

林庆西,姜喜姣. 2023. 震源机制解谱聚类方法的研究及应用. 中国地震, 39(1):64~77.

震源机制解谱聚类方法的研究及应用

林庆西 姜喜姣

广东省地震局,广州 510070

摘要 多参数、多维度的震源机制解很难通过直观观察或简单的对比分析来进行有效的类型划分。谱聚类是一种基于谱图理论的聚类方法,对震源机制解这类非线性可分数据具有良好的划分效果。本文尝试使用该方法对震源机制解进行聚类分析,采用最小旋转角为相似度矩阵,利用规范割集准则(Ncut)完成类型判别,以间隔统计量法(Gap statistic)确定聚类数的最优解,从而对海量的震源机制解数据进行快速准确的类型划分。本文不仅通过一组随机样本数据集验证了这种方法的可行性和可靠性,还分别以海城 $M_s7.3$ 地震序列和川滇及周边地区的震源机制解集作为研究对象,验证了此方法的实用性。结果表明,该方法合理细分了区域内的震源机制解类型,不同类型解之间的差异性主要体现在受不同的区域构造背景控制,有利于区域地震活动性的研究。总体上看,基于震源机制解的谱聚类方法是区分震源机制解类型较为有效的方法,具有一定的实用价值。

关键词: 震源机制解 谱聚类 间隔统计量法 海城 $M_s7.3$ 地震序列 川滇及周边地区

[文章编号] 1001-4683(2023)01-0064-14 [中图分类号] P315 [文献标识码] A

0 引言

震源机制解是描述震源几何特性的重要参数,是研究断层错动方式和分析地壳应力场的基础资料(Gephart et al, 1984; 许忠淮, 1985)。为捕捉强震前的震兆信息,既可以利用不同阶段不同类型的震源机制解来反映强震前后构造应力场的时空变化(许忠淮等, 1989; 刁桂苓等, 2011),也可以利用震源机制的一致性规律判别强震前的小震群序列类型(陈颙, 1978; 程万正等, 2006)。但无论以上何种方法,均需要对震源机制解进行有效的类型划分。早期对震源机制解的对比分析多依赖专业知识和经验,如今,对于大量震源机制解的产出,直观观察震源机制解的异同或单纯利用人工处理分析显然不够精确。因此,利用机器学习对震源机制解进行准确快速的类型区分具有重要意义。

聚类分析是一种多元统计分析方法,其针对大量数据或者样品,根据数据本身特性将个体或者对象分类,使得同一类对象之间相似性比其他类对象的相似性更强(方开泰等, 1982; 何晓群, 2004)。震源机制解是一组具有多参数、多维度的数据,聚类分析方法可以有效地对

[收稿日期] 2021-11-16 [修定日期] 2022-02-21

[项目类别] 地震监测、预测、科研三结合课题(3JH-202301026)、震情跟踪定向工作任务(2023010108)、广东省地震局青年地震科学基金(重点实验室开放基金)项目(GDDZZ202306)共同资助

[作者简介] 林庆西,男,1986年生,博士研究生,主要从事震源机制解和应力场反演研究及应用工作。

E-mail: forkiter@163.com

大样本下的震源机制解集进行聚类划分,并提取出潜在的地震学特征。刁桂苓等(1992)提出了一种借助模式识别中系统聚类的方法,把两个解的P轴夹角与T轴夹角之和作为距离衡量,对震源机制解进行聚类划分。通过研究海城地震序列震源机制解,验证了该方法的可靠性和实用性,但由于其相似性距离的度量标准过于简单,迭代终止条件选取的较为主观,对聚类结果产生了一定影响。随后,许多学者针对震源机制的聚类做了大量研究工作,例如采用体波谱振幅相关系数聚类分析震源机制的变化(Lund, 2002; 朱航等, 2006; 崔子健等, 2012)、利用波形相关性和聚类分析确定小震群的震源机制解(Shelly et al, 2016)、通过震源机制节面聚类提取断层几何参数(万永革等, 2019)、震源机制一致性的显著性检验法研究(郭祥云等, 2019)等。上述工作多是针对大震前后或小区域事件集的震源机制相似性程度进行研究,无法体现大空间和大时间跨度下震源机制的差异性。另外,聚类的方法仍采用层次聚类等传统方法,类与类之间的划分取舍也存在一定的主观性。

谱聚类(Spectral Clustering,简称SC)算法是一种基于图论的聚类方法,是目前聚类分析中的一个研究热点,其本质是将聚类问题转化为图的最优划分问题。其优势在于能在任意形状的样本空间上聚类,且收敛于全局最优解,因此该方法适用于三维球空间的震源机制解(高琰等, 2007; 蔡晓妍等, 2008)。同时,由于谱聚类采用了主成分分析(PCA)中降维的思想,这对于处理多维的震源机制数据也十分有利。基于此,本文尝试采用谱聚类方法对震源机制解进行聚类分析,利用最小旋转角为相似度矩阵,以间隔统计量法的最优解确定聚类数,从而实现对海量震源机制解的自动快速类型划分。为检验这种方法的准确性与实用性,本文利用小区域、小数据量以及大区域、大数据量这两组震源机制解实例,分别进行了谱聚类分析,并针对聚类结果,给出该方法的优势以及适用范围。

1 方法

1.1 谱聚类

谱聚类算法来源于谱图划分理论(Fiedler, 1973),其主要思想是将n个样本数据(x_1, x_2, \dots, x_n)看作空间中的点,用边将这些点之间连接起来,组成一个基于样本相似度的无向加权图 $G=(V, E)$ 。其中, $V(v_1, v_2, \dots, v_n)$ 表示图 G 的顶点集合, E 表示具有权重值的边的集合,定义权重 w_{ij} 为点 v_i 和点 v_j 之间边的权重,由于是无向图,所以 $w_{ij}=w_{ji}$ 。假设距离较远的两点间的边权重较小,而距离较近的两点间的边权重较大,寻找一种切图的方式令切图后子图的边权重和尽可能低,而子图间的权重和尽可能高,从而达到聚类的效果。目前,谱聚类的算法已在很多场景中得到应用,其具体原理也在前人研究中经过反复讨论(Ng et al, 2002; 高琰等, 2007; 蔡晓妍等, 2008; 张宪超, 2017)。因此,本文首先将在简要介绍该方法原理基础上,研究针对复杂的震源机制解数据所采取的聚类策略和方法。

