DOI：10.11779/CJGE201408008

基于凝聚层次聚类分析法的岩体随机结构面产状优势分组

张 奇[1]，王 清[1]，阙金声[2]，李严严[1]，宋盛渊[1]
（1. 吉林大学建设工程学院，吉林 长春 130026；2. 华北电力设计院工程有限公司，北京 100120）

摘 要：在岩体斜坡稳定性分析和岩体水力学分析中，岩体随机结构面的优势分组是一项十分重要的内容。提出一种基于凝聚层次聚类分析的岩体随机结构面产状优势分组的新方法，这种方法的优点在于事先无需确定聚类中心，在分类结果生成后还可明显剔除数据的孤点与野值。应用人工随机生成的结构面产状数据对这种新方法和模糊 C 均值法进行了对比验证。结果表明，凝聚层次聚类分析法不仅在无孤值点的情况下分组结果优于模糊 C 均值算法，而且还可以有效地剔除孤值点对于分组结果的不利影响。最后将这种方法应用于松塔水电站坝肩结构面优势分组中，同样得到了比较满意的结果。

关键词：结构面；优势分组；凝聚层次聚类；模糊 C 均值
中图分类号：TU45 **文献标识码**：A **文章编号**：1000－4548(2014)08－1432－06

作者简介：张 奇(1989－)，男，吉林辽源人，硕士研究生，主要从事岩土力学及特殊土方面的研究。E-mail: zhangqi632005@163.com。

Dominant partitioning of discontinuities of rock masses based on AGNES

ZHANG Qi[1], WANG Qing[1], QUE Jin-sheng[2], LI Yan-yan[1], SONG Sheng-yuan[1]
(1. College of Construction Engineering, Jilin University, Changchun 130026, China; 2. North China Power Engineering Co., Ltd., Beijing 100120, China)

Abstract: A large number of random discontinuities are widely distributed in rock masses and have significant influences on the mechanical and hydraulic properties of fractured rock masses. In the analysis of the mechanical and hydraulic properties of fractured rock masses, the dominant partitioning of discontinuities of rock masses is an important part, and it is still a key for establishing the three-dimensional (3-D) network model of random discontinuities. A new method is proposed for the dominant partitioning of discontinuities of rock mass based on AGNES. In the proposed method we do not need to determine the centers of every cluster before clustering, and the acnodes or outliers can be eliminated effectively after clustering. Through the comparison of the proposed method and the fuzzy *C*-means method applied in the artificial and randomly generated data of discontinuities, the following conclusions can be drawn. The proposed method is a better method than the fuzzy *C*-means method in general cases, and it can get more accurate results by eliminating the acnodes or outliers. Finally, the proposed method is applied to a practical project, and the results are shown to be satisfactory.

Key words: discontinuity; dominant partitioning; AGNES; fuzzy *C*-mean

0 引 言

岩体经常是工程建筑的地基和环境。然而岩体的结构特征使岩体具有了不连续性和不均匀性，这也使得对于深入研究岩体的变形性、强度特征及裂隙水分布特征等问题变得极为复杂[1]。因此，为了解岩体结构特征的复杂性与不可知性，岩体随机不连续面三维网络数值模拟技术[2]应运而生。但对于这种模拟技术基础工作的岩体随机结构面产状优势分组的理论，则众说纷纭，并没有讨论出一个比较完善的方法。传统的方法一般采用极点图、等密度图以及玫瑰花图等作为分析的依据来判断，这种方法简单明了但却主观性强并不够准确[3]，而且对于数据量大且分布较为复杂的结构面，传统方法则无能为力[4]。目前，基于 K 均值法发展而来的模糊 C 均值聚类方法的应用较为普遍，且针对其需要事先确定初始聚类中心而影响其分类的准确性的不足，许多学者提出了改进意见[3, 5-7]。

