文章编号: 1008- 2786- (2006) 4- 442- 04

区域泥石流危险度评价的投影寻踪动态聚类方法

倪长健[1, 2], 崔 鹏[1]
(1. 中国科学院水利部成都山地灾害与环境研究所, 四川 成都 610041; 2 成都信息工程学院, 四川 成都 610041)

摘 要: 在对投影寻踪聚类方法深入分析的基础上, 首次把投影寻踪聚类的思想和动态聚类方法结合起来构造投影指标, 提出了投影寻踪动态聚类 (PPDC)新方法, 新方法在整个运算过程中完全由自身数据驱动, 毋需人为给定参数, 具有聚类结果客观、明确以及稳定性好、操作简便等特点。计算结果表明, PPDC法应用于区域泥石流危险度评价不仅切实可行, 而且能克服了现有评价方法的不足, 取得了令人满意的实际应用效果。上述应用为同类问题的研究开创了一条新途径, 也进一步为生态保护、防灾减灾提供科学依据和技术支持。

关键词: 投影寻踪; 动态聚类; 泥石流; 危险度评价
中图分类号: P642 23 文献标识码: A

泥石流灾害的发生、发展受到多种不可控制的随机因素的影响。由于它对土地资源、生态环境和人民生命财产安全具有极大的破坏力, 因此, 建立区域泥石流危险度评价的指标体系, 并通过特定的数据处理技术对其危险程度进行客观、准确的评价, 不仅是泥石流研究的核心问题之一, 也是环境保护和减灾对策研究的重要内容。泥石流危险度区划是在区域泥石流规律研究的基础上, 根据区域泥石流危险度和发育状况确定其危险等级。针对影响泥石流危险度评价的诸多因素的不确定性, 研究人员先后提出了一些新的研究方法, 如关联度分析法、模糊综合评价法和神经网络法等[1-3]。上述方法的提出深化了泥石流危险度评价这一问题的研究, 但是他们往往存在评价过程中诸如权重确定没有统一的理论和计算公式、神经网络结构试算确定、评价过程相对复杂等不足, 故在客观性和可操作性等方面尚存在一定的局限性。投影寻踪[4] (projection pursuit, 简称 PP)是直接由样本数据驱动的探索性数据驱动方法, 可用于具有任何结构的高维数据的分析。基于投影寻踪的聚类 (projection pursuit cluster, 简称 PPC)方法的基本思想是把高维数据样本通过某种组合投影到低维空间中, 对于投影到的构形, 采用目标函数 (投影指标函数)来衡量投影暴露某种分类结构的可能性的大小, 寻找出使目标函数达到最优 (即能反映高维数据结构或特征)的投影值, 然后根据该投影值对样本集进行相应的分类。PPC法较好地克服了上述评价方法的不足[5], 但通过对其分析, 不难发现密度窗宽这一重要参数依赖经验给定、寻优过程和聚类结果的脱节以及模型的稳定性差是 PPC法存在的问题。在上述分析的基础上, 我们基于投影寻踪聚类的思想, 结合动态聚类方法[6], 提出了投影寻踪动态聚类 (projection pursuit dynamic cluster, 简称 PPDC)新方法。本文将详细介绍投影寻踪动态聚类法的实现过程, 并将其应用于区域泥石流危险度评价, 获得了令人满意的结果。

1 投影寻踪动态聚类方法

1.1 投影寻踪动态聚类方法实现的思想

对于任一投影方向, 基于某一聚类准则的样本

收稿日期 (Received date): 2006- 02- 01; 改回日期 (Accepted): 2006- 05- 16。
基金项目 (Foundation item): 国家自然科学基金重点项目 (90202007)、成都信息工程学院科技发展基金 (CSRF200501)。 [Supported by Key National Science Foundation (90202007) and the Scientific and Technical foundation of Chengdu University of Information Technology.]
作者简介 (Biography): 倪长健 (1970-), 男, 博士, 副教授, 现主要从事山地生态研究。 [Ni Changjian (1970-), male, born in 1970, PH. D, associate professor, works mainly on mountain environment and ecology.]