谱聚类的基础之一就是确定相似矩阵 W 的生成方式。相似矩阵主要通过样本点之间的距离度量来构建,在三维空间中一般采用欧式距离等函数。但对于震源机制解,普通的空间距离无法真实客观地描述不同断层的空间差异。Kagan(1991)采用了一种三维旋转方法,可以利用4种旋转模式来实现一个震源机制到另一个震源机制的旋转。这种旋转角度客观地反映了两个震源机制在三维空间的差别。因此,本文采用Kagan(1991)的算法,将两震源机制的最小旋转角作为距离度量。

利用距离构建相似矩阵一般有三种方法： ϵ -邻近法、 K 邻近法和全连接法(张宪超, 2017)。 ϵ -邻近法和 K 邻近法均给出了一种近似关系, 针对大数据量具有计算速度快的优势, 但不可避免会遗漏掉部分细节信息。为真实反映每个震源机制解的差异性, 本文采用全连接法中的高斯核函数(RBF)将距离值转换为相似矩阵, 即

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \quad (1)$$

其中, $x_i - x_j$ 为两个震源机制的最小旋转角; σ 为高斯核比例因子, 用来控制样本点的邻域宽度。将 w_{ij} 组成矩阵形式, 就可以得到相似矩阵 \mathbf{W} 。

根据边的权重值还可以计算出度矩阵 \mathbf{D} 。度 d 的定义是样本点和其相连的所有边的权重之和, 公式表示为

$$d_i = \sum_{j=1}^n w_{ij} \quad (2)$$

利用度可以组成一个 $n \times n$ 的对角线矩阵, 即度矩阵 \mathbf{D} 为

$$\mathbf{D} = \begin{pmatrix} d_1 & \cdots & \cdots & \cdots \\ \vdots & d_2 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & d_n \end{pmatrix} \quad (3)$$

其中, 度矩阵的对角线表示每个样本点的度, 其他矩阵位置均为 0。

谱图理论实现数据聚类问题的关键在于分割子图, 即通过划分准则寻找最优解, 这是一个 NP 难问题(Non-deterministic Polynominal-Hard), 通常解决的方法是采用谱方法将该问题松弛化为连续实数值, 使其转换为求解拉普拉斯矩阵的谱分解, 然后再将松弛化的问题离散化, 得到最终的聚类结果, 其本质是对图划分准则的逼近(Zhou et al, 2003)。目前, 划分准则包括最小割集准则(Minimum cut)、规范割集准则(Normalized cut)和比例割集准则(Ratio cut)等(蔡晓妍等, 2008; Wu et al, 1993; Shi et al, 2000; Hagen et al, 2002)。本文采用常见的规范割集准则(简称 NCut)计算归一化的随机游走拉普拉斯矩阵, NCut 切图的函数表达式为

$$\text{NCut}(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \overline{A}_i)}{\text{vol}(A_i)} \quad (4)$$

这里假设聚类数为 k , A_i 表示第 i 类样本数据集, \overline{A}_i 为除 A_i 子集外其他子集的并集, $W(A_i, \overline{A}_i)$ 表示 A_i 和 \overline{A}_i 之间切图的权重之和, $\text{vol}(A_i)$ 表示第 i 类所有样本的度之和。 $W(A_i, \overline{A}_i)$ 和 $\text{vol}(A_i)$ 表示为

$$W(A, B) = \sum_{i \in A, j \in B} w_{ij} \quad (5)$$

$$\text{vol}(A) = \sum_{i \in A} d_i \quad (6)$$

谱聚类最终的优化目标是最小化 NCut 函数, 这里需要用到维度规约的思想去近似求解 NP 难问题, 即计算标准拉普拉斯矩阵最小的前 k 个特征, 然后求出对应的特征向量, 并标准化, 再对最后的特征向量矩阵进行传统的 k -means 聚类, 即可得到最终的聚类结果。标准化拉普拉斯矩阵为

$$\mathbf{L}_{sym} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}} \quad (7)$$

其中, \mathbf{D} 表示度矩阵, \mathbf{L} 为拉普拉斯矩阵, 数学表达式为 $\mathbf{L} = \mathbf{D} - \mathbf{W}$, \mathbf{L}_{sym} 为标准化后的拉普拉斯矩阵。

1.2 间隔统计量法确定聚类数

在聚类分析中, 聚类数目的多少直接影响聚类的质量。从谱聚类的原理中可以发现, 聚类数 k 需要用户给出, 这对于简单数据可以很直观或有目的地得到。但是对于震源机制这种多维多属性的复杂数据集, 显然无法直观获取聚类的最优数目。因此如何自动地确定最佳聚类数目至关重要。通常解决这一关键问题的方法是利用评价指标穷举聚类数, 然后对每一个聚类结果进行评估, 根据评估结果确定最优的聚类数目。这一类方法虽然计算量大且耗时严重, 但给出的结果是最为可靠的。目前评价方法主要有平方误差和法 (SSE)、Calinski-Harabasz 准则法、轮廓系数法和间隔统计量法 (Gap statistic) 等 (Calinski et al, 1974; Rousseeuw, 1987)。本文主要采用间隔统计量法 (以下简称 GS) 评估震源机制解聚类数目。

GS 聚类评价方法是由 Tibshirani 在 2000 年提出的, 其思想主要来源于方差分析 (Tibshirani et al, 2001), 通过待分数据集的离散度与由参考分布生成的数据集的离散度的差异构造一个间隔统计量, 用以刻画间隔统计量关于聚类数目的变化情况, 当间隔统计量取得最大值时, 对应的 k 为最佳聚类数 (窦婷, 2017)。

在应用于震源机制解的聚类数时, 假设震源机制解数据集已有谱聚类算法聚成 k 类: C_1, C_2, \dots, C_k , 震源机制解 x_i 和 x_j 的距离平方即最小旋转角的平方表示为 d_{ij} , 则 C_r 类 ($r=1, 2, \dots, k$) 的类内紧凑度为

$$D_r = \sum_{x_i, x_j \in C_r} d(x_i, x_j) = \sum_{x_i, x_j \in C_r} \|x_i - x_j\|^2 \quad (8)$$

所有类内紧凑度总和, 即平方误差和 W_k 可以表示为

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r \quad (9)$$

其中, n_r 表示第 r 类中震源机制解的数量。从式(9)可以看到, 随着聚类数的增大, 每一类震源机制解越来越相似, 距离平方和越来越小, 因此 W_k 值必定随着震源机制的聚类数增大而减小。但是, 当 W_k 减小的较为缓慢时, 可以认为增加聚类数并不能进一步改善聚类效果, 这个减小缓慢点即为“肘部”拐点, 此时的聚类数 k 可以视为最佳的震源机制解类别数。如何定量判断“肘部”拐点是 GS 方法中的一项关键工作。