基金项目：国家自然科学基金重点项目（41330636）
收稿日期：2013－12－10

但由于模糊 C 均值聚类分析方法对噪声和野值敏感这一最为主要的缺点并未得到有效的解决[8]。本文提出一种基于凝聚层次聚类分析的岩体随机结构面产状优势分组的新方法，这种方法事先无需确定聚类中心，在分类结果生成后可明显剔除数据的孤点与野值，并通过人工随机生成的结构面产状数据对这种新方法进行说明，最后将这种方法应用于松塔水电站坝肩结构面优势分组中，得到了比较满意的结果。

1　凝聚层次聚类分析法

凝聚层次聚类分析法是层次聚类分析法中较为常用的一种[9]，其思想是先将数据中的每一个样本做为一个聚类，然后每一次合并距离最近的两个聚类，直到达到终止条件或最终化为一类为止。该算法为典型的贪婪算法，先取局部最优最终优化为全局最优。

1.1　模型的建立

岩体随机结构面的产状通常由倾向和倾角两个随机变量来表示，实际野外结构面的调查也是通过测量结构面的倾向和倾角来描述结构面产状的。结构面的倾向表示空间上结构面所倾斜的方位，一般将正北设为 0°，顺时针旋转与正北的夹角即为倾向值，其值范围在 0°～360°。倾角为结构面与水平面之间的夹角，其值范围为 0°～90°。假设结构面为一平面，其产状就可以通过结构面的单位法向量表示出来。如图 1，X 轴与 Y 轴正方向分别表示正北和正东方位，Z 轴正方向为垂直向上。通过几何分析可以得出，结构面的单位法向量的表达式为 X=（x，y，z），其中

$$x=\cos\alpha\sin\beta \quad , \tag{1}$$

$$y=\sin\alpha\sin\beta \quad , \tag{2}$$

$$z=\cos\beta \quad , \tag{3}$$

式中，α 为倾向 β 为倾角。

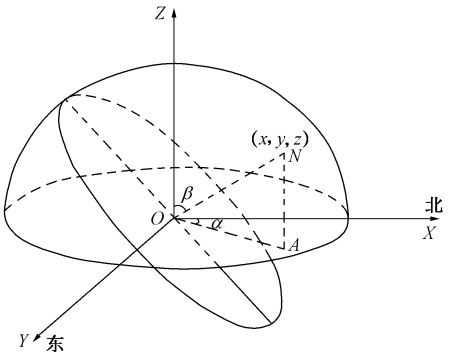

图 1　结构面产状空间模型

Fig. 1 Spatial model of discontinuity occurrence

度量结构面产状数据之间距离的方法有很多，为满足结构面产状数据分组具体特征的要求，本文采用结构面法向量夹角的正弦值，其具体表达式为

$$\sin\theta=[1-(X_i\bullet X_j^{\mathrm{T}})^2]^{\frac{1}{2}} \quad 。 \tag{4}$$

凝聚层次聚类分析法先将每一个结构面作为一组聚类，假设共有 n 个结构面，那么就有 n 组聚类。通过式（4）距离的计算，可以得到每一个结构面与其他结构面之间的距离，即 $n(n-1)$ 个距离值，并可以找到距离最短的两个结构面，把他们聚为一组，这时结构面聚类组数就变为了 n−1。然后，计算这 n−1 组聚类相互之间的距离，仍然将距离最小的两组合并为一组，如此往复，聚类数将达到预先设定的组数或最终聚为一组，其基本思路见图 2。

