

模糊ISODATA在草地植物群落分类上的应用*

高琼 郑慧莹

(中国科学院植物研究所, 北京 100044)

摘 要

模糊集合理论在元素与集合的关系中引入了连续变化的隶属度, 因而能适应植物群落作连续和明显间断分布同时存在的情况。本文运用以模糊集合理论为基础的 ISODATA 聚类分析方法对松嫩平原草地植物群落进行了数量分类, 结果说明模糊 ISODATA 聚类比传统的分类方法有更为客观、灵活的特点, 是一种较为理想的植被数量分类方法。

关键词 模糊集合; 植被; 生态; 数量分类

一、引 言

植物群落数量分析中所用的聚类法, 长期以来以经典的集合论为其理论基础, 聚类元素(如群落样方或植物种)被按其属性的相近程度分划成一些典型的类别。对非等级制聚类而言, 各类别之间互不相通, 任一样方(或植物种)对任一类型的关系只能是属于或者不属于, 而不存在任何中间状态。这种简单的两态关系, 数学上通常以“0”和“1”来表示, 并且只能近似地描述呈明显间断分布的植物群落。但实际上自然界里植物群落的分布既有间断又有连续的情况, 对于连续分布的植物群落, 传统的分类方法很难给出满意的结果。此外, 如果由于各种数值计算上的原因(如参数选择不当, 或计算精度不够, 以致积累误差淹没了分析对象的主要梯度等), 分类结果出现了不可解释之处, 一般只有两种解决问题的办法: 一是将聚类结果中个别不能解释的样方(或植物种)作为异常情况(Outlier)剔除, 二是调整某些参数, 重新分类。前者易受主观因素干扰, 而后者有赖于分析者的经验, 需要付出更多的时间。著名的 TWINSpan^[10] 聚类实质上也是一种经典的集合理论分类方法, 经常需要反复调整程序中的某些参数才能获得令人满意的结果。加之该程序本身不能作分析结果的图形显示。这样, 分析—显示—调整参数—再分析的反循环过程不可避免地需要程序之间的切换, 当数据量较大时, 以上过程将需要相当长的时间。

模糊集合论自从六十年代诞生以来, 在我国的各学科领域中得到了广泛的应用^[2-4]。究其根本原因, 在于它较为客观地处理了元素对集合(或样方对类型)的隶属关系, 模糊集合理论对于植物群落数据的无明显间断或连续、间断相杂的情况从理论上讲有着无与伦比的适应性。模糊集合是经典集合的推广, 所以它包含了经典集合, 可以用来处理植物群落的间断分布。另一方面, 它在元素与集合的关系中引入了隶属度的概念, 用来表示元素

本文于1990年6月收到, 1990年8月收到修改稿。

* 本文为中国科学院植被数量生态开放实验室支持的课题: “植被数量分析与不确定建模的研究”的一部分。

(样方或植物种)属于某一类型的程度,其值可以是“0”和“1”之间的任意实数。因此模糊集合理论又具有某些连续函数的特点,元素可以不同的隶属度同时属于两个或两个以上的在概念上互相排斥的类型。在模糊集合理论基础上发展的模糊 ISODATA 聚类法,是一非等级式的聚类分析方法,本文重点探讨该方法对于群落分类的适用性。

二、模糊 ISODATA 聚类法

模糊 ISODATA (Iterative Self Organizational Data Analyzing Technique A)的核心内容是根据反复叠代的原理,对给定的群落数据 X_{jk} , $j=1,2,\dots,n$; $k=1,2,\dots,m$; 和给定的分类类别数 N (其中 X_{jk} 表示第 j 个样方中第 k 种植物的观测量,如盖度等, n 是样方总数, m 是植物种数),构造一软分划矩阵 r_{ij} 和一聚类中心矩阵 V_{ik} , 使得:

$$R = \sum_{i=1}^N \sum_{j=1}^n r_{ij}^\alpha \sum_{k=1}^m (X_{jk} - V_{ik})^2 \quad (1)$$

为最小。式中 $\alpha > 1$, 称作收敛因子, r_{ij} 的取值范围是从 0 到 1, r_{ij}^α 可以看作是第 j 个样方对第 i 类型的隶属度。这样,当 r_{ij}^α 退化为经典集合中的“0”或“1”时, R 表示的实际上是所有样方在 m 维空间中到其各自所属类型的聚类中心的总平方和。使 R 为最小,实质上是使样方尽可能地靠近聚类中心^[4]。关于 r_{ij} 和 V_{ik} 的计算方法,可参考文献^[4]。