根据震源机制解的维度, 利用蒙特卡洛抽样方法建立参考数据集, 并计算每个参考数据集平方误差和, 建立 Gap 值的关系式, 即

$$Gap_n(k) = E_n^* \{ \lg(W_k) \} - \lg(W_k) \quad (10)$$

$$E^* \{ \lg(W_k) \} = \frac{1}{B} \sum_{b=1}^B \lg W_{kb}^* \quad (11)$$

其中, $E^* \{ \lg(W_k) \}$ 表示 $\lg(W_k)$ 的期望, B 为蒙特卡洛抽样次数, W_{kb}^* 表示第 b 个参考数据集内平方误差和。由于参考数据集随机数的非线性特征, 需要利用标准误差 $sd(k)$ 进行修正, 即

$$sd(k) = \left[\frac{1}{B} \sum_{b=1}^B ((\lg(W_{kb}^*)) - \bar{l})^2 \right]^{\frac{1}{2}} \quad (12)$$

$$\bar{l} = \frac{1}{B} \sum_{b=1}^B \lg W_{kb}^* \quad (13)$$

定义 $s_k = \sqrt{1 + \frac{1}{B} sd(k)}$, 使式(14)成立的最小的 k_{\min} 就是通过 GS 法得到的最佳聚类数。

即当最佳聚类数为 k_{\min} 时, Gap 值应为 $\text{Gap}(k)$ 中的最大值

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1} \quad (14)$$

1.3 震源机制平均解的求取

在完成震源机制解的聚类后,需要对每一类震源机制的空间状态进行描述,这就需要寻找一个平均解来体现断层面的空间几何特性。本文主要采用刁桂苓等(1992)求平均解的方法,对同类震源机制的 P 轴或 T 轴进行矢量合成,计算出同类解的力轴平均位置,从而得到震源机制的平均解。

1.4 数值实验

为研究将谱聚类方法应用于震源机制解的聚类中是否可行以及结果是否可靠,本文根据生成的一个合成震源机制目录进行方法验证。通过改变 3 个断层面参数:走向 = {90°, 135°, 180°}、倾角 = {30°, 60°} 和滑动角 = {-90°, 0°, 90°} 来生成一个三维的格状网 ($3 \times 2 \times 3$) (Martinez-Garzon et al, 2014)。这些参数的不同组合生成了多种断层模式,从纯走滑到斜滑再到纯正断或逆断,每个网格点随机生成符合正态分布的 100 个震源机制。为直观区分每一类的震源机制,将标准差 σ 设置为 5°。利用上文给出的震源机制解谱聚类方法和 GS 方法,得到结果如图 1 所示。

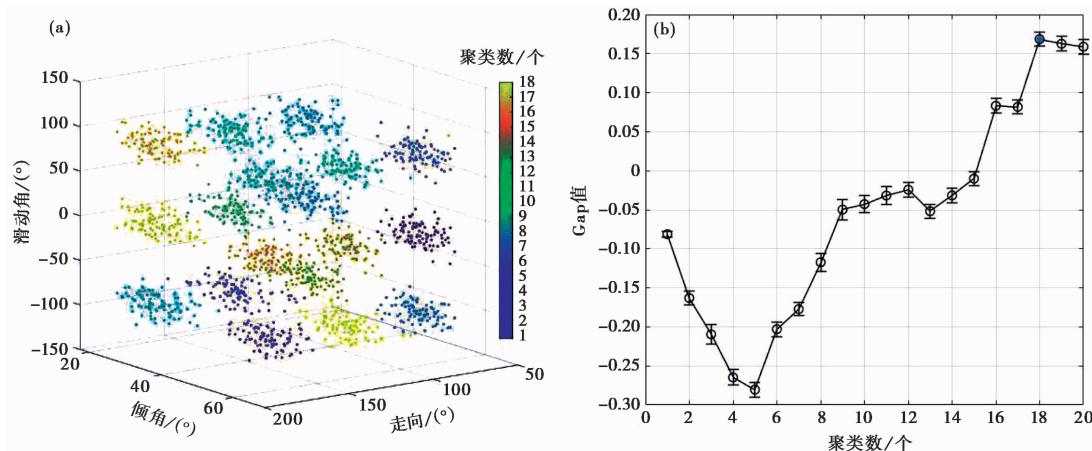


图 1 合成震源机制数据集的谱聚类结果(a)与 Gap 统计(b)

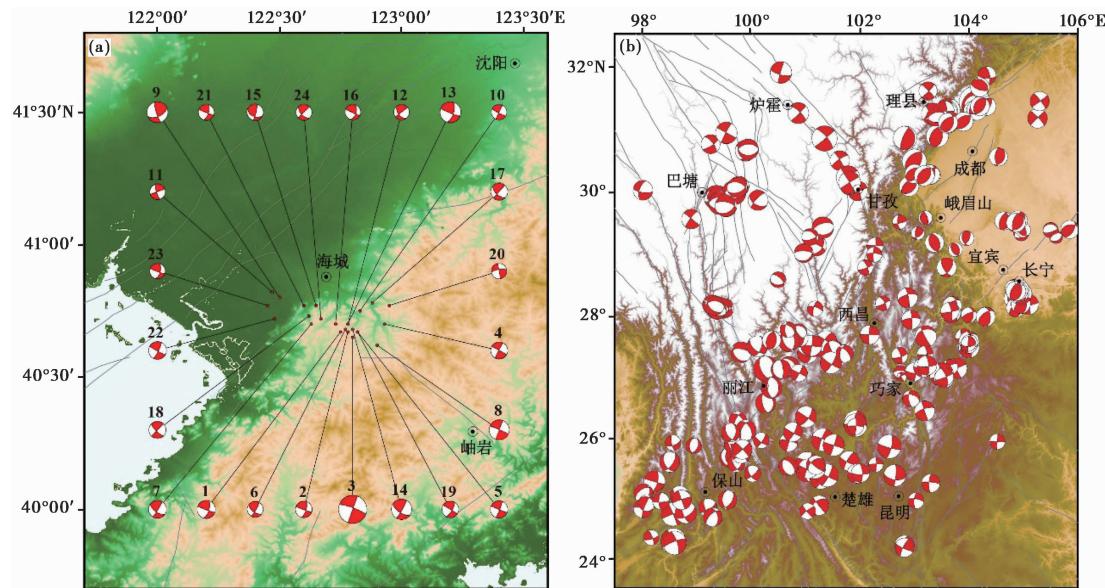
图 1(b)为利用 GS 方法得到的 Gap 值随聚类数变化曲线,图中可以明显看到,在聚类数 k 为 18 时,Gap 值最大,因此可以认为 18 是此震源机制数据的最佳聚类数,这也符合之前的预设(网格点 $3 \times 2 \times 3 = 18$)。图 1(a)是当 $k=18$ 时,合成震源机制的聚类情况,图中不同颜色表示不同的类别,可以看到每一簇聚类都基本被区分出,聚类效果良好。通过数值实验可以看到,本文研究的震源机制解谱聚类方法是可行的,结果稳定且可靠。