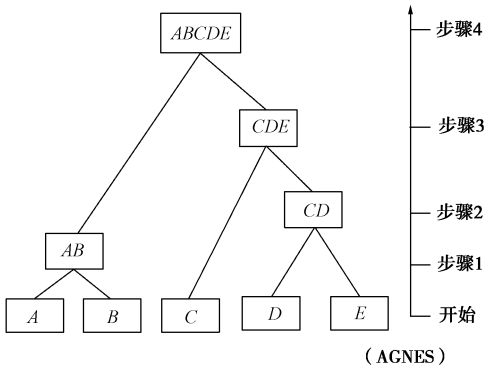

图 2　凝聚层次聚类图

Fig. 2 Model of AGNES

当计算聚类与聚类（如图 2 中 AB 与 C 或 AB 与 CD 或 AB 与 CDE）之间的距离的方法有很多，如最小距离法、最远距离法、中心距离法和平均距离法等。其中，最小距离法可以发现数据分布形状拉长或不规则的类别，但不适合寻找数据分布形状紧密的类别；最远距离法偏向于产生直径大致相等的类别，并且受异常值的影响较大；中心距离法比较不容易受异常值的影响，但其他方面并不如平均距离法。故本文采用较为常用平均距离法，即用两类中所有两两结构面之间距离的平均值作为两组聚类的距离，其计算公式为

$$D_{pq}=\frac{1}{n_p n_q}\sum_{i\in G_p}\sum_{j\in G_q}d_{ij} \quad 。 \tag{5}$$

式中　G_p 和 G_q 分别表示两个聚类，D_{pq} 表示两个聚类间的距离值，n_p 和 n_q 表示 G_p 和 G_q 聚类中结构面的条数，d_{ij} 表示 G_p 中的 i 结构面与 G_q 中的 j 结构面之间的距离，可通过式（4）计算求得。

因此，凝聚层次聚类分析法是典型的贪婪算法，它先将每一个数据点定义为一组，然后以类与类之间的距离作为量度，逐步将距离较小的类聚在一起，并最终优化为全局的特定组数的聚类结果。图 2 表示了最终分为一组的二叉树枝图，计算时是由每一个数据点为一组的 5 组经过 4 步的计算最终分为 1 组。但分

1 组并不一定是最终的最优分组数，如果没有预先规定分组数，最优的分组数需要由不同分组数的聚类有效性指标的比较来确定，相关内容将会在模型的检验中具体阐述。从图 2 中可以明确看出 *ABCDE* 5 个数据点不同分组数的具体情况，如分 1 组即为 *ABCDE*，分 2 组为 *AB* 和 *CDE*，分 3 组为 *AB*，*C* 和 *DE*，分 4 组为 *AB*，*C*，*D* 和 *E*，分 5 组即为每 1 个数据点 1 组，故可以得到不同聚类数的具体分组情况。

1.2 模型的检验

由于评价聚类有效性的指标有很多[10]，本文检验模型好坏与确定最佳聚类数采用的指标为 Xie-Beni 有效性指标。Xie 等从数据集的几何结构出发，在 1991 年提出了 Xie-Beni 有效性指标[11]，其具体表达式为

$$\mathrm{XB}(U,V,c)=\frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-x_j\|^2}{\min\limits_{i\neq j}\|v_i-v_j\|}\quad,\quad(6)$$

式中，U 为隶属矩阵，V 为聚类中心矩阵，c 为聚类数，m 为模糊因子，u_{ij} 为 U 矩阵中的元素，v_i 为 V 矩阵中的第 i 行元素。XB 为类内紧凑度和类间分离度的比例，在类内紧凑度和类间分离度之间找一个平衡点，使其达到最小，从而获得最好的聚类效果。因此，同一方法具有较小的 XB 指标的聚类数为最优分组数，不同方法最优分组数 XB 指标较小的方法为较好的方法。

2 人工随机生成数据的模拟

依据结构面产状的倾向倾角值服从二维正态分布[1]，本文随机生成了 4 组结构面，其具体的二维正态分布参数见表 1。

表 1 倾向倾角的二维正态分布参数表

Table 1 Parameters of bivariate normal distributions of clusters

组号	倾向平均值/(°)	倾角平均值/(°)	结构面条数
1	165	55	80
2	345	40	20
3	30	65	40
4	90	75	60

通过人工随机生成的数据绘制极点图如图 3。为了避免初始聚类中心对模糊 *C* 均值聚类方法的影响，本文采用人工模拟数据的均值作为该方法的初始聚类中心，即初始聚类中心为（165，55），（345，40），（30，65）和（90，75）。通过模糊 *C* 均值方法模拟得到的分组情况绘制极点图如图 4。通过凝聚层次聚类算法对人工生成的数据进行模拟，得到的分组情况绘制的极点图如图 5。模糊 *C* 均值法与凝聚层次聚类法分组的具体数据对比见表 2。