模糊 ISODATA 分析的关键在于类别数 N 的确定。这一参数通常是以主观,或以尝试—误差—再尝试的方法来取得。因此聚类结果中不可避免地掺入了一些主观、随意的成分,对数量分析的重要目的(即给客观世界以客观的评价)的实现有一定的不良影响。作者以计算理论中的最小码原理解决了客观地确定类别数的问题^[9]。基于这一思想的模糊 ISODATA 分析已被实施于中国科学院植物研究所植被数量生态学开放研究实验室发展的数量分析软件系列中。该软件集数值计算和图形显示于一体,以窗口、菜单为基本操作方式,分析、调整简便,无须进行程序间的切换。其使用性能达到国内外同类软件的先进水平。

三、松嫩平原草地群落的模糊 ISODATA 聚类

我们以松嫩平原草地植物群落为实例进行了模糊 ISODATA 聚类分析。松嫩平原地处温带,具有温带半湿润大陆性气候的特点,春季干旱多风,夏季暖热多雨,秋季短暂早霜,冬季寒冷少雪^[6]。地带性植被为草甸草原。平原内河流纵横,水量充沛,地势低洼(海拔在 150—200m 之间)^[6],周围数百条无尾河流向平原中部漫散,每年积盐约达 15 万吨^[7]。平原中分布有大大小小的数以千计的盐湖(水泡子),以碱湖为中心,植被分布呈同心园向外伸展,近碱湖周边处主要生长着一些盐生植物如碱蓬(*Suaeda glauca*)、角碱蓬(*S. carniculata*)、碱茅(*Paccinellia tenuiflora*)等,向外地势渐高,依次生长有羊草(*Aneurolepidium chinense*)及羊草-杂类草群落,贝加尔针茅(*Stipa baicalensis*)群落和干旱沙岗上的大针茅(*Stipa grandis*)和榆树(*Ulmus pumila*)疏林群落等。

我们以随机取样方法,对松嫩平原的草地植被进行了全面调查。群落样方分三种不同面积:榆树疏林为 $10 \times 10 \text{m}^2$,灌丛为 $2 \times 2 \text{m}^2$,草类为 $1 \times 1 \text{m}^2$,调查所用的植物特征值是常用的 Braun-Blanquet 的多盖度级^[8]。我们从全部 731 个样方中随机抽取 65 个样

方,共计 138 个植物种,作为实例进行模糊 ISODATA 分析。

本例共有 65 个样方,138 个种,数据矩阵为 65 行,138 列。因为植物种数超过 30 的样方极少,所以矩阵的绝大部分元素为 0。以这样的数据矩阵直接进行聚类分析,既浪费计算机储存空间,又影响计算的速度和精度。所以数据处理的第一步是采用 CA (Correspondence Analysis),即对应分析对数据进行降维排序处理。取 CA 分析的前五轴座标为新的属性指标,这样样方的属性变量就从 138 下降至 5,数据矩阵的元素个数也减至 65×5 ,大大地减少了下一步聚类分析的工作量。当然,这种降维处理损失了一部份信息,但保留了原始数据信息中的大部份。一般来说,保留的大部份是有规律的变化,而摒弃的大多是随机的、无规律的变化,从某种意义上讲,损失一部份无规律的随机变化信息不仅无损于分析结果,反而有利于抓住群落变化的本质机理。

第二步是按文献[9]中的原理作确定类别数的模糊 ISODATA 聚类分析,定出类别数为 4,即根据最小码原理,本例数据中的 65 个样方以分 4 类为妥。如前所述,分析结果中 r_{ij} 表示的是样方 j 对群落类型 i 的隶属度(本例取 $\alpha = 2$), V_{ik} 表示类型 i 的中心在第 k 轴上的投影。其中 $i = 1, 2, 3, 4, j = 1, 2, \dots, 65, k = 1, 2, 3, 4, 5$ 。

上述分析,均用中国科学院植被数量生态开放实验室的植被数量分析软件 FOIS 完成。

为了将聚类结果作一简单、明瞭的表示,我们将各样方按下列原则最终分别归于 4 个不同的类型:

