第 23 卷第 10 期
2003 年 10 月

生 态 学 报
ACTA ECOLOGICA SINICA

Vol. 23, No. 10
Oct., 2003

基于投影寻踪的天然草地分类模型

金菊良，张礼兵，潘金锋
（合肥工业大学土木建筑工程学院，安徽 合肥 230009）

摘要：提出了基于投影寻踪的天然草地分类模型(GQC-RAGAPP)，利用该模型可把各天然草地多维分类指标值综合成一维投影值，投影值越大表示该草地的环境综合质量越高，根据投影值的大小就可对草地样本集进行合理分类。建议用实码加速遗传算法进行 GQC-RAGAPP 的建模，简化了投影寻踪技术的实现过程，克服了目前投影寻踪技术计算过程复杂、编程实现困难的缺点。实例计算的结果说明，直接由样本数据驱动的 GQC-RAGAPP 模型用于天然草地分类简便可行，适用性和可操作性较强，可应用于各种非线性、非正态高维数据分类、评价等区域可持续发展研究中。

关键词：天然草地；分类；评价；可持续发展；投影寻踪；遗传算法

Classification model of natural grassland quality based on projection pursuit

JIN Ju-Liang, Zhang Li-Bing, Pan Jin-Feng (*College of Civil Engineering, Hefei University of Technology, Hefei* 230009, *China*). ***Acta Ecologica Sinica*, 2003, 23(10): 2184～2188.**

Abstract: A new method based on projection pursuit for natural grassland quality classification, called GQC-RAGAPP for short, is presented in this paper. The basic idea of GQC-RAGAPP model is to project high dimension indexes of natural grassland quality classification to projective values in only one dimension space, to describe classification structure by using a projective index function, to search the optimal projective directions according to the projective index function, and to analyze the classification structure characters of the high dimension data by the projective values. The problem to construct and optimize projective index function is suggested to be optimized by using real coded accelerating genetic algorithm developed by the authors, called RAGA for short.

The modeling of GQC-RAGAPP is the key in this paper, which includes four steps as following.

Step 1 is to standardize each index of natural grassland quality classification according to the minimum and the maximum in the classification sample set.

Step 2 is to construct the projective index function, which aim is to synthesize many dimensions sample data to one dimension $z(i)$ named projective value with projective direction a, where a is an unit length vector. The demands of the scattering characters of the projective values of $z(i)$ are that local projective dots should be denseness, that it is best to condense the dots to some groups, and that the dot groups should be dispersed. Based on the demands, a projective index function $Q(a)$ is designed as $Q(a)=S_zD_z$, where S_z is the standard variance of $z(i)$, and D_z is the local density.

基金项目：四川大学高速水力学国家重点实验室开放基金资助项目(0201)；安徽省优秀青年科技基金资助项目；安徽省自然科学基金资助项目(01045102)

收稿日期：2002-12-03；**修订日期**：2003-09-09

作者简介：金菊良(1966～)，男，江苏吴江人，博士，教授，主要从事水资源系统工程研究。

Foundation item: Open Fund of the State Key Lab. of High Speed Hydraulics in Sichuan University(No. 0201); Anhui Provincial Excellence Youth Science and Technology Foundation; Anhui Provincial Natural Science Foundation (No. 01045102)

Received date: 2002-12-03; **Accepted date**: 2003-09-09

Biography: JIN Ju-Liang, Ph. D., main research field: systems engineering of water resources.

Step 3 is to optimize the projective index function so that the optimal projective direction a^* can be estimated. As a general kind of optimization methods based on the mechanics of natural selection and natural genetics in biology, RAGA can be applied to deal with the optimal problem of maximizing the projective index function $Q(a)$ both easily and effectively.

Step 4 is to classify natural grassland samples according to the projective values $z^*(i)$ of the samples, which can be gained by substituting the optimal projective direction a^* according to Step 3. The series of $\{z^*(i)\}$ can be sorted orders from big to small. The bigger the value of $z^*(i)$ is, the better the natural grassland quality is, based on which the sample set can be classified.

The computation results of the case study can include four terms as following.