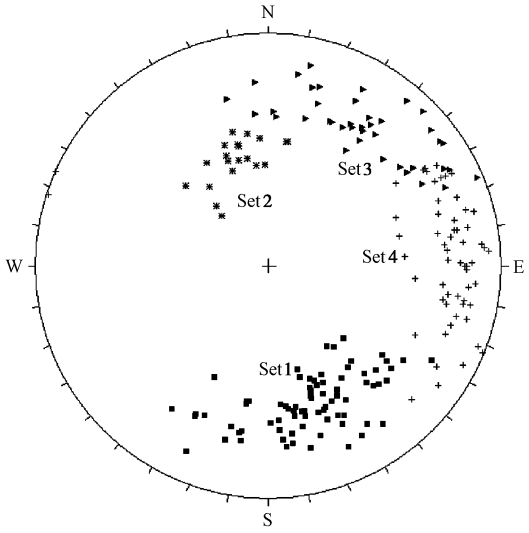

图 3 随机生成数据的分组极点图

Fig. 3 Pole of artificial discontinuity data

图 3 表示随机生成数据的分组极点图，从图中可以看出随机生成的分组并不是最佳的优势分组结果，如第 3 组和第 4 组边缘相交的部分就没有很好的分开。但其大致的分组情况及聚类中心值仍具有很好的参照价值。

从图 4，5 与图 3 的对比可以看出，凝聚层次聚类分组的极点图比模糊 *C* 聚类方法分组的极点图更为接近随机生成数据的极点图，即凝聚层次聚类方法更能反映随机生成数据的分组情况；从表 2 两种方法分组所得到的具体数据也可以看出，凝聚层次聚类方法的聚类中心与随机生成数据的均值更为接近；凝聚层次聚类模拟的 XB 指标不仅比模糊 *C* 均值模拟的 XB 指标小很多，而且也小于随机生成的分组情况的 XB 指标值，这说明了凝聚层次聚类方法使得各组聚类更为紧密，组与组之间更为分离，聚类结果更为有效和可靠。在大量结构面产状测量的实际过程中，由于各种

表 2 模糊 *C* 均值法与凝聚层次聚类法分组的具体数据对比表

Table 2 Comparison of specific clustering results between FCM and AGNES

方法	第一组			第二组			第三组			第四组			XB 指标
	倾向/(°)	倾角/(°)	条数	倾向/(°)	倾角/(°)	条数	倾向/(°)	倾角/(°)	条数	倾向/(°)	倾角/(°)	条数	
随机生成	165	55	80	345	40	20	30	65	40	90	75	60	0.1248
C 均值模拟	163.47	52.46	78	354.21	43.61	33	56.55	66.16	44	98.73	69.17	45	0.1577
凝聚层次模拟	163.56	52.82	80	348.45	43.14	26	44.22	66.35	43	95.07	71.03	51	0.1001

外在因素使得结构面产状测量结果产生一些误差是在所难免的，这就有可能使结构面产状数据中出现噪声点和野值。模糊 *C* 聚类方法对于这种噪音点和野值极为敏感，有时甚至会对整体的聚类效果影响巨大。而凝聚层次聚类算法则会有效的将噪声点和野值剔除，使结构面优势分组结果有效且可靠。