2 应用实例与计算

2.1 数据选取

为检验本文震源机制解聚类方法的实用性,本文选取了两个区域的地震震源机制解集作为主要研究对象,用以分析震源机制解的聚类效果,分别为1975年海城 $M_s7.3$ 地震序列和川滇及周边区域地震事件集。这主要基于以下3个原因:①两组不同区域的地震均有明确的震源机制解集,其计算结果已被不同学者或机构验证,可以确保震源机制解的准确性和可靠性;②1975年海城 $M_s7.3$ 地震序列已有学者做过类似的聚类分析(刁桂苓等,1992),通过聚类结果可以对比分析震源机制解聚类方法相较于其他方法的优劣;③川滇块体和周边区域研究范围较广,且数据量较大,区内断裂特征明显,有利于通过聚类结果分析区域地震活动与构造特征的关系,从而探讨揭示区域内不同的孕震环境。

1975年海城 $M_s7.3$ 地震序列采用顾浩鼎等(1976)的结果,共有24个震源机制解,其具体分布如图2(a)所示。川滇及周边区域震源机制数据主要采用罗钧(2013)的结果,其利用CAP方法反演得到了该区域2007—2013年间共75次3.5级以上地震震源机制解。为了扩大聚类时的数据量,另外补充了1976年至2020年12月31日GCMT记录的该区域141个震源机制解,总计216个震源机制结果,具体分布如图2(b)所示。



注:文中所有震源球均采用下半球投影。

图2 海城 $M_s7.3$ 地震序列(a)和川滇及周边区域地震(b)震源机制解空间分布

2.2 谱聚类计算

震源机制解是一个包含节面和 P 轴、 T 轴和 B 轴的12维的数据:2个节面的走向(strike)、倾角(dip)和滑动角(rake),以及 P 轴、 T 轴和 B 轴的方位角(trend)和倾伏角(plunge)。尽管12个参数中只有3个是独立的,即可以由节面的3个参数或 P 轴2个参数加上 T 轴的1个参数求出其余的参数(中国地震局监测预报司,2017),但是为了讨论震源机

制的异同,需要考虑数据的整体完备性。

利用谱聚类方法对震源机制解进行聚类分析,其具体步骤如下:

(1) 对震源机制数据按地震事件的时间顺序进行排列,并标注序号,用以建立震源机制数据的索引。

(2) 构建相似矩阵和度矩阵。利用 Kagan 算法求出两两地震对的最小旋转角,组成一个 $n \times n$ 的对称矩阵 \mathbf{W} (n 为区域内震源机制解的数目)。这里的对称矩阵是由无向图的性质所决定。根据全连接方法中高斯核函数的定义,对距离矩阵进行转换,得到同样对称的相似矩阵。这里规定核参数 $\sigma=40$,可以保证震源机制数据之间的权重值处于 0~1 区间内。同样,度矩阵也可以通过高斯核函数求得的权重值计算得到。

(3) 利用相似矩阵和度矩阵,计算得到拉普拉斯矩阵,然后对其进行标准化,得到标准化后的拉普拉斯矩阵 \mathbf{L}_{sym} ($n \times n$)。

(4) 计算标准化拉普拉斯矩阵最小的 k 个特征值以及对应的特征向量 f 。将各自对应的特征向量 f 组成矩阵形式,然后按行进行标准化,最终组成 $n \times k$ 维的特征矩阵 \mathbf{F} 。这时,震源机制解转换为了一组具有 k 维的数据矩阵。

(5) 利用传统的 k -means 聚类方法,对 k 维的 n 个新数据进行聚类,得到 k 个类别,利用序号索引还原到这 k 个类别中,可以得到最终的聚类结果。

在聚类过程中,采用 GS 方法分别对两个区域震源机制解的聚类结果进行评价,获取最佳聚类个数,如图 3 所示。可以看到当聚类数 k 均为 7 时聚类效果最佳。比较两个区域数据聚类时的 Gap 值和误差条显示,海城地震数据 Gap 值较小且误差范围较大,这主要是因为地震样本量过少导致的;川滇及周边地区地震样本量较多,且地震分布范围较广,类间相似程度差异大,聚类结果较为准确且稳定。从以上两个区域地震数据 Gap 值分布状态分析可知,地震样本数量直接影响 Gap 值的误差分布,这主要是由震源机制的特性所决定的。将最佳聚类数代入震源机制谱聚类的计算步骤中,可以得到最终的聚类结果。下面将着重对聚类结果进行研究和分析。

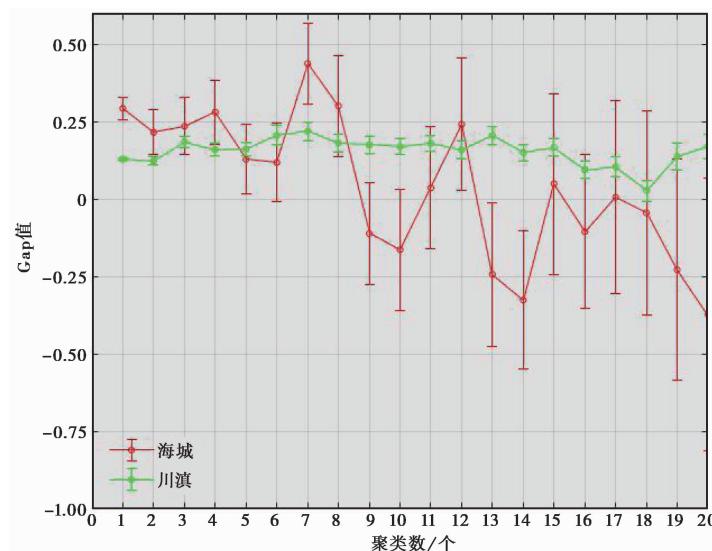


图 3 Gap 值统计分析

3 结果分析

3.1 1975 年海城 $M_s 7.3$ 地震序列

表 1 给出了 1975 年海城 $M_s 7.3$ 地震序列的地震目录和震源机制解(顾浩鼎等,1976)。根据上文提到的方法和步骤,对海城区域 24 个震源机制解聚类分析,计算得到 7 类震源机制解,聚类结果见表 1。刁桂苓等(1992)曾利用系统聚类中的最长距离法将该区域 24 个震源机制划分为了 3 种类型,其中,系统聚类得到的第 I 类解包含于本文的第 I 类解,第 II 类解对应了本文第 I、II 和 III 类解,第 III 类解对应于本文第 IV、V 和 VII 类解。观察两个方法得到的聚类结果,主要有以下 3 点不同:

(1) 序号 20 的震源机制解其聚类结果间存在明显差异。系统聚类法将其归于前震相似的震源机制类别(第 I 类),而谱聚类方法将其划分为一个新的类别(第 VI 类)。通过分析相似度矩阵可知,谱聚类结果中第 I 类解类内平均最小旋转角为 6.1° ,而序号 20 的震源机制

表 1 海城 $M_s 7.3$ 地震序列的震源机制解(顾浩鼎等,1976)以及
系统聚类(刁桂苓等,1992)和谱聚类的聚类结果

序号	日期 (年-月-日)	发震时刻 (时:分:秒)	震级 (M_L)	地理坐标		节面 I			系统聚类结果	谱聚类结果
				经度 /(°E)	纬度 /(°N)	走向 /(°)	倾角 /(°)	滑动角 /(°)		
1	1975-02-04	07:50:47	4.7	122.75	40.67	110	85	25.1	I	I
2	1975-02-04	10:35:35	4.3	122.78	40.67	109	86	22.1	I	I
3	1975-02-04	19:36:06	7.3	122.80	40.65	290	81	-15.2	II	II
4	1975-02-05	01:01:45	4.4	122.93	40.70	298	88	4.0	II	II
5	1975-02-05	02:56:29	4.5	122.82	40.67	112	88	6.0	II	II
6	1975-02-05	12:33:00	4.1	122.77	40.68	126	61	11.5	II	III
7	1975-02-05	23:52:54	4.6	122.63	40.70	125	64	5.6	II	III
8	1975-02-06	05:43:42	5.2	122.90	40.62	292	87	-2.0	II	II
9	1975-02-06	12:24:57	5.4	122.50	40.80	162	80	34.6	III	VII
10	1975-02-06	13:56:16	4.0	122.83	40.75	116	80	-128.9	III	V
11	1975-02-08	02:30:23	4.0	122.47	40.82	161	78	-31.8	III	IV
12	1975-02-12	20:42:46	4.0	122.78	40.70	143	52	-118.2	III	V
13	1975-02-15	21:08:02	5.4	122.78	40.70	111	84	26.2	I	I
14	1975-02-16	22:01:26	5.3	122.80	40.68	123	62	6.8	II	III
15	1975-02-18	18:51:49	4.2	122.65	40.77	114	44	14.5	II	III
16	1975-02-22	15:45:14	4.0	122.73	40.70	113	85	36.2	I	I
17	1975-02-24	05:07:20	4.4	122.88	40.78	132	57	-52.9	III	IV
18	1975-02-25	04:52:10	4.4	122.62	40.73	134	74	-4.2	II	II
19	1975-02-26	05:09:53	4.3	122.82	40.67	121	90	20.0	II	I
20	1975-03-21	11:32:59	4.0	122.95	40.77	260	72	-29.6	I	VI
21	1975-03-29	23:16:36	4.1	122.60	40.77	111	85	24.1	I	I
22	1975-04-10	03:55:37	4.6	122.48	40.72	118	85	9.0	II	II
23	1975-04-21	00:17:06	4.0	122.45	40.77	109	84	30.2	I	I
24	1975-07-04	07:06:29	4.1	122.67	40.72	137	56	-19.4	II	III

解和第Ⅰ类平均解最小旋转角为 39° ,两者相差甚远。因此将序号20的震源机制解划分为第Ⅰ类解是不合适的。同样情况存在于序号为19的地震,经分析相似度矩阵,其与第Ⅰ类解距离更近。出现这种偏差可能是由于系统聚类采用的样本距离仅考虑了P、T轴,虽然简化了计算,但是对于空间的震源机制解未做到更好的约束。

(2)谱聚类方法将系统聚类中的第Ⅱ类解进一步细分为两类,如表1和图4所示。为了突出每一类震源机制的特征,利用P轴和T轴的矢量和计算得到了每一类震源机制的平均解,图4中用黑色圆弧和大实心圆来表示。从图4中可以清晰看到,两类解还是有较大的差异,谱聚类方法得到的第Ⅱ类解地震为纯走滑型,P轴和T轴倾伏角较小,接近于水平;而第Ⅲ类解地震具有一定的倾滑分量,P轴和T轴倾伏角则明显变陡。谱聚类得到的结果较于系统聚类要更为精确一些。