(1) In GQC-RAGAPP model, many classification indexes values of grassland samples can be synthesized projection value with only one dimension, which indicates environmental comprehensive quality of grassland samples.

(2) The grassland samples can be naturally classified according to the projection value of each grassland sample.

(3)By using real coding based accelerating genetic algorithm developed by the authors, the modeling of GQC-RAGAPP can be predigested the realized process of projection pursuit technique, and can overcome the shortcoming. of large computation amount and difficulty of computer programming in traditional projection pursuit methods.

(4) Applying GQC-RAGAPP model driven directly by samples data to classifying natural grassland samples is simple and feasible, its computed result is steady, and its applicability and maneuverability is good.

(5) GQC-RAGAPP model can be applied to classification and evaluation of nonlinear, non-normal distribution and high dimensional index data in regional sustainable development study.

Key words: natural grassland; classification; evaluation; sustainable development; projection pursuit; genetic algorithm

文章编号:1000-0933(2003)10-2184-05　**中图分类号**:S812.8　**文献标识码**:A

天然草地是由草地生态子系统、草地生产子系统和草地经济子系统等复合而成的复杂系统[1],不同类型的天然草地,它的草地稳定水平、缓冲水平、自净水平、抗逆水平和恢复水平等草地生态环境能力具有显著差异[1,2],从而它对草地生产子系统和草地经济子系统的承载能力也具有明显的不同,特别是它的载畜能力也将不同。中国的新疆、内蒙古、黑龙江等地区广泛分布着各种天然草地,对这些天然草地进行科学分类是这些地区可持续发展研究的一项基本内容,对这些天然草地的日常管理也具有重要的指导意义[1,3]。由于该问题涉及到许多不确定性因素,而各单因素指标的分类结果往往是不相容的,已有学者对此提出了分类方法,但这些方法对于天然草地高维、非线性、非正态指标数据处理效果并不理想,在模型应用的客观性、可操作性等方面尚存在一定的局限性[1,2]。近 20a 来国际统计界兴起的投影寻踪(Projection Pursuit,简称 PP)新技术[4,5],是一种直接由样本数据驱动的探索性数据分析方法,可用于具有任何结构和特征的高维数据的分类处理。但是常规 PP 技术的计算过程复杂、编程实现困难,至今仍限制了它的应用价值[4~6]。鉴于此,本文针对天然草地分类的具体情况,提出了基于实码加速遗传算法(Real coded Accelerating Genetic Algorithm,简称 RAGA)[7]的投影寻踪天然草地分类模型(Grassland Quality Classification model based on RAGA and Projection Pursuit,简称 GQC-RAGAPP 模型),并进行了应用研究。

1　基于投影寻踪的天然草地分类模型(GQC-RAGAPP 模型)

基于 PP 的分类模型的基本思想就是,把高维数据样本通过某种组合投影到低维子空间中,对于投影到的构形,采用投影指标函数(目标函数)来衡量投影暴露某种分类结构的可能性大小,寻找出使投影指标函数达到最优(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。其中,投影指标函数的构造及其优化问题是应用 PP 分类方法能否成功的关键。该问题很复杂,目前的 PP 实现方法的计算量相当大,在一定程度上限制了 PP 分类方法的深入研究和广泛应用[4~6]。为此,笔者建议用

RAGA 处理该问题，进而提出天然草地的投影寻踪分类模型(GQC-RAGAPP 模型)，其建模过程包括如下 4 个步骤：

步骤 1 建立分类指标体系，对各分类指标的样本数据进行标准化处理。根据所研究的天然草地的实际情况，从系统、应用和可操作的角度建立天然草地的复合型分类指标体系，一般包括草地生态指标子系统、草地生产指标子系统和草地经济指标子系统[1,3]。设研究地区天然草地分类指标的数据样本集为$\{x(j,i)|j=1\sim p,i=1\sim n\}$，其中 n、p 分别为草地样本的数目和分类指标的数目。为消除各分类指标的量纲效应，使建模具有通用性，需对$\{x(j,i)|j=1\sim p\}$进行标准化值处理[1]，其中对越大越优的正向指标的标准化值处理公式为：

$$y(j,i)=[x(j,i)-x_{\min}(j)]/[x_{\max}(j)-x_{\min}(j)] \tag{1}$$

对越小越优的逆向指标的标准化值处理公式为：

$$y(j,i)=[x_{\max}(j)-x(j,i)]/[x_{\max}(j)-x_{\min}(j)] \tag{2}$$

式中，$x_{\min}(j)$、$x_{\max}(j)$分别为样本数据集中第 j 个指标的最小值和最大值，$y(j,i)$为标准化后的数据样本值，$j=1\sim p,i=1\sim n$。