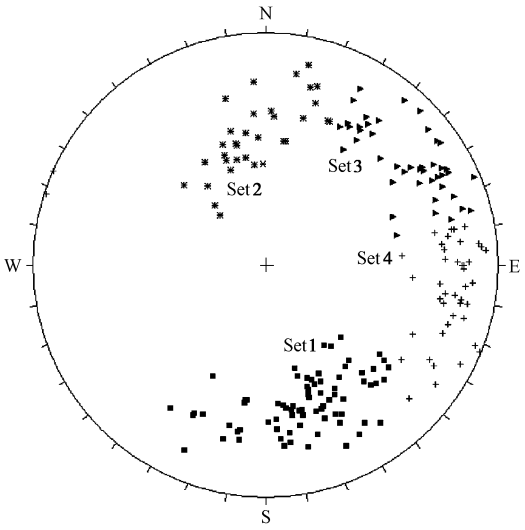

图 4 模糊 *C* 聚类方法分组极点图

Fig. 4 Pole of discontinuity data in FCM

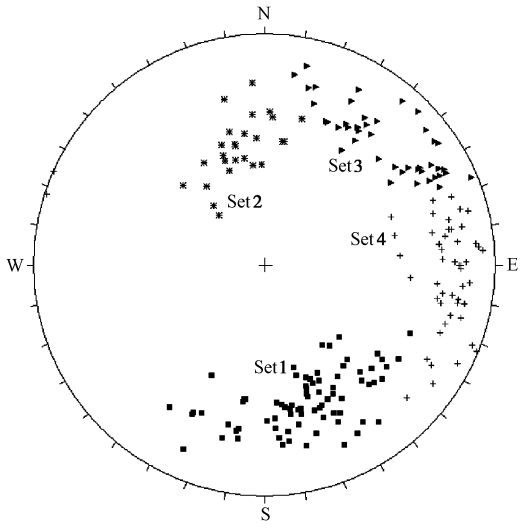

图 5 凝聚层次聚类法分组极点图

Fig. 5 Pole of discontinuity data in AGNES

为了说明凝聚层次聚类法对于含有噪声点和野值数据聚类的有效性，人为的在由表 1 生成的随机结构面产状数据中加入了 3 个噪声点，并绘制加入孤值点的随机生成数据的极点图如图 6。分别应用凝聚层次聚类算法和模糊 *C* 均值算法对加入孤值点的数据进行聚类分析，将聚类组数为 2～10 的 XB 指标列于表 3。由表 3 可以看出凝聚层次聚类算法的最佳分组数为 5，绘制分 5 组后的极点图如图 7；模糊 *C* 均值算法的最佳分组数为 3，分别绘制分 3 组、4 组和 5 组的极点图如图 8～10。

