基于防震减灾公益服务 12322 号码的地震灾情短信息速报试验平台, 将地震现场的灾情上报人员的手机或其他无线终端与后台指挥部建立一种互动联系, 并规定了灾情信息短信的格式 $\text{Dot}(x, y, z)$, 其中 x 为离散灾情点的经度, y 为离散灾情点的纬度, z 为灾情点属性信息(表 1)。

z 的值分为四个等级: 1 为无震感或震感轻; 2 为震感强无破坏; 3 为有破坏无伤亡; 4 为有伤亡破坏重。帅向华等^[6]基于 Delaunay 三角网的最邻近点插值法对离散点短信上报的灾情信息进行分析并发展了适合基于地震灾情离散点的网格化分析算法。但由于地震灾情速报员为个体点, 在灾情上报的时候看到的或者得到的信息是有限的, 往往与真实情况出入很大, 不能准确反映地震的破坏程度, 因此, 为较为准确地表达地震灾情分布, 需要根据上述特点对灾情信息数据进行修改, 并采用适合于灾情采样点的插值和空间拟合算法。为此本文致力于对汶川地震中提取的有效短信的聚类分析以便更好的

* 来稿日期: 2013-04-06 修回日期: 2013-06-06
基金项目: 国家科技支撑计划“地震灾情服务及应急决策支撑平台研究”(2012BAK15B06)
作者简介: 程显洲(1987-), 男, 山东青岛人, 硕士研究生, 主要研究方向地震动衰减关系。E-mail: mn1472008@163.com

表 1		汶川地震提取的有效短信示例 ^[6]					
序号	地震编号	经度 / (°)	纬度 / (°)	灾情代码	短信内容		
1	200805121428	101. 17	32. 1904	3	200805121428	3	阿坝县
2	200805121428	110. 1063	27. 1775	2	200805121428	2	桥头溪乡
3	200805121428	109. 1853	28. 1427	2	200805121428	2	白岩乡
4	200805121428	109. 1641	28. 1714	2	200805121428	2	保靖县
5	200805121428	108. 1938	34. 1259	2	200805121428	2	西安市

反应地震的破坏状况。

2 椭圆类间距阈值 c 均值聚类法

2.1 模糊 c 均值聚类 (FCM)

在普通分类的基础上,J C Bezdek^[7] 引入了模糊集的概念首次提出了模糊 c 均值的分类方法 (FCM),他认为被样本集合中的每一个样本均以不同的隶属度隶属于某一类,而不同向量可以属于同一类。此后 J C Bezdek 等^[8] 人通过权重的分析和对模糊 c 均值算法 (FCM) 总结完善了这一算法。算法简介与步骤如下所示。

设一个样本数据集合: $Y = \{y_1, y_2, \cdots, y_N\}$; y_k 是第 k 个特征向量; y_{kj} 是 y_k 的第 j 个特征。根据聚类数 $c(2 < c < N)$ 将 Y 分成 c 个子集(Y_1, Y_2, \cdots, Y_c)。其中子集必须满足如下条件:

$$Y_i \neq \phi, 1 \leq i \leq c; \tag{1}$$

$$Y_i \cap Y_j = \phi, i \neq j; \tag{2}$$

$$\bigcup_{i=1}^c Y_i = Y. \tag{3}$$

设一个 c 行 N 列的实数矩阵 $U, U = [u_{ik}]$ 作为为 Y_i 的矩阵形式满足下列三个条件:

$$u_i(y_k) = u_{ik} = \begin{cases} 1; y \in Y_i; \\ 0, y \notin Y_i. \end{cases} \tag{4}$$

$$\forall i \quad \sum_{i=1}^N u_{ik} > 0. \tag{5}$$

$$\forall k \quad \sum_{i=1}^N u_{ik} = 0. \tag{6}$$

用隶属度函数定义的最小二乘聚类残差函数表示为:

$$J_m(U, v) = \sum_{k=1}^N \sum_{i=1}^c (u_{ik})^m \|y_k - v_i\|_A^2. \tag{7}$$

式中: m 为权重指数; U 为子集; v 为聚类中心; A 为残差范数。基于 A 范数的 y_k 与 v_i 的距离平方可以表示为:

$$d_{ik}^2 = \|y_k - v_i\|_A^2 = (y_k - v_i)^T A (y_k - v_i). \tag{8}$$

对于大多数数据,将权重 m 取在 1.5 ~ 3.0 之间得到的结果会比较理想。聚类中心 v 在每次迭代后都会变化,其计算公式为:

$$\hat{v}_i = \sum_{k=1}^N (\hat{u}_{ik})^m y_k / \sum_{k=1}^N (\hat{u}_{ik})^m, \quad 1 \leq i \leq c; \tag{9}$$

$$\hat{u}_{ik} = \left(\sum_{j=1}^c \left(\frac{\hat{d}_{ik}}{\hat{d}_{jk}} \right)^{2/(m-1)} \right)^{-1}, \quad 1 \leq k \leq N; 1 \leq i \leq c. \tag{10}$$

FCM 算法就是通过最小二乘法迭代求出 J_m 最优化解,其过程如下:

Step1 设置聚类数 c 、变量权重值 m 、最小二乘范数 A 和残差容许值 e ,选择待聚类的矩阵 U 。

Step2 从 $k = 0, 1, \cdots, k_{\max}$ 开始根据式 (9) 计算聚类中心 $\hat{v}^{(k)}$ 。

Step3 从 $i = 1, 2, \cdots, c$ 开始根据公式 (10) 计算关系矩阵 $\hat{U}^{(k+1)} = [\hat{u}_{ik}^{(k+1)}]$ 。

Step4 计算 $\hat{U}^{(K+1)}$ 与 $\hat{U}^{(k)}$ 之间的距离,如果大于残差容许值大于 e ,则返回 Step2,如果距离小于 e 则结束。

这种基于最小二乘法寻找最优目标函数解的方法只考虑了类中距,并没有考虑到类间距,容易陷入局部极值,不能得到全局最优解。并且,聚类数也很难确定。

2.2 以包含烈度信息椭圆为类间阈值空间散点聚类

根据中国地震地震活动受活动构造的控制这一特点,地震破坏分布往往与活动构造的走向一致,等震线常呈椭圆状,其长轴方向与活动构造的走向一致,一般情况下内圈等震线比较狭长,外圈等震线逐渐趋于圆形。为此,地震危险性分析的衰减模型采用椭圆衰减模型,所以,在此情况下除地震震级和距离外,等震线长轴取向对场点地震危险性起着一定的作用,在近场尤为显著。为此,本文提出了针对地震烈度椭圆为聚类阈值的空间散点方位信息聚类,此聚类椭圆包含了发震断层的方位信息和烈度信息,对于包含方位信息的散点的灾情点属性信息进行修改。

设烈度椭圆的长轴为 a ,短轴为 b ,离心角为 θ ,发震断层的走向作为椭圆长轴走向与水平线的夹角为 θ' (图 1)。其中烈度椭圆的长短轴由《中国地震烈度区划图说明》^[9] 中的烈度回归模型得出:

中国东部:

长轴 $I = 6.046 + 1.480M - 2.081\ln(R + 25), S = 0.49。$

短轴 $I = 2.617 + 1.435M - 1.441\ln(R + 7), S = 0.56。$

中国西部:

长轴 $I = 5.643 + 1.538M - 2.109\ln(R + 25), S = 0.64。$

短轴 $I = 2.941 + 1.363M - 1.494\ln(R + 7), S = 0.61。$

式中: I 为地震烈度; M 为面波震级; R 为长短轴半径; S 为标准误差。

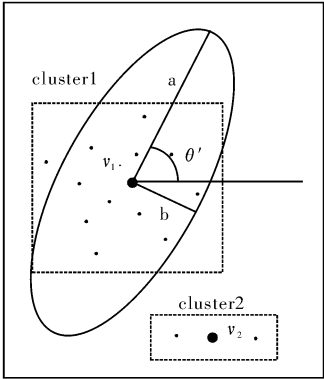


图1 椭圆阈值空间散点聚类示意图

徐卫晓等^[10] 提出了判定等震线走向的三种方法:① 利用断裂体系数字化;② 通过震源机制解判断等震线长轴走向;③ 通过余震分布确定等震线长轴走向。

对散点数据中的经纬度数据进行 c 均值聚类,求得第一类的类中心为 $v(v_x, v_y)$, 那么以第一类类中心的椭圆的参数方程为:

$$x = v_x + a \cos(\theta + \theta'), y = v_y + b \sin(\theta + \theta')。 \quad (11)$$