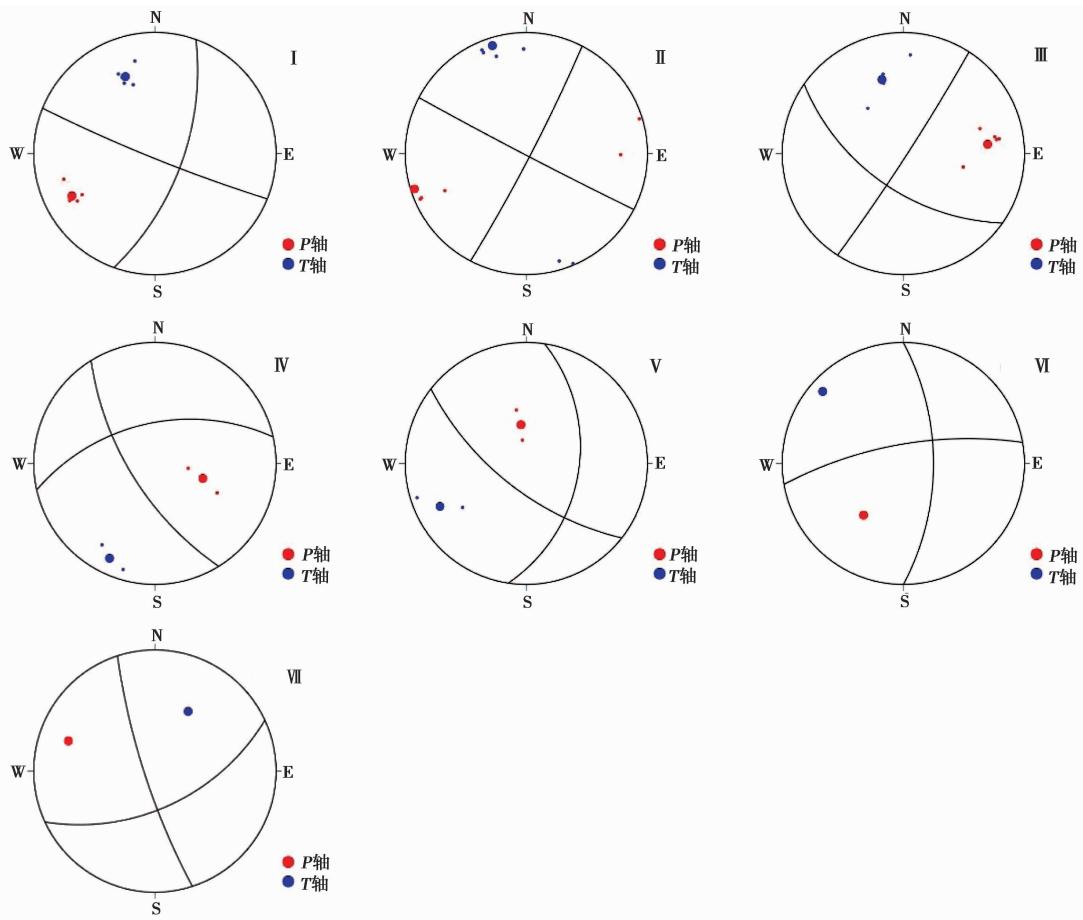


图4 海城 $M_s 7.3$ 地震序列震源机制解聚类结果

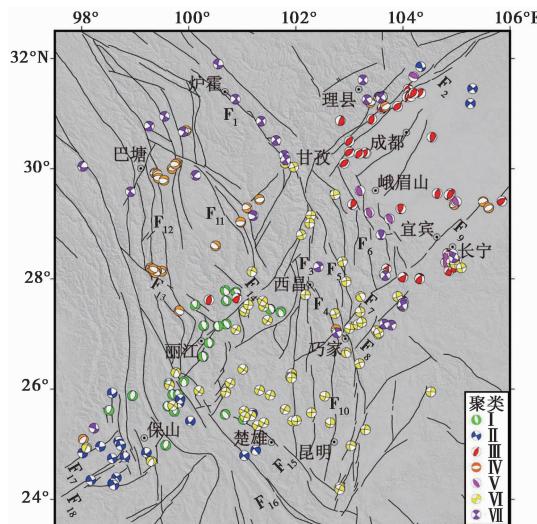
(3)顾浩鼎等(1976)和刁桂苓等(1992)提到的5个“错动过头”的地震在谱聚类中也区分了出来,而且进一步细分了这5个地震的类型(第Ⅳ、Ⅴ和Ⅶ类),如图4所示。三类解体现出明显的差异性,其中第Ⅳ和第Ⅴ类解具有明显的正断层倾滑性质,可能为“错动过头”所致,也可能受不同的区域构造背景所控制,如褒窝水库的垂向水重力影响。如果简单地归于

一类解,体现不出部分震源机制的差异性,不利于区域内复杂地震活动的力学成因分析和研究。

通过上述分析可以发现,谱聚类方法在处理复杂震源机制解聚类过程中所得结果较好,不仅给出了类间震源机制的差异,而且利用GS法给出了聚类的最优解。但从图3的Gap值波动范围可知,由于海城研究区仅有24个地震事件,样本量过少,因此结果有一定的误差,但无论如何,震源机制解的谱聚类方法得到的聚类结果仍是较为准确且可靠的。

3.2 川滇及周边区域

川滇地区位于青藏高原的东南缘,是近年来中国大陆地震活动最强烈的地区之一(闻学泽等,2003)。该区域范围较广,其构造活动以大规模剪切变形为主,兼有强烈的隆升运动,发育多条活动大断裂,构造背景非常复杂(罗钧等,2014)。对上文提到的216个震源机制解进行聚类分析,得到聚类结果空间分布(图5),图中震源机制解的聚类结果用不同颜色进行区分,每种类型的平均解如图例所示。



注: F₁:鲜水河断裂带; F₂:龙门山断裂带; F₃:安宁河断裂带; F₄:则木河断裂带; F₅:大凉山断裂带; F₆:荥经—马边—盐津断裂带; F₇:莲峰断裂带; F₈:昭通断裂带; F₉:华蓥山断裂带; F₁₀:小江断裂带; F₁₁:理塘断裂带; F₁₂:金沙江断裂带; F₁₃:德钦—中甸一大具断裂带; F₁₄:丽江一小金河断裂带; F₁₅:曲江断裂带; F₁₆:红河断裂带; F₁₇:大盈江断裂带; F₁₈:瑞丽断裂带。

图5 川滇地区主要断裂带及震源机制解聚类结果空间分布

根据研究区内主要断裂带,简要分析和讨论聚类结果的空间分布特征:

金沙江断裂带主要分布两种类型的震源机制解:第IV类正断型和第VII类走滑型。其北段与理塘断裂带交汇区域以及南段与德钦—中甸一大具断裂带交汇区域均发育为正断型的第IV类解,而该断裂带东侧的理塘断裂带中段地震也为第IV类,说明两者的震源特征相同,这与罗钧等(2014)给出的结论一致。第VII类解多分布于金沙江断裂带以北区域,受构造活动影响走滑特征显著。鲜水河断裂带北段和中段同样分布第VII类解,但从甘孜往南,震源机制解变为第VI类。尽管两类解均为走滑型,但是分属两个不同的类别,说明鲜水河断裂带南

北段受到不同的构造背景影响,差异特征明显,这在前人研究中得到了验证(罗钧等,2014;吴萍萍等,2016)。

龙门山断裂带主要分布第Ⅲ类逆冲型解,这主要受青藏高原隆起、深部物质向东流动的影响,形成“挤压”型地震。但由于该区域应力环境十分复杂,由图5可以清晰地看到自SW至NE方向震源机制解类型发生多次变化,分段特征显著(杨宜海等,2021)。荥经—马边—盐津断裂带以逆冲型的第Ⅴ类解为主,和龙门山断裂带西南段地震形成原因相似。但由于两条断裂带断层走向不同,因此两个区域内震源机制解应分属于不同的类型,聚类结果准确且合理。

安宁河—则木河—小江断裂带及周边主要分布走滑型的第Ⅵ类解,其NS向节面的左旋走滑特征与该区域地质活动一致(段梦乔等,2019;苏珊,2019)。在巧家附近可以明显看到两个不同于附近震源机制解类型的地震,主要表现为正断型和正断兼走滑分量的特征,说明该区域两次地震均受到一定拉张作用的影响(罗钧等,2014)。谱聚类方法准确地将这两个地震区分出来。昭通—莲花山断裂带西南区域分布着与小江断裂带一致的第Ⅵ类解,但在东北区域受坚固的华南块体阻挡,地震断层类型变为NE向节面右旋走滑且部分带有逆冲分量,在更靠近四川盆地交接地时,震源机制解变为第Ⅲ类逆冲型,体现出了明显的分段特征(李君等,2019)。