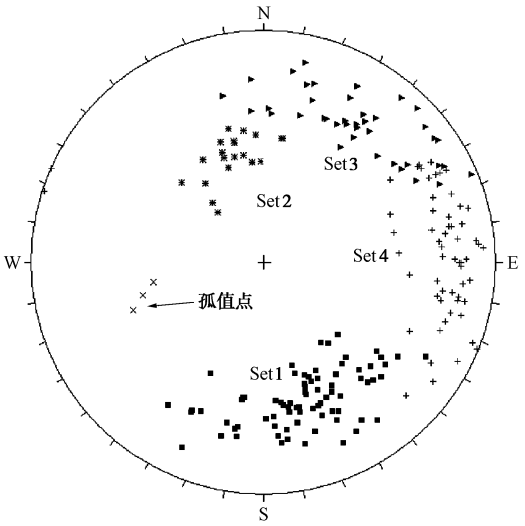

图 6 加入孤值点的随机生成数据的极点图

Fig. 6 Pole of artificial discontinuity data with single points

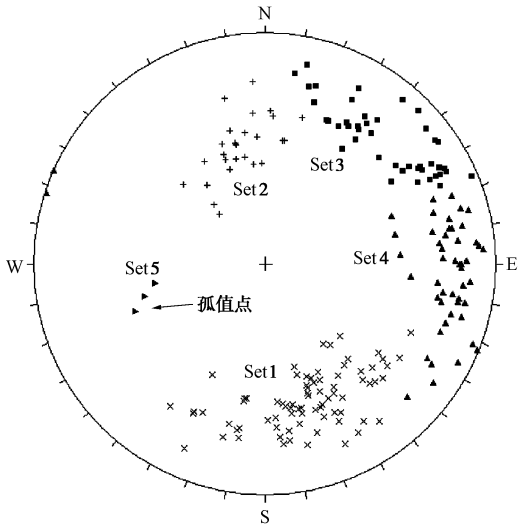

图 7 凝聚层次聚类分析法分组的极点图

Fig. 7 Pole of discontinuity data in AGNES

表 3 两种方法不同聚类数的 XB 指标表

Table 3 Values of XB of different clusters in two methods

聚类数	2	3	4	5	6	7	8	9	10
凝聚层次法	0.1703	0.098	0.094	0.0888	0.1196	0.1183	0.1046	0.1413	0.1256
模糊 *C* 均值法	0.2107	0.1241	0.1628	0.3301	0.2732	0.1947	0.453	0.3849	0.3105

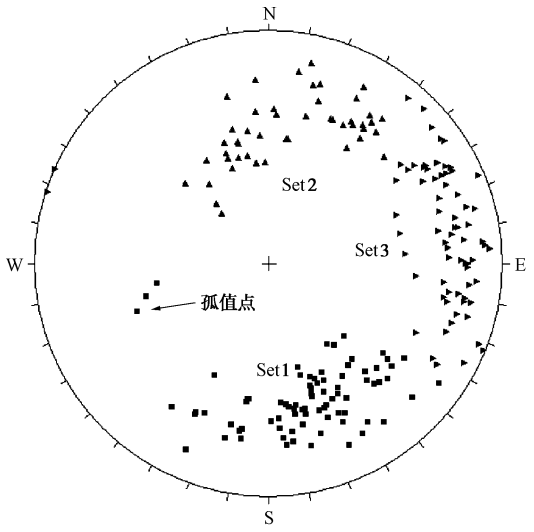

图 8 模糊 *C* 均值法分 3 组极点图

Fig. 8 Pole of discontinuity data in FCM with three clusters

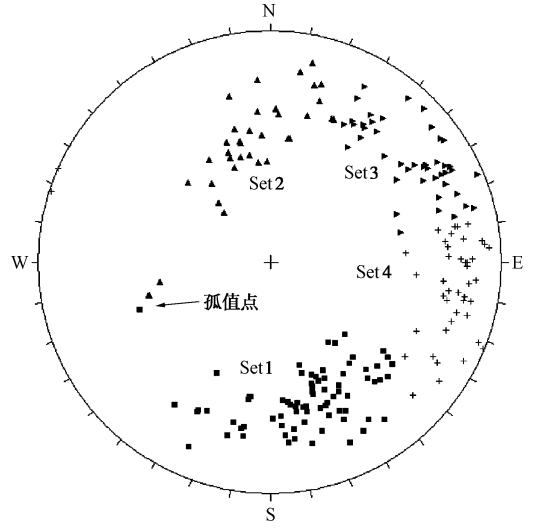

图 9 模糊 *C* 均值法分 4 组极点图

Fig. 9 Pole of discontinuity data in FCM with four clusters

从图 7 中可明显的看出，凝聚层次分析法很好的将 3 个孤值点化为 1 组，其余 4 组的分组情况与未加入孤值点的情况相同，这就有效的剔除了孤值点对整个数据聚类及聚类中心的影响。相反对于模糊 *C* 均值聚类分析的方法，无论是 XB 指标最小的分 3 组的情况，还是分 4 组和 5 组的情况，该方法都没能有效的避免孤值点对于聚类分组的影响，甚至可看出，应用模糊 *C* 均值聚类方法对于加入 3 个孤值点的数据的模拟结果与未加入孤值点的模拟结果是有很大不同的。

