模糊聚类分析在水土流失分类中的应用

汪有科　　王培森

（中国科学院西北水土保持研究所）

模糊聚类分析是模糊数学的一个分支。它的主要功能是进行多指标的综合分类。由于它能分析自然界中的模糊现象，给人以清晰的答案，因而越来越受到人们的重视。目前已在很多领域中得到了应用，成为人们处理模糊问题的有力工具。水土流失分类是以防治为目的，由于影响水土流失的因素很多，单独采用侵蚀模数作为分级指标，虽然能看出水土流失的强弱，但不能反映造成各类侵蚀的主导因子是什么，所以满足不了制定水土保持规划的要求。要在分类中正确的反映众多的侵蚀因子信息，就需要采用数学方法，应用计算机进行多指标的综合分类。数学分类方法较多，模糊聚类分析是其中较好的方法之一。本文试图介绍模糊聚类分析的基本思想和运用步骤，并以水土流失分类为例来说明模糊聚类分析在水土流失分类中的应用。

一、模糊聚类分析的基本思想及运算步骤

要进行分类的对象称为样本。模糊聚类分析是首先计算样本的相似性尺度矩阵，然后采用基于模糊等价关系的聚类方法进行样本的聚类。

从集合论的角度来看，所谓一个分类，实际上是将集合G分成若干个子集g，每个子集叫做一类。G中任一元素都必须属于而且仅属于某一类。同一类的元素具有以下3个性质：

1、自反性：$r_{ii}=1$；2、对称性：$r_{ij}=r_{ji}$，其中$0\leqslant r_{ij}\leqslant 1$；3、传递性：$R\cdot R\subseteq R$。

为了对样本进行合理的分类，首先要将样本的种类性质数量化（如编码），这种数量化的性质称为样本的指标。如果某样本有m种指标，就可用m维空间的一个点来描述该样本；若有n个样本，其中每个样本有m个指标，可将样本列成表1的形式，表中x_{ij}表示第i个样本的第j个指标，第i个样本的向量表示为：$X_i=(x_{i1}, x_{i2}, x_{i3}, \cdots\cdots x_{im})$。

表1　　样本指标及其向量表示

样本	指标		
	y_1 y_2 …… y_m	向量表示	平均值
x_1	x_{11} x_{12} …… x_{1m}	X_1	$\bar{X}_1$
x_2	x_{21} x_{22} …… x_{2m}	X_2	$\bar{X}_2$
⋮	⋮ ⋮ ⋮	⋮	⋮
x_n	x_{n1} x_{n2} …… x_{nm}	X_n	$\bar{X}_n$
向量表示	y_1 y_2 …… y_m		
平均值	$\bar{y}_1$ $\bar{y}_2$ …… $\bar{y}_m$		

在选定了样本的指标之后，进行模糊聚类分析的方法大致分以下三步：

第一步，把代表点的样本指标的数据标准化，也称为正规化。计算公式如下：

$$x=\frac{x'-\overline{x'}}{c}$$

式中：x'——原始数据；

$\overline{x'}$——原始数据的平均值；

c——原始数据的标准差。

若把标准化数据压缩到〔0，1〕闭区间，可用极值标准化公式：

$$x=\frac{x'-x'_{min}}{x'_{max}-x'_{min}}$$

当$x'=x'_{max}$时，则$x=1$。

当$x'=x'_{min}$时，则$x=0$。

第二步，叫做标定，即算出衡量被分类对象间相似程度的统计量r_{ij}（$i=1,2,3\cdots\cdots n$；$j=1,2,3\cdots\cdots n$；n为被分类对象的个数），从而确定论域U上的相似关系R。

$$R=\begin{pmatrix} r_{11} & r_{12}\cdots\cdots r_{1n} \\ r_{21} & r_{22}\cdots\cdots r_{2n} \\ \vdots & \vdots \quad\quad \vdots \\ r_{n1} & r_{n2}\cdots\cdots r_{nn} \end{pmatrix}$$

计算统计量r_{ij}的方法很多，我们可在实践中选用。

第三步，基于模糊等价关系上的聚类。采用标定方法构成的模糊矩阵，往往只满足自反性和对称性，而能否满足传递性则不易看出，需要计算$R\cdot R$（表示R和它自身的合成R^2），看其能否满足$R\cdot R\subseteq R$的关系，即将R自乘得$R\cdot R=R^2$；如不满足再自乘$R^2\cdot R^2=R^4$，如此继续下去，至某步出现：$R^{2K}=R^K$

至此，则R^K便是一个模糊等价关系。然后根据水土保持工作的需要，选择适当的λ水平，即可得到相应的分类。

二、分类举例

我们采用杨艳生等（1982，《土壤学报》第19卷第1期）选定的主要侵蚀因子作为水土流失分类的指标。这些指标都可看作有序多态指标，因此编码取连续排列的非负整数，如表2所示。