(一) 在样方对全部 4 个类型的 4 个隶属度中,若最大的隶属度值大于 0.7,则样方应归于最大隶属度所对应的类型,此即为所谓最大隶属度原则。

(二) 若样方的最大隶属度小于或等于 0.7,则样方的最终归类要根据实际情况而定。理由是最大隶属度的微弱优势可能反映了实际差异,但也可能是由于计算误差或由于前一步中降维分析中的信息损失所致。

表 1 中列出的是 CA 分析的结果,即各样方在前五轴上的座标值。对样方所作的聚类结果列在表 2。4 个类别分别代表了 4 个不同生境,潮湿盐碱地,湿润,肥沃的立地,排水良好且较肥沃的立地以及干旱而肥力较低的立地上的植物群落类型,即碱蓬-碱茅群落,羊草群落,贝加尔针茅群落和榆树-大针茅群落。这与已有的研究结果相一致^[5,81]。实际上,这是该地区特征明显,有显著差异和间断的四个群落类型。这一结果,从客观上说明了以最小码原理作为模糊 ISODATA 聚类的类别数的确定是较为合理的。

应该指出的是最大隶属度在 0.1—0.5 之间的样方,如按最大隶属度原则归类,则 25, 18, 19, 20, 21, 22 号样方都得不到可解释的分类。因此,这几个样方是按其实际生境-植物关系,即按归类原则(二)来归类的。

以上根据最小码原理来确定模糊聚类的类别数而得出的分类是一种非常初级的分类。为了对该地区的群落作进一步研究,我们还作了类别数为 10 的模糊 ISODATA 聚类,其结果列于表 3,类型以小写字母表示,其与前四类分类的隶属关系亦显示在表中。说明不同类别数的聚类结果有可能存在一定的等级关系。

表 1 65 个样方的 CA 分析结果

Table 1 Result of CA for the 65 samples

轴 样方 Axis Samples	轴 Axis				
	1	2	8	4	5
2	-3.457	0.517	0.506	-8.355	0.506
4	-3.636	-0.521	0.583	-8.199	-4.880
6	-3.396	0.454	0.625	-7.138	11.228
8	-3.078	0.128	1.379	0.070	9.416
10	-3.391	0.510	0.463	-11.279	14.211
12	-3.531	0.520	0.499	-10.068	0.071
14	-3.508	0.498	0.463	-10.245	-0.308
16	-3.214	0.540	0.280	-13.854	10.925
18	29.572	34.277	-6.870	-13.729	10.862
20	31.012	39.021	-0.473	-13.021	-0.473
22	17.883	-23.588	-1.302	1.979	-1.302
24	30.732	-21.080	-0.307	-0.107	-0.307
26	31.059	-21.023	0.555	-0.078	-0.078
28	-3.955	0.523	6.199	11.507	1.051
30	-3.927	0.539	-1.042	9.565	0.981
32	-3.906	0.529	-2.779	7.312	0.940
34	-3.938	0.527	1.289	10.217	0.987
36	-3.971	0.534	5.407	16.427	1.053
38	-3.823	0.887	3.275	14.404	1.040
40	-3.901	0.515	3.627	14.418	1.047
42	-3.987	0.535	5.943	18.629	1.086
44	-3.885	0.540	-2.805	-0.645	0.739
46	-3.885	0.495	-5.319	6.069	0.963
48	-3.904	0.520	5.490	5.980	1.038
50	-3.873	0.547	3.570	14.871	1.053
52	-3.873	0.524	-11.490	1.631	0.873
54	-3.899	0.574	-12.226	1.388	0.876
56	-3.899	0.544	-12.380	1.061	0.851
58	-3.894	0.546	-12.218	-2.294	0.794
60	-3.884	0.539	-12.194	-2.139	0.799
62	-3.857	0.534	-15.159	-6.892	0.720
64	-3.849	0.528	-24.604	-10.247	0.748
64	-3.849	0.528	-15.082	-7.168	0.728
1	-3.234	0.519	0.284	-13.850	10.925
3	-2.942	-0.071	-0.505	-19.902	23.136
5	-3.014	0.287	0.065	-13.621	10.780
7	-3.522	0.463	0.763	-2.264	6.802
9	-3.388	0.441	0.688	-6.554	10.581
11	-3.268	0.512	0.582	-5.269	11.559
13	-3.202	0.505	0.288	8.018	8.018
15	-2.822	0.483	0.234	-13.812	10.884
17	31.001	0.658	-0.005	-13.729	10.862
19	17.909	39.060	-7.471	0.816	-0.473
21	30.673	-23.618	-34.802	1.979	-1.302
23	16.082	-21.130	31.752	-0.107	-0.307
25	-3.833	0.484	-30.841	-0.078	-0.078
27	-3.938	0.536	1.051	11.507	1.051
29	-3.932	0.538	0.981	9.565	0.981
31	-3.918	0.527	0.940	7.312	0.940
33	-3.968	0.526	0.987	10.217	0.987
35	-3.974	0.533	1.053	16.427	1.053
37	-3.966	0.529	1.040	14.404	1.040
39	-4.001	0.535	1.047	14.418	1.047
41	-3.914	0.522	1.086	18.629	1.086
43	-3.770	0.525	0.739	-0.645	0.739
45	-3.703	0.533	0.963	6.069	0.963
47	-3.912	0.549	1.038	5.980	1.038
49	-3.897	0.539	1.053	14.871	1.053
51	-3.865	0.540	0.873	1.631	0.873
53	-3.889	0.540	0.876	1.388	0.876
55	-3.835	0.530	0.851	1.061	0.851
57	-3.914	0.544	0.794	-2.294	0.794
59	-3.835	0.540	0.799	-2.139	0.799
61	-3.843	0.536	0.720	-6.892	0.720
63	-3.843	0.536	0.748	-10.247	0.748
65	-3.843	0.536	0.728	-7.168	0.728