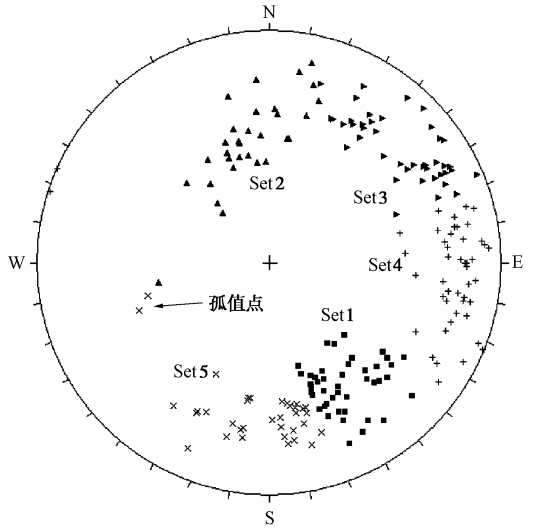

图 10 模糊 *C* 均值法分 5 组极点图

Fig. 10 Pole of discontinuity data in FCM with five clusters

3 凝聚型层次聚类法在松塔水电站坝址勘察中的应用

本文应用凝聚层次聚类分析方法，采用松塔水电站坝址勘察过程所采集到的结构面产状数据进行结构面产状的优势分组，其中坝肩一平洞中结构面共计 305 条，聚类数为 2～10 的 XB 指标为 0.2286，0.1032，0.1014，0.1048，0.16，0.1168，0.1282，0.1104 和 0.1055。最佳分组情况见图 11，每一组的平均产状和条数见表 4。

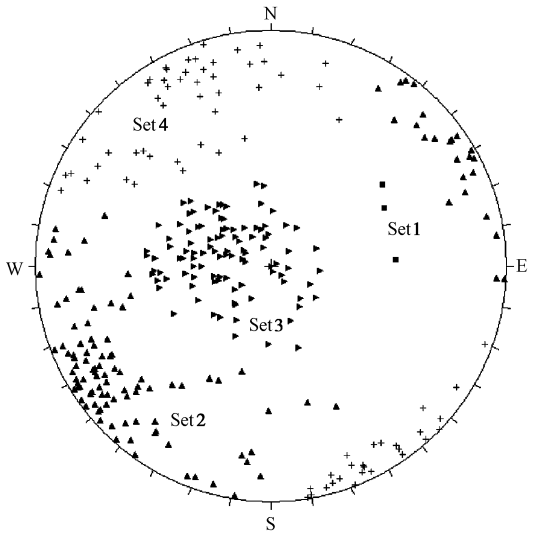

图 11 松塔数据优势分组结果

Fig. 11 Clustering results of Songta data

表 4 松塔数据优势分组结果

Table 4 Custering results of Songta data

第一组			第二组			第三组			第四组			XB
倾向 /(°)	倾角 /(°)	条数	倾向 /(°)	倾角 /(°)	条数	倾向 /(°)	倾角 /(°)	条数	倾向 /(°)	倾角 /(°)	条数	指标
67.51	45.15	3	235.56	77.76	124	283.38	16.19	105	330.68	80.2	73	0.1014

从表 4 和图 11 可以看出，第一组节理明显是噪声点应予以剔除，所以平洞结构面的优势分组数为 3 组，各组的平均产状及条数不变。

4 结　论

应用凝聚层次聚类方法进行结构面的优势分组不仅不需要事先设定初始聚类中心，而且可以剔除掉噪声点和野值对于聚类结果的影响，其方法明显优于模糊 *C* 均值聚类方法，因此通过本文的研究可以得到以下结论：

（1）在对人工生成的随机机构面产状数据进行模拟的过程中，凝聚层次聚类方法明显优于模糊 *C* 均值聚类算法。

（2）在对添加了孤值点的随机结构面产状数据进行模拟的过程中，模糊 *C* 均值并没有剔除噪音点，而且噪音点对其聚类分组结果影响很大。凝聚层次聚类算法则很好的避开了孤值点的影响，对数据进行了有效的分组。