表2　　侵蚀因子强度分级及编码

侵蚀因子	强度分级				
土壤可蚀性	山地土壤	黑垆土	暗黄土	绵黄土	
地面坡度	0—5°	5—10°	10—15°	15—25°	>25°
土壤利用状况	林地或川地	草地或台地	疏草地或缓坡农地	光坡地或耕地	
土壤利用干扰强度	轻	中	强	很强	
编码	1	2	3	4	5

假设我们在水土流失地区取得了17个样本资料，如表3所示。

表3　　样本指标统计

样本号	指标			
	土壤可蚀性	地面坡度	土壤利用状况	土壤利用干扰强度
1	1	3	2	1
2	2	1	4	1
3	1	4	3	2
4	2	3	2	4
5	2	4	2	3
6	2	4	2	3
7	1	5	1	1
8	1	4	1	2
9	1	4	1	3
10	1	5	4	3
11	1	4	3	4
12	1	4	3	4
13	1	5	3	4
14	2	1	1	1
15	3	1	1	1
16	2	1	1	1
17	2	1	1	1

计算样本间的相似程度的统计量（r_{ij}），选用夹角余弦法：

$$r_{ij}=\frac{\sum_{k=1}^{m} x_{ik}\cdot x_{jk}}{\sqrt{\left(\sum_{x=1}^{m} x_{ik}^{2}\right)\left(\sum_{k=1}^{m} x_{jk}^{2}\right)}}$$

经上述运算（全部运用由微型计算机——Apple Ⅱ完成），当$R^{32}=R^{16}$时，得模糊等价关系矩阵（矩阵表略），然后选不同的λ值可得表4分类结果，据此作出样本的模糊聚类树系图1。

由表4和图1可看出，随λ值（$0\leqslant\lambda<1$）的增大，相应的分类也越细。水土流失分类是以治理为目的，分类太细就失去了分类的意义，分类太粗又达不到因地制宜的效果。因此，选择适当的分类结果是十分重要的。如何选择分类结果？我们认为在样本及指标少的情况下，一般可凭工作经验直接从树系图上选择分类结果；当遇到样本和指标繁多，分类结果难选择时，建议用徐克学曾用的选择结合次数出现飞跃前后的中值(1983,《植物分类学报》第21卷第1期)。如在此例中有两次飞跃（图2），第一次飞跃前后的中值为0.8275，第二次飞跃前后的中值为0.9075，选择中值0.8275时，分类结果太粗，应选择中值为0.9075时的分类结果。这样，此例可分为三类。这个分类结果对比表2和表3可知，样本（14、15、16、17）主要是土壤本身性质影响它的侵蚀状况，土壤侵蚀不明显，可作为不治理地段。样本（2）单独为一类，（下转第40页）

程不完善，除南山坡水下来后，因排水不畅进入市区内老河道外，市区内因缺少排水设施，也加重了内涝程度。“85·8·16”大水过后，一些低洼地段的居民区因一时积水排不出，长期受淹。

第三、在防汛抢险工作中经验不足，造成了一些失误。伊春市已有十多年没有发生大水，参加这次防汛指挥的领导同志又多年没有经历过抗洪斗争的实践，难免出现一些工作的失误。比如，在汛前检查时，忽略了西部薄弱环节的检查；在洪水进入市区时，没有随时派专人监视水情的变化和汛情的发展；在防洪抢险上，缺乏适应不同情况的处理方案，当发生“85·8·16”大水时就显得措手不及。上述这些情况如果市里领导具有一定的防汛抢险经验，及时采取果断措施，将会减少一些损失。

第四、平时对河道和防洪工程管理注意不够。这次现场调查中发现不少漏洞。在河道管理上省里三令五申不准在河道内设障，省人大颁发的《黑龙江省河道管理条例》对这个问题专门作了规定，已经立了法，而伊春市却在伊春河的堤坝上修建了水上俱乐部，在行洪断面内堆筑了“吕梁山”，以其为中心还打算建立一个水上公园。在河道内设置了这些碍障物，严重的影响了行洪。在城市堤防工程管理上，存在各自为政的现象：汤旺河、伊春河堤防归市政部门管理，乌马河堤防和红光农场堤归农场管理；应该归市政管理的堤防又交给了伊春区代管，没有设置专管机构，人员也不适应，给堤防管理工作带来了很大被动。

四、两点建议

鉴于“85·8·16”大水反映出的问题和应吸取的教训，建议：

1、伊春市委、市政府应该很好地组织干部、群众和有关部门认真总结这次大水的教训，加强防洪除涝工程的基础建设，提高堤防工程的防洪能力。对暴露出来的问题要切实研究好挽救措施，防止发生类似大水再受其害。从干部、群众，特别是市委、市政府领导同志，对伊春市防涝问题应该有充分的认识，决不能掉以轻心，麻痹大意。