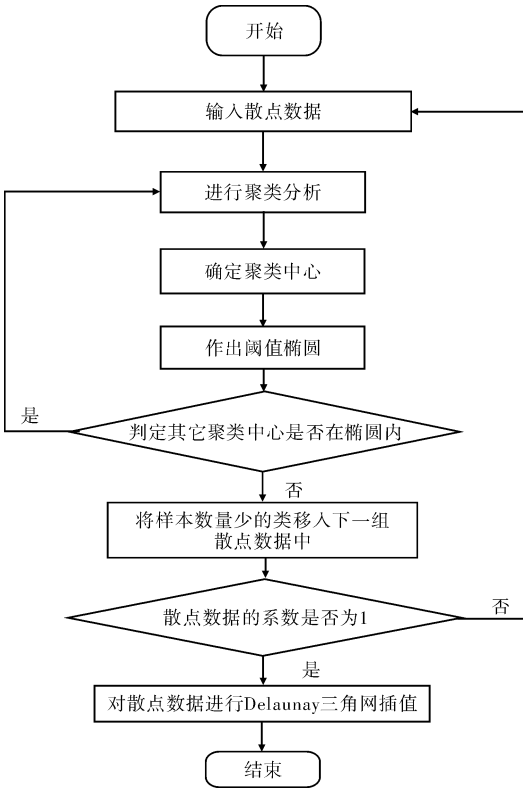
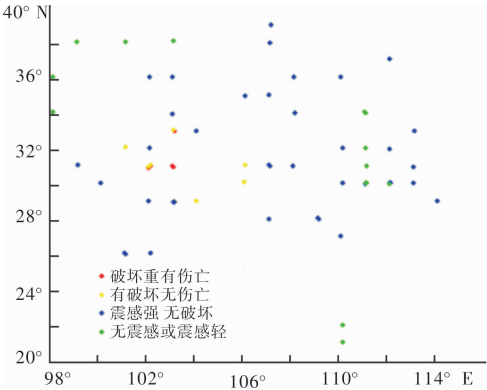
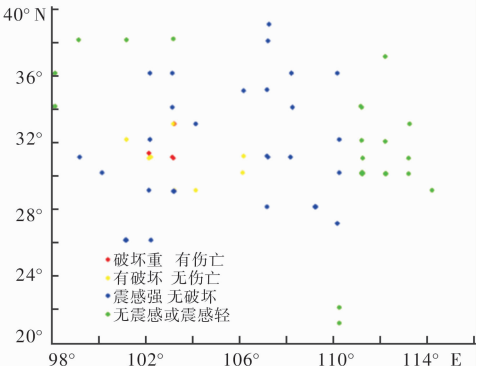


图2 椭圆阈值空间散点聚类程序流程图

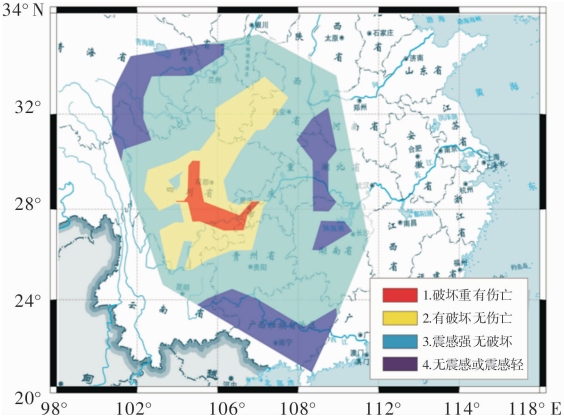


(a) 修正前

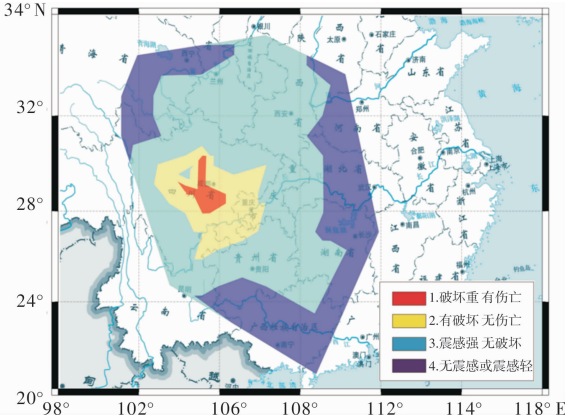


(b) 修正后

图3 空间散点聚类修正对比图



(a) 修正前



(b) 修正后

图4 空间散点聚类后插值对比图

式中: v_x, v_y 表示类中心的方位信息。根据参数方程得到椭圆的标准方程为:

$$[v_x + \alpha \cos(\theta + \theta')]^2/a^2 + [v_y + b \sin(\theta + \theta')]^2/b^2 = 1. \quad (12)$$