丽江—小金河断裂带中北段南侧靠近安宁河断裂带主要分布走滑型的第Ⅵ类解,而北侧靠近德钦—中甸一大具断裂带主要分布正断型的第Ⅰ类解。两类解以丽江—小金河断裂带为界分隔开,呈现明显的空间分布特征,说明丽江—小金河断裂带中北段两侧受不同构造动力作用的影响。丽江—小金河断裂带南段分布有三种类型震源机制解,其中第Ⅰ类为该断裂带主要分布的震源机制解,第Ⅱ类为西南端大盈江和瑞丽断裂带主要分布的震源机制解,第Ⅵ类为东侧滇中次级块体主要分布的震源机制解,丽江—小金河断裂带南段夹于三个区域交汇处,体现出该段断层活动和发震断层破裂的复杂性(罗钧等,2014;Tian et al,2019)。

总体上看,川滇及周边区域震源机制解的聚类类型分区性明显,其空间分布符合各个断裂带的构造活动和块体间相互作用的控制,聚类结果准确可靠。

4 结论与讨论

本文利用谱聚类方法对震源机制解数据进行了聚类分析,该方法针对震源机制解的数据结构特点,采用震源机制之间的空间最小旋转角作为相似度矩阵,利用GS方法得到的最优解约束聚类的数量,基本解决了震源机制复杂数据解集的聚类问题。在谱聚类方法研究中,主要得到以下认识:

(1)震源机制解的数据结构复杂多维,仅从单个或几个元素或者直观观察很难分辨出震源机制解之间的异同。谱聚类是近年来较为流行的一种聚类方法,针对多维数据有着良好的适应性,这对于解决多维多属性的震源机制解无疑是有利的。通过模拟数据验证,可以观察到不同类别之间明显的差异性,说明谱聚类方法针对震源机制解的聚类结果是可信的。

(2)相似矩阵的构建是谱聚类算法的前提,震源机制解表达的是一种空间球体的概念,他们之间的最小空间旋转角可以真实反映两者之间的亲疏远近,因此采用最小空间旋转角

表达震源机制解之间的差别最为科学(万永革,2019)。

(3)本研究中的构图方式采用的是全连接法,其可以真实反映每个震源机制解的差异性,适用于区域性震源机制解的聚类,即样本量不高的情形。但如果样本过大,此方法迭代速度会受到严重制约,此时可考虑 ϵ -邻近法, K 邻近法等,尽管可能会造成部分信息缺失,但计算速度会大幅提升。

(4)切图准则是谱聚类的核心思想,在现阶段应用方法很多,每个方法都有其解决NP难问题的优势范围。本文采用的NCut方法应用较为广泛,切图结果可靠性也较高,其归一化的特征可以平衡类间的相异度,避免对小区域的分割。针对震源机制解,NCut方法有其合理的解决范围,但其他切图方法或许更为适用,这需要在以后的实践中具体验证。

(5)解决聚类数目问题的GS方法,其本质是统计样本类内和类间的离差平方和,通过统计量刻画样本观察值和参考分布下期望值之间的差异。这种复杂的大量统计方法可靠性高,应用在小样本中可以快速实现,但对于大样本的统计其时间成本过高,还需要高性能硬件的支撑。现阶段确定聚类数目方法很多,效果也因人而异,如何挑选适用于自身数据集的方法仍是当前聚类应用的一个难点工作,这需要更为深入的研究,从而寻找解决问题的最优解。

本研究不仅选择利用小区域小数据量的震源机制解数据进行谱聚类分析,还对大区域大数据量的震源机制解集进行了聚类分析。研究结果表明,相较于系统聚类方法,震源机制解的谱聚类方法在实际地震中更为适用,同时,GS方法也对聚类数做出了合理的判断,聚类结果更为准确且符合实际;在大数据量处理上,震源机制解的谱聚类方法不需要考虑震源机制类型,如Zoback(1992)的震源机制解分类方法,也可以合理准确地划分出震源机制解的不同类型,应用在川滇及周边地区中可以明显看到其聚类效果。由此可见,震源机制解的谱聚类方法不仅适用于小区域小数据量的地震事件,还适用于大区域大数据量的地震事件,且聚类结果基本准确且可靠。但在实际应用过程中,震源机制解的谱聚类方法仍存在一定的局限性,如川滇及周边地区的安宁河断裂带和曲江断裂带,断层实际走向差别巨大,但由于地震的震源机制很相似,谱聚类方法将其聚为同一类,说明分类结果的解释仍然需要结合断层实际走向,否则可能出错。解决该问题是改进谱聚类方法及今后发展的一个方向。

参考文献

- 蔡晓妍,戴冠中,杨黎斌. 2008. 谱聚类算法综述. 计算机科学,35(7):14~18.
- 陈颙. 1978. 用震源机制一致性作为描述地震活动性的新参数. 地球物理学报,21(2):142~159.
- 程万正,阮祥,张永久. 2006. 川滇次级地块震源机制解类型与一致性参数. 地震学报,28(6):561~573.
- 崔子健,李志雄,陈章立,等. 2012. 判别小震群序列类型的新方法研究——谱振幅相关分析法. 地球物理学报,55(5):1718~1724.
- 刁桂苓,徐锡伟,陈于高,等. 2011. 汶川 M_w 7.9 和鲁甸 M_w 7.6 地震前应力场转换现象及其可能的前兆意义. 地球物理学报,54(1):128~136.
- 刁桂苓,于利民,李钦祖. 1992. 震源机制解的系统聚类分析——以海城地震序列为例. 中国地震,8(3):86~92.
- 窦婷. 2017. ISODATA 模型及其 Gap 统计应用研究. 硕士学位论文. 南京:南京理工大学,9~10.
- 段梦乔,赵翠萍. 2019. 金沙江下游水库区地震震源机制特征. 地震地质,41(5):1155~1171.
- 方开泰,潘恩沛. 1982. 聚类分析. 北京:地质出版社,3~21.
- 高琰,谷士文,唐琎,等. 2007. 机器学习中谱聚类方法的研究. 计算机科学,34(2):201~203.