分类结果是确定的, 记整个样本的投影特征值序列组成的集合为 $\Omega=\{z_1, z_2, \cdots, z_n\}$, 要将它们分成 K 类, 采用动态聚类法[6], 实现步骤如下:

①随机选取 K 个点作为 K 个聚核, 记为 $L^0=(A_1^0, A_2^1, \cdots, A_K^0)$;

②根据 L^0, 把 Ω 中的点分为 K 类, 记为 $P^0=(P_1^0, P_2^0, \cdots P_K^0)$;

其中

$$P_i^0=\{z\in\Omega \mid d(A_i^0-z)\leqslant d(A_j^0-z), \forall j=1,2,\cdots,K, j\neq i\}$$

$d(A_i^0-z)$ 为点 A_i^0 和集合 Ω 中任一点的绝对值距离。

③由 P^0 发, 计算新的聚核 L^1, $L^1=(A_1^1, A_2^1, \cdots, A_K^1)$;

其中 $A_i^1=\frac{1}{n_i}\sum_{z\in P_i^0} z_i$, 类 P_i^0 中有 n_i 个点。

④重复以上步骤, 由此得到一个分类结果序列 $V^m=(L^m, P^m)$, $m=1, 2, \cdots$。记 $D(A_i^m, P_i^m)=\sum_{z\in P_i^m}|z-A_i^m|$, $u_m=\sum_{i=1}^{K}D(A_i^m, P_i^m)$, 则算法的终止判断条件是 $\frac{|u_{m+1}-u_m|}{u_{m+1}}\leqslant\varepsilon$, 式中, ε 是充分小的误差允许值。理论证明这种算法是收敛的[6]。

投影寻踪聚类思想的具体实现方式是在力求整个样本序列投影特征值尽可能散开的同时, 又使相似样本投影特征值尽可能聚在一起, 目标函数的构造是实现这一思想的关键。由于聚类分析是根据待评价样本的数据特性将样本进行分类和评价, 因此可以用投影特征值来构造目标函数。基于上述动态聚类结果, 目标函数可以用整个样本投影特征值的分散程度和类内样本投影特征值的聚集程度之差来表示, 即 $ss(a)-dd(a)$, 前者定义为投影分散度 $ss(a)=\sum_{\substack{zm\in\Omega\\ zj\in\Omega}} d(zm, zj)$, 其值愈大, 则整个样本投影特征值离散程度越高; 后者定义为类内聚集度 $dd(a)=\sum_{i=1}^{K}D(P_i)$, 其中 $D(P_i)=\sum_{\substack{zm\in P_i\\ zj\in P_i}} d(zm, zj)$, $dd(a)$ 愈小, 则相似样本的聚集程度越高。对此目标函数的求解就是寻求一投影方向满足 $ss(a)-dd(a)$ 取得最大值, 显然, 投影分散度越大或类内聚集度越小, 则目标函数越大, 这正是投影寻踪聚类建模思想的体现。

由上可见, 此目标函数消除了 PPC法中密度窗宽确定的人为任意性[5], 把寻优过程和聚类结果紧密结合起来, 因此, 以其为核心的 PPDC 方法能有效地克服 PPC法存在的不足, 具有结构清晰、稳定性好、分类结果明确、客观性强等特点。

1.2　投影寻踪动态聚类方法实现的步骤

设待研究区域泥石流危险度评价指标的样本数据集为 $x_{ij}^0(i=1, \cdots, n; j=1, \cdots, m)$; n 为样本个数, m 为评价指标的个数), 投影寻踪动态聚类方法的实现步骤如下:

步骤 1: 预处理。由于各同类评价指标数值范围可能相差较大, 为了消除极端异常值对聚类结果造成的影响, 找出满足下式的样本个体, 并将其从样本集中加以剔除

$$x_{ij}^0 > 10^* \bar{x}_j \tag{1}$$

另外, 为消除各评价指标的量纲效应, 使建模具有通用性, 采用下式对剔除后的评价指标值进行极值归一化处理

$$x_{ij}=[x_{ij}^0-x_{j\min}]/[x_{j\max}-x_{j\min}] \tag{2}$$

$\bar{X}_j$, $x_{j\min}$、$x_{j\max}$分别为样本数据集中第 j个指标值的平均值、最小值和最大值。

步骤 2: 线性投影。针对区域泥石流危险度评价这一多因素影响问题, 线性投影实际就是给出一投影方向, 把多因素复杂问题转化为一维投影特征值序列, 将高维数据投影到线性空间进行研究, 设 $\vec{a}$ 为 m 维单位向量, 则 x_{ij}的投影特征值 z_i 可用式 (3) 描述

$$z_i=\sum_{j=1}^{m}a_j x_{ij} \tag{3}$$

步骤 3: 目标函数的构造。综合投影值时, 要求投影值 z_i 的散布特征为: 类内的投影点尽可能密集, 而类间投影点尽可能散开。基于 1.1的分析, 目标函数 $QQ(a)$定义为投影分散度与类内聚集度之差, 即

$$QQ(a)=ss(a)-dd(a) \tag{4}$$

步骤 4: 优化投影方向。当给定区域泥石流危险度评价指标样本数据时, 目标函数 $QQ(a)$只随投影方向 $\vec{a}$的变化而变化。模型求解的关键是找到能反映系统特征的最优投影方向, 根据上述分析可知, 当式 (4)取得最大值时所对应的 $\vec{a}$就是最优投影方向向量。所以, 此问题可转化为式 (5)描述的优化问题

$$\begin{cases}\max QQ(a)\\ \|a\|=1\end{cases} \tag{5}$$

免疫进化算法[7]是受生物免疫机制启发而得到的一种具有优良性能的进化算法, 可被用于解决上述复杂的非线性优化问题。

步骤 5 综合分析。把由式(5)求得的最佳投影方向 a^* 代入式(3)即得到各样本的投影特征值 $z^*(i)$。$a^*(i)$值一方面反映了区域泥石流危险度的大小,另一方面,基于动态聚类法,由其还可以进一步得到最终的聚类结果。