表2 类数为4的模糊ISODATA聚类结果
Table 2 Fuzzy isodata dividing samples into 4 classes

类型 Class	隶属度 Membership degree			生 境 Habitat	优 势 种 Dominant species
	0.7—1.0	0.5—0.7	0.1—0.5		
A	23, 24, 26		20, 21, 22, 18, 19, 25	碱湖周边 Around salty lakes	角碱蓬、碱蓬等 <i>S. carniculata</i> <i>S. glauca</i> , etc.
B	1, 2, 5, 6, 9 10, 11, 13, 15, 16 17	3, 7, 8, 12, 14	4	肥沃, 湿润生境, A外围 Fertile, wet, outside of class A	羊 草 等 <i>A. neurolepidium</i> <i>chinese</i> , etc
C	27, 28, 29, 32, 33 34, 35, 36, 37, 38 39, 40, 41, 42, 46 47, 48, 49	30, 31, 43	45	肥沃, 排水较好, B以外 Fertile, well drained, outside of class B	贝加尔针茅等 <i>Stipa</i> <i>baicalensis</i> , etc.
D	44, 50, 51, 52, 53 54, 55, 56, 57, 58 59, 60, 61, 62, 64 65	63		沙岗, 旱生生境, C以外 Sand dune, dry, outside of class C	大针茅、榆树等 <i>Stipa grandis</i> , <i>Ulmus pumila</i> , etc.

表3 类数为10的模糊ISODATA聚类结果
Table 3 Fuzzy isodata dividing samples into 10 classes

类 型 Class	隶属度 Membership degree			优 势 种 Dominat species	前 四 类 Former 4 classes
	0.7—1.0	0.5—0.7	0.1—0.5		
a	23, 24, 26			角 碱 蓬 <i>S. carniculata</i>	A
b	21, 22, 25			碱 蓬 <i>S. glauca</i>	
c	18, 19, 20			碱 茅 <i>P. tenuiflora</i>	
d	1, 5, 15 16, 17	10 13	8	羊 草 <i>Aneurolepidium chinense</i>	B
e	6, 9, 11	2	7, 8	杂 类 草 Mixture	
f	12, 14	4		羊草、长梗葱 Mixture	
g	29, 30, 31 32, 46	33 45	47	贝加尔针茅 <i>Stipa baicalensis</i>	C
h	27, 28, 34 35, 36, 37 38, 39, 41 42, 48, 49		40 43	贝加尔针茅 + 杂草 <i>S. baicalensis</i> and others	
i	50, 51, 53 54, 55, 56 57, 58, 59 60	44	52	大 针 茅 <i>Stipa grandis</i>	
j	61, 62, 64 65	63		榆 树 <i>Ulmus pumila</i>	D