- 顾浩鼎,陈运泰,高祥林,等. 1976. 1975年2月4日辽宁省海城地震的震源机制. 地球物理学报, **19**(4):270~285.
- 郭祥云,陈学忠,李艳娥,等. 2019. 震源机制一致性的显著性检验方法——以1999年11月29日辽宁岫岩 M_s 5.4 地震前震序列为例. 地震学报, **41**(6):709~722.
- 何晓群. 2004. 多元统计分析. 北京:中国人民大学出版社,58~60.
- 李君,王勤彩,崔子健,等. 2019. 川滇菱形块体东边界及邻区震源机制解与构造应力场空间分布特征. 地震地质, **41**(6): 1395~1412.
- 罗钧. 2013. 川滇块体及周边现今震源机制和应力场特征研究. 硕士学位论文. 北京:中国地震局地震预测研究所,46~49.
- 罗钧,赵翠萍,周连庆. 2014. 川滇块体及周边区域现今震源机制和应力场特征. 地震地质, **36**(2):405~421.
- 苏珊. 2019. 安宁河断裂及周边地区震源机制解与应力场分布特征. 硕士学位论文. 北京:中国地震局地球物理研究所, 35~45.
- 万永革. 2019. 同一地震多个震源机制中心解的确定. 地球物理学报, **62**(12):4718~4728.
- 万永革,靳志同. 2019. 根据大量震源机制节面聚类提取活动断层几何参数. 见:第二届地球物理信息前沿技术研讨会论文摘要集. 哈尔滨:中国地球物理学会信息技术专业委员会,9~10.
- 闻学泽,易桂喜. 2003. 川滇地区地震活动统计单元的新划分. 地震研究, **26**(增刊I):1~9.
- 吴萍萍,王阳,朱洁,等. 2016. 1970年以来鲜水河断裂带地震活动特征与2014年康定 M_s 6.3 地震. 中国地震, **32**(4): 776~786.
- 许忠淮. 1985. 用滑动方向拟合法反演唐山余震区的平均应力场. 地震学报, **7**(4):349~362.
- 许忠淮,汪素云,黄雨蕊,等. 1989. 由大量的地震资料推断的我国大陆构造应力场. 地球物理学报, **32**(6):636~647.
- 杨宜海,张雪梅,花茜,等. 2021. 龙门山断裂带的分段性特征——来自密集震源机制解的约束. 地球物理学报, **64**(4): 1181~1205.
- 张宪超. 2017. 数据聚类. 北京:科学出版社,157~189.
- 中国地震局监测预报司. 2017. 测震学原理与方法. 北京:地震出版社,433.
- 朱航,刘杰,陈天长. 2006. 采用体波谱振幅相关系数方法研究地震序列的震源机制变化过程. 地震, **26**(2):1~11.
- Calinski T, Harabasz J. 1974. A dendrite method for cluster analysis. Commun Stat, **3**(1):1~27.
- Fiedler M. 1973. Algebraic connectivity of graphs. Czech Math J, **23**(2):298~305.
- Gephart J W, Forsyth D W. 1984. An improved method for determining the regional stress tensor using earthquake focal mechanism data: application to the San Fernando earthquake sequence. J Geophys Res Solid Earth, **89**(B11):9305~9320.
- Hagen L, Kahng A B. 2002. New spectral methods for ratio cut partitioning and clustering. IEEE Trans Comput Aided Des Integr Circuits Syst, **21**(9):1074~1085.
- Kagan Y Y. 1991. 3-D rotation of double-couple earthquake sources. Geophys J Int, **106**(3):709~716.
- Lund B. 2002. Correlation of microearthquake body-wave spectral amplitudes. Bull Seismol Soc Am, **92**(6):2419~2433.
- Martinez-Garzon P, Kwiatek G, Ickrath M, et al. 2014. MSATSI: A MATLAB package for stress inversion combining solid classic methodology, a new simplified user-handling, and a visualization tool. Seismol Res Lett, **85**(4):896~904.
- Ng A Y, Jordan M I, Weiss Y. 2002. On spectral clustering: analysis and an algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver: ACM, 849~856.
- Rousseeuw P J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math, **20**: 53~65.
- Shelly D R, Hardebeck J L, Ellsworth W L, et al. 2016. A new strategy for earthquake focal mechanisms using waveform-correlation-derived relative polarities and cluster analysis: application to the 2014 Long Valley Caldera earthquake swarm. J Geophys Res Solid Earth, **121**(12):8622~8641.
- Shi J B, Malik J. 2000. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell, **22**(8):888~905.
- Tian J H, Luo Y, Zhao L. 2019. Regional stress field in Yunnan revealed by the focal mechanisms of moderate and small earthquakes. Earth Planet Phys, **3**(3):243~252.
- Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. J Royal Stat Soc Ser B, **63**(2):411~423.

- Wu Z, Leahy R. 1993. An optimal graph theoretic approach to data clustering; theory and its application to image segmentation. *IEEE Trans Pattern Anal Mach Intell*, **15**(11) :1101~1113.
- Zhou D Y, Bousquet O, Lal T N, et al. 2003. Learning with local and global consistency. In: Proceedings of the 16th International Conference on Neural Information Processing Systems. Whistler: ACM, 321~328.
- Zoback M L. 1992. First- and second-order patterns of stress in the lithosphere; the world stress map project. *J Geophys Res Solid Earth*, **97**(B8) :11703~11728.

Research and Application of Spectral Clustering Method for Focal Mechanism Solutions

Lin Qingxi, Jiang Xijiao

Guangdong Earthquake Agency, Guangzhou 510070, China

Abstract The focal mechanism solutions, as a set of data with multiple dimensions, could be classified hardly by visual observation or simple comparative analysis. However, spectral clustering based on spectrogram theory has been approved to be well suited for nonlinear separable data, such as for focal mechanism solutions. In this paper, we attempted to use this method, that could cluster mass focal mechanisms fast and accurately by using the minimum rotation angle as the similarity matrix, to conduct Gap statistics for optimal numbers of clusters in order to normalize cut criterion (Neut) for classification discrimination. We verified the feasibility and reliability of this method through a set of random data sets, as well as verified the practicality of this method through the focal mechanisms for Haicheng M_s 7.3 earthquake sequences and Sichuan-Yunnan and its adjacent regions. The results showed that this new method reasonably subdivided the types of focal mechanisms and mechanism variations were strongly associated with corresponding hypocentral structure. In addition, this method has good applicability no matter for the focal mechanism solutions of a small area with a small amount of data or a large area with a large amount of data. In general, the spectral cluster analysis for focal mechanism solutions could serve as a method that has practical utility in determining the type of earthquake clusters.

Keywords: Focal mechanism solution; Spectral cluster; Gap statistics; The Haicheng M_s 7.3 earthquake sequences; Sichuan-Yunnan and its adjacent regions