2 实例分析

文献[1]从泥石流分布情况和影响泥石流形成条件的地质、地貌、水文气象、森林植被和人类活动5个方面着手,选择了18项候选指标,采用灰色关联度分析方法,确定出8个定量指标来综合判定区域泥石流危险度,具体的指标体系由泥石流分布密度 x_1(条/10^3 km^2)、岩石风化程度系数 x_2、断裂带密度 x_3[km/(10^3 km^2)]、≥25°坡地面积百分比 x_4(%)、洪灾发生频率 x_5(%)、月降雨量变差系数 x_6、年平均≥25 mm大雨日数 x_7(日)和≥25°坡耕地面积百分比 x_8(%)8个指标构成。基于四川省阿坝州10个县市的泥石流危险区划8项定量指标的基础资料[1],样本数据见表1,应用投影寻踪动态聚类法对其进行评价,将这10个样本聚为4类,依据该方法的计算步骤,其中 $n=10, m=8$,经过运算得最优目标函数值为24.5114,最优投影方向向量为

$$a^* = (0.7515, 0.0435, 0.1622, 0.0007, 0.5514, 0.0004, 0.0003, 0.3208)$$

各样本的投影特征值和聚类结果也一并列在表1,投影特征值越大,表示该区域泥石流危险程度越高。

表1 泥石流危险性的评价指标及评价结果

Table 1 Indexes and results of regional dangerous degree of debris flow evaluation

序号	县名	区域泥石流危险度评价指标								投影特征值	PPDC法聚类结果
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8		
1	汶川	24.76	1.85	95.38	32.96	23.08	0.78	1.28	31.72	1.0403	显著危险
2	理县	42.02	1.80	18.94	41.90	57.14	0.68	0.93	4.56	1.4047	高度危险
3	茂县	26.69	1.89	151.85	30.80	42.86	0.78	1.45	8.32	1.1697	显著危险
4	黑水	23.92	1.79	20.34	44.54	25.00	0.77	2.68	13.49	0.8016	中度危险
5	松潘	7.88	1.76	65.97	61.56	14.29	0.71	1.40	19.72	0.4709	轻度危险
6	马尔康	5.99	1.69	8.19	76.01	16.67	0.86	2.50	1.63	0.2345	轻度危险
7	壤塘	5.45	1.67	8.11	91.66	16.67	0.93	1.83	1.31	0.2201	轻度危险
8	金川	40.99	1.73	12.59	65.91	57.14	0.89	1.38	36.89	1.6565	高度危险
9	小金	19.08	1.82	2.82	48.45	14.29	0.83	1.342	2.66	0.6605	中度危险
10	九寨沟	19.13	1.82	79.70	33.89	42.86	0.80	1.73	5.62	0.8711	中度危险

由表1可见:从投影特征值的大小来看,金川这一地区的泥石流危险度最大,壤塘的泥石流危险度最小。从聚类结果来看,金川和理县聚为一类,二者的泥石流分布密度分别为40.99条/(10^3 km^2)和42.02条/(10^3 km^2),区域内泥石流十分发育,以金川为例,曾分别于1926、1980~1988年共10次发生过群发性大规模泥石流[1],应为泥石流的高度危险区;茂县和汶川聚为一类,二者的泥石流分布密度分别为28.69条/(10^3 km^2)和24.76条/(10^3 km^2),区域内泥石流也很发育,历史上数次发生过群发性大中规模泥石流,应为泥石流显著危险区;黑水、小金和九寨沟聚为一类,其对应的泥石流分布密度分别为23.92条/(10^3 km^2)、19.08条/(10^3 km^2)、19.13条/(10^3 km^2),区域内泥石流比较发育,历史上数次发生过群发性中等规模泥石流,应为泥石流中度危险区;松潘、马尔康和壤塘聚为一类,其对应的泥石流分布密度分别为7.88条/(10^3 km^2)、5.99条/(10^3 km^2)、5.45条/(10^3 km^2),区域内泥石流较少发育,曾分别于1940、1949、1953、1957、1982、1983和1987年共7次发生过群发性中小规模泥石流[1],应为泥石流轻度危险区。基于PPDC法得出的区域泥石流危险度排序和聚类结果和文献[1]是基本一致的,和实际情况也是吻合的。