（3）应用凝聚层次聚类分析法对松塔水电站坝肩岩体结构面进行了有效的模拟，剔除了某平洞中的孤值点，得到了比较满意的分组结果。

参考文献：

[1] 陈剑平, 肖树芳, 王　清. 随机不连续面三维网络计算机模拟原理[M]. 长春: 东北师范大学出版社, 1995. (CHEN Jian-ping, XIAO Shu-fang, WANG Qing. Computer simulation principle of 3D network numerical modeling tenique[M]. Changchun: Northeast Normal University Press, 1995. (in Chinese))

[2] 陈剑平. 岩体随机不连续面三维网络数值模拟技术[J]. 岩土工程学报，2001, **23**(4): 398－403. (CHEN Jian-ping. 3D network numerical modeling technique for random discontinuities of rockmass[J]. Chinese Journal of Geotechnical Engineering, 2001, **23**(4): 398 － 403. (in Chinese))

[3] 蔡美峰, 王　鹏, 赵　奎, 等. 基于遗传算法的岩体结构面的模糊 *C* 均值聚类方法[J]. 岩石力学与工程学报, 2005, **24**(3): 371－376. (CAI Mei-feng, WANG Peng, ZHAO Kui, et al. Fuzzy C-means cluster analysis based genetic algorithm for automatic of joint sets[J]. Chinese Journal of Rock Mechanics and Engineering, 2005, **24**(3): 371 － 376. (in Chinese))

[4] 陈剑平, 石丙飞, 王　清. 工程岩体随机结构面优势方向的表示法初探[J]. 岩石力学与工程学报, 2005, **24**(2): 241 －245. (CHEN Jian-ping, SHI Bing-fei, WANG Qing. Study on the dominant orientations of random fractures of fractured rock masses[J]. Chinese Journal of Rock Mechanics and Engineering, 2005, **24**(2): 241 － 245. (in Chinese))

[5] 周玉新, 周志芳, 孙其国. 岩体结构面产状的综合模糊聚类分析[J]. 岩石力学与工程学报, 2005, **24**(13): 2283－2287. (ZHOU Yu-xin, ZHOU Zhi-fang, SUN Qi-guo. Synthetic fuzzy clustering analysis for joints occurrence of rock mass[J]. Chinese Journal of Rock Mechanics and Engineering, 2005, **24**(13): 2283 － 2287. (in Chinese))

[6] XU L M, CHEN J P, WANG Q, et al. Fuzzy *C*-means cluster analysis base on mutative scale chaos optimization algorithm for the grouping of discontinuity sets[J]. Rock Mechanics and Rock Engineering, 2013(46): 189 － 198.

[7] 冯　羽, 马凤山, 巩城城, 等. 节理岩体结构面优势产状确定方法研究[J]. 工程地质学报, 2011, **19**(6): 887 － 892. (FENG Yu, MA Feng-shan, GONG Cheng-cheng, et al. Data analysis method for optimized and dominant orientations of joints in rock mass[J]. Journal of Engineering Geology, 2011, **19**(6): 887 － 892. (in Chinese))

[8] 陈东辉. 基于目标函数的模糊聚类算法关键技术研究[D]. 西安: 西安电子科技大学, 2012. (CHEN Dong-hui. Research of key techniques in fuzzy clustering based on objective function[D]. Xi'an: Xidian University, 2012. (in Chinese))

[9] 李仁义. 数据挖掘中聚类分析算法的研究与应用[D]. 成都: 电子科技大学, 2012. (LI Ren-yi. Research and application of the algorithm of clustering analysis in data mining[D]. Chengdu: University of Electronic Science and Technology of China, 2012. (in Chinese))

[10] 唐明会. 模糊聚类有效性研究[D]. 西安: 西安交通大学, 2010. (TANG Ming-hui. Research on fuzzy clustering validity[D]. Xi'an: Southwest Jiaotong University, 2010. (in Chinese))

[11] XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, **8**(13): 841 － 847.