将其余的聚类中心 $v_j, i \in 2, 3, \dots, c$ 代入到标准方程中, 根据方程的结果判断其余聚类中心是否在椭圆外, 程序过程如图所示(图2)。如在椭圆内则继续聚类, 如在椭圆外则停止聚类, 将样本最多的一类定义为基本类, 其他类则合并到下一系数的聚类中继续进行聚类, 直到待聚类的数据的灾情属性信息为1时停止并对散点进行 Delaunay 三角插值得出实验结果。

3 实验结果与对比

本文的聚类模型以 MATLAB 为实验工具, 最终得出修正后的灾情损失评估结果, 通过与修正前的灾情损失评估做对比, 可以清楚的看出修正后的结果在河南、湖南、湖北等地的灾情属性点信息做了修改, 修改后的灾情信息更符合实际情况(图3)。经过与实际的烈度衰减图进行对比, 通过聚类修正后的灾情信息在没有足够多的数据拟合的情况下能够基本反映实际的地震损失状况(图4)。

4 结论与讨论

本文在模糊 c 均值聚类的基础上, 运用包含地震烈度信息的椭圆类间阈值对 12322 平台的 73 条有效灾情短信的方位值进行聚类, 该方法充分利

用了空间几何散点的方位信息, 基本消除了对灾情信息的人为影响。通过实验进一步修正了灾情损失评估结果, 使得结果更趋于合理, 能更好的显示各地区的受灾状况, 完善了地震速报平台的灾情评估系统, 为今后空间离散点的几何性质的聚类 and 地震应急救援提供了参考。

参考文献:

- [1] Rose K, Gurewitz E, Fox GC. Constrained clustering as optimization method[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1993, 15(8): 785-794.
- [2] 张慧哲, 王坚. 基于初始聚类中心选取的改进 FCM 聚类算法[J]. 计算机科学学, 2009, 36(6): 206-209.
- [3] 于剑, 程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学: E 辑, 2002, 4(2): 274-280.
- [4] 谢宇, 程维明. 一种基于类间阈值的模糊聚类算法[J]. 计算机应用与软件, 2012, 32(4): 34-39.
- [5] 孙才志, 王敬东, 潘俊. 模糊聚类分析最佳聚类数的确定方法研究[J]. 模糊系统与数学, 2001, 15(1): 89-92.
- [6] 帅向华, 侯建盛, 刘钦. 基于地震现场离散点灾情报告的灾害空间分析模拟研究[J]. 地震地质, 2000, 31(2): 321-333.
- [7] Bezdek J C. Pattern recognition with fuzzy objective function algorithms; Plenum[M]. New York; Suppinger, 1981: 256.
- [8] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm [J]. Computers & Geosciences, 1984, 10(23): 191-203.
- [9] 中国地震烈度区划图编委会. 中国地震烈度区划图说明[M]. 北京: 地震出版社, 1990.
- [10] 许卫晓, 孙景江, 林淋, 等. 烈度快速评估中的实时修正方法研究[J]. 地震工程与工程振动, 2012, 32(4): 34-39.

Elliptic Threshold Space Scatter Clustering Research Based on Intensity Attenuation

—Taking Quick Disaster Platform 12322 as Example

Cheng Xianzhou^{1,2}, Xiao Lanxi^{1,3}, Dong Xiang³ and Li Bo³

(1. Lanzhou Earthquake Research Institute, China Earthquake Administration, Lanzhou 730000, China;

2. Earthquake Administration of Shanghai, Shanghai 250000, China;

3. Earthquake Administration of Shandong Province, Jinan 250014, China)

Abstract: The inter elliptic threshold of ellipse intensity centered on epicenter is used, and by spatial dispersion of clustering analysis on latitude and longitude data of report information on 12322 damage report platform of Wenchuan M8.0 earthquake, the actual clustering numbers are determined according to the threshold, and remove unreasonable data through clustering to make the final scatter information to show the actual situation.

Key words: spatial clustering; elliptical threshold; intensity attenuation; disaster quick report