四、讨 论

本文将模糊 ISODATA 聚类法用于松嫩平原植物群落的分类,结果表明两种不同类别数的分类基本合理,说明模糊 ISODATA 分析不失为一种较好的植被数量分类的方法,其主要特点是:

(一) 较为客观地处理了样方对类别的隶属关系。模糊集合的连续变化的隶属度的概念使得介于两个或两个以上的典型植被类型之间的一些过渡性的样方和其所对应的植被类型有了较为客观,合理的表达方式。一个样方可以 0.5 的隶属度属于一种类型,而以 0.5 的隶属度属于另一种类型,这种情况,只能说明该样方介于两种类型的过渡区间。

(二) 较为灵活地处理了最终归类问题。前述的归类原则(二),即最大隶属度小于 0.7 的样方的归类具有一定的灵活性和可塑性,要求根据用户的实际经验来进行。即要求主观的,经验的参予。模糊 ISODATA 分析的软划分矩阵 r_{ij} ,即隶属度矩阵决定了样方、种、类型之间是一种可紧可松的“弹性”关系。若一样方以 0.9 的隶属度属于某一类型,则无论如何,该样方应归于这一类。另一方面,若一样方以 0.5 的隶属度属于类型 1,而对类型 2 的隶属度为 0.49,则只能说明样方基本上是介于这两个类型之间,至于两个隶属度之间的微小差异,可能是由于实际上的偏移,但也有可能是由于计算误差或降维分析中的信息损失所致。我们没有理由将该样方简单地归于类型 1,而应该视实际情况确定其归属。这就是模糊集合理论所独到的地方。在以经典集合论为基础的分类中,有时出现了不可解释的结果,例如某样方被不合理地分在某一类型中,我们没有任何理由和根据(数量上的)将其从误分的类型中取出来,归于其应该属于的类型。而模糊 ISODATA 聚类则为解决这一问题提供了数量上的依据,所以它比经典集合聚类方法更为灵活。这样一种灵活性在实际分析中是非常必要的,因为数量分类方法只能提出模型或假说,而不能检验假说,对于数量分类的结果的优劣,要以生态学的专业知识去进行解释和判断^[1]。

另外,以最小码原理作的总类别数选择也基本符合实际群落的结构特点。

参 考 文 献

- [1] 阳含熙、卢泽恩,1981: 植物生态学的数量分类方法,科学出版社,350—351。
- [2] 陈贻原,1984: 模糊数学,华中工学院出版社,武汉。
- [3] 汪培庄、韩立岩,1989: 应用模糊数学,北京经济学院出版社,北京。
- [4] 袁嘉祖、冯晋臣,1988: 模糊数学及其在林业中的应用,中国林业出版社,北京。
- [5] 郑慧莹、李建东,1990: 松嫩平原草原植被分类系统的探讨,植物生态学与地植物学学报,14(4) 297—304。
- [6] 李崇皓,1980: 松嫩平原草原植被的基本类型和特征,第三次东北草原学术会议论文集,47—58。
- [7] 李建东,1980: 东北平原植被类型及植被区划中的几个问题,第三次东北草原学术会议论文集,42—46。
- [8] 李建东、郑慧莹,1988: 松嫩平原南部植被与环境相关性的探讨,植物学报30(4) 420—429。
- [9] Gao, Q., 1989: The minimum description length principle as applied to the multivariate clustering analysis, coenoses, in press.
- [10] Gauch, H. Jr., 1982: Multivariate Analysis in Community Ecology, Cambridge University Press, Cambridge, UK.

AN APPLICATION OF FUZZY ISODATA TO THE CLASSIFICATION OF GRASSLAND COMMUNITY OF SONGLUN PLANE

Gao Qiong Zheng Hui-ying
(*Institute of Botany, Academia Sinica*)

Abstract

The concept of membership degree of the fuzzy set theory is well suited for the analysis of plant communities, where both abrupt and smoothly continuous distribution variations of plants and environmental quantities are often found. This work is an application of ISODATA based on fuzzy set theory to the classification problem of the grassland community in Songlun Plane, Northeast China. The result shows that the technique is objectively correct and is manipulative compared to the classical methods. In addition, the classification guided by the Minimum Description Length Principle proved that the principle is suitable for this type of classification problems.

Key words Fuzzy sets; Vegetation; Ecology; Quantitative classification