2、伊春市做为山区城市，要切实注意做好水土保持工作。对市区附近山区分水岭以内的林木植被一定要严加保护，坚决做到封山育林，保持水土，尽量减轻发生山洪造成的危害。

（上接第95页）

其主要是“土地利用状况”差所造成，在治理上应加强林草措施。样本（1、3、4、5、6、7、8、9、10、11、12、13）为一类，这类水土流失严重，它主要是由于坡度大，植被稀少，人为破坏严重而造成，所以应以综合治理为主，如修水平阶，造林种草，加强保护措施等。

三、小　结

1、模糊聚类分析能综合评判我们提供的各种水土流失因子，从中计算样本之间的相似关系，以样本之间的相似程度进行分类，可使水土流失分类有较严格的数值依据。

2、模糊聚类分析能将侵蚀因子相近的样本归为一类，这在制定水土保持规划方面较单一采用侵蚀模数分类更具有实用价值。

3、本文中的举例可作为应用模糊聚类进行水土流失分类时参考。在实际应用中，样本指标还可根据情况进行增减或重新选定，如要进行全国水土流失分类或分区，应选定统一的样本指标，这样的分类结果无疑对国土整治是十分有益的。

4、λ值选定多大为宜，还可进行探讨。

表4　　不同λ水平下的分类结果

λ 值	样 本 分 类
0.999	(1), (2), (3), (4), (5, 6), (7), (8), (9), (10), (11, 12), (13), (14, 16, 17), (15)
0.997	(1), (2), (3, 10), (4), (5, 6), (7), (8), (9), (11, 12), (13), (14, 16, 17), (15)
0.990	(1), (2), (3, 10), (4), (5, 6), (7), (8). (9), (11, 12, 13), (14, 16, 17), (15)
0.985	(1, 3, 10), (2), (4), (5, 6), (7), (8), (9), (11, 12, 13), (14, 16, 17), (15)
0.982	(1, 3, 10), (2), (4), (5, 6), (7), (8, 9), (11, 12, 13), (14, 16, 17), (15)
0.980	(1, 3, 10, 11, 12, 13), (2), (4), (5, 6), (7), (8, 9), (14, 15, 16.17)
0.975	(1, 3, 5, 6, 10, 11, 12, 13), (2), (4), (7), (8, 9), (14, 15, 16, 17)
0.970	(1, 3, 5, 6, 8, 9, 10, 11, 12, 13), (2), (4), (7), (14, 15, 16, 17)
0.967	(1, 3, 4, 5. 6, 8, 9, 10, 11, 12, 13), (2), (7), (14, 15, 16, 17)
0.960	(1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13), (2), (14, 15, 16, 17)
0.855	(1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17), (2)
0.800	(1. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17)

注：表中每个括号为一类，括号内的数字为样本号

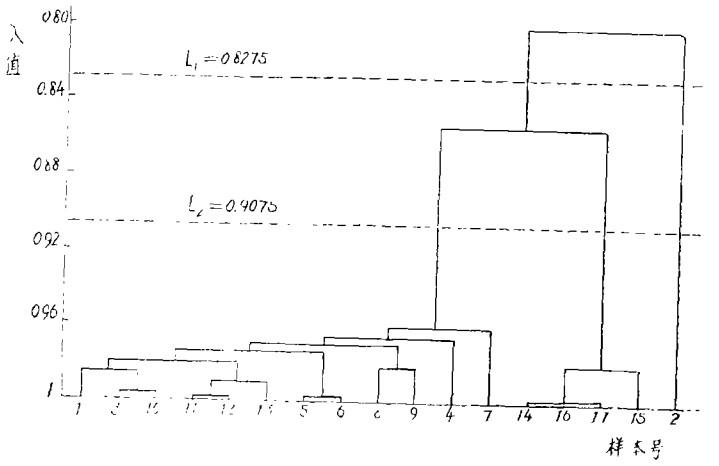

图1　样本模糊聚类树系图

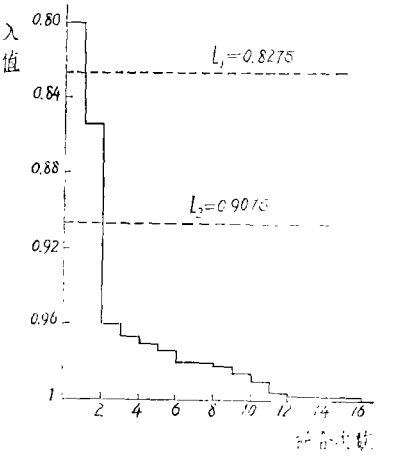

图2　样本模糊聚类的结合线图