应用分析表明,投影寻踪动态聚类方法具有如下特点:(1)在整个运算过程中,仅仅需要预先给定样本的聚类数,而不需要人为给定其他任何参数,避免了投影寻踪聚类方法中密度窗宽确定的人为任意性,因此,投影寻踪动态聚类法具有客观性和普适

性, 便于实际应用推广。(2)与文献[15]相比较, 投影寻踪动态聚类法不但可以直接得到投影特征值和最优投影方向向量, 而且可以得到明确的聚类结果, 避免了结果的经验判定。

3 结论

1. 投影寻踪动态聚类方法在继承投影寻踪聚类的思想和优点的同时, 克服了在实际应用中其密度窗宽确定人为任意性以及聚类结果依赖直观判定等弱点, 因而该方法具有结构清晰、稳定性好、分类结果明确、客观性强等特点。

2. 投影寻踪动态聚类方法在区域泥石流危险度评价中的应用是成功的, 为这方面的研究提供了一条新途径。

3. 此研究为区域泥石流危险度指标体系的精简奠定了基础。

参考文献 (References)

[1] Liu Xilin, Mo Duowen. Risk Assessment on Debris Flow [M]. Chengdu: Sichuan Science and Technology Press, 2003. [刘希林, 莫多闻. 泥石流风险评价 [M]. 成都: 四川科学技术出版社, 2003.]

[2] Wei Yongming. Application of Relativity Analysis Method and Fuzzy Synthetic Assessment Method in classification of dangerous degree of debris flow—a case study of Huairou and Miyun county in Beijing Suburb [J]. *Journal of Natural Disasters*, 1998, 7(2): 109~117 [魏永明. 关联度分析法和模糊综合评判法在泥石流沟谷危险度划分中的应用—以北京市郊区怀柔、密云县为例 [J]. 自然灾害学报, 1998, 7(2): 109~117]

[3] Wang Mingwu. Regional classification of dangerous degree of debris flow based on Neural Network [J]. *Hydrogeology and Engineering-geology*, 2000, 27(2): 18~19 [汪明武. 基于神经网络的泥石流危险度区划 [J]. 水文地质与工程地质, 2000, 27(2): 18~19]

[4] Friedman J H, Turkey J W. A projection pursuit algorithm for exploratory data analysis[J]. *IEEE Trans on Computer*, 1974, 23(9): 881~890

[5] Wang Mingwu. Application of new Projection Pursuit Method to evaluation of dangerous degree of debris flow [J]. *Journal of Soil and Water Conservation*, 2002, 16(6): 79~81 [汪明武. 投影寻踪新方法在泥石流危险度评价中的应用 [J]. 水土保持学报, 2002, 16(6): 79~81]

[6] Ren Ruoen, Wang Huiwen. Multi-Dimensional Statistics Data Analysis-Theory, Method and Practice[M]. Beijing: National Defence Industry Press, 1999. 76~80 [任若恩, 王惠文. 多元统计数据分析—理论、方法、实例 [M]. 北京: 国防工业出版社, 2000. 148~151]

[7] Ni Changjian, Ding Jing. Immune Evolutionary Algorithm[J]. *Journal of Southwest Jiaotong University*, 2003, 38(1): 87~91 [倪长健, 丁晶. 免疫进化算法 [J]. 西南交通大学学报, 2003, 38(1): 87~91]

Projection Pursuit Dynamic Cluster Method for Evaluating Regional Dangerous Degree of Debris Flow

NI Changjian[1, 2], Cui Peng[1]

(1. *Institute of Mountain Hazards and Environment, China Academy of Sciences, Chengdu* 610041, *China*;
2. *Chengdu University of Information and Technology, Chengdu* 610041, *China*)

Abstract: Based on penetrating analysis of the projection pursuit cluster method (PPC), a projection pursuit dynamic cluster (PPDC) method, which combines dynamic cluster method with projection pursuit principle, is proposed for the first time in this study. As to PPDC method, we construct a new projection index, which successfully avoids the problem of parameter calibration in PPC method and makes the cluster results more objective and definite; besides that, PPDC method is also robust and easy to operate in practice. PPDC method is applied to study multi-factor evaluation of regional dangerous degree of debris flow, and the tests show that it is not only feasible, but also can overcome the short coming of current evaluating methods and achieve satisfying practical results. This application initiates a new approach for solution of similar problems, and furthermore also provides scientific foundation and technical support for ecological protection, disaster prevention and disaster relief.

Key words: projection pursuit; dynamic cluster; debris flow; dangerous degree evaluation