步骤 2 构造投影指标函数。PP 分类方法就是把 p 维数据$\{y(j,i)|j=1\sim p\}$综合成以 $a=(a(1),a(2),\cdots,a(p))$为投影方向的一维投影值 $z(i)$：

$$z(i)=\sum_{j=1}^{p}a(j)y(j,i) \tag{3}$$

然后根据 $z(i)\sim i$ 的一维散布图进行分类。式(3)中：$a(j)>0$，$\sum_{j=1}^{p}a(j)=1$。

在综合投影值时，要求投影值 $z(i)$的散布特征应为：局部投影点尽可能密集，最好凝聚成若干个点团，而在整体上投影点团之间尽可能散开。为此，投影指标函数可构造为[4]

$$Q(a)=S_z D_z \tag{4}$$

式中，S_z 为投影值 $z(i)$的标准差，D_z 为投影值 $z(i)$的局部密度，即：

$$S_z=\left[\sum_{i=1}^{n}(z(i)-Ez)^2/(n-1)\right]^{0.5} \tag{5}$$

$$D_z=\sum_{i=1}^{n}\sum_{j=1}^{n}(R-r_{ij})u(R-r_{ij}) \tag{6}$$

式中，Ez 为序列$\{z(i)\mid i=1\sim n\}$的均值；R 为求局部密度的窗口半径[4]，它的选取既要使包含在窗口内的投影点的平均个数不太少，避免滑动平均偏差太大，又不能使它随着 n 的增大而增加太快，R 的设置目前仍是经验性的，笔者的应用经验也表明 R 一般可取值为 $0.1\,S_z$；距离 $r_{ij}=|z(i)-z(j)|$；$u(t)$为单位阶跃函数，当 $t<0$ 时其函数值为 0，否则其函数值为 1。

步骤 3 优化投影指标函数。当给定天然草地分类指标样本数据时，投影指标函数 $Q(a)$只随投影方向 a 的变化而变化。不同的投影方向反映不同的数据结构特征，最佳投影方向可最大可能暴露高维样本数据的某分类特征结构。因此可通过求解投影指标函数最大化问题来估计最佳投影方向，即：

$$\max Q(a)=S_z D_z \tag{7}$$

$$\text{s.t.}\quad a(j)>0,\ \sum_{j=1}^{p}a(j)=1, \tag{8}$$

这是一个以$\{a(j)|j=1\sim p\}$为变量的非线性优化问题，常规方法处理很困难，这在很大程度上限制了投影寻踪技术的广泛应用[4~6]。模拟生物优胜劣汰规则与群体内部染色体信息交换机制的加速遗传算法(RAGA)，是一种通用的全局优化方法，用它来求解上述问题则较为简便。RAGA 的具体算法可参见文献[7]。

步骤 4 分类。把由步骤 3 求得的最佳投影方向 a^*代入式(3)后即得各天然草地的投影值 $z^*(i)$。$z^*(i)$值可反映各天然草地的综合质量特征，通过 $z^*(i)$值大小的比较，可对各天然草地进行分类。

2 实例研究

取中国新疆某地区的 31 个天然草地分类作为本文的实例[1],根据研究分析,该地区天然草地分类指标体系由天然草地植被覆盖度、可食风干牧草产量 2 个草地生态指标,草地牧草利用率、草地可利用面积系数 2 个草地生产指标,以及草群中优良牧草比率、羊单位需草地面积 2 个草地经济指标共 6 个指标组成,这些指标的样本数据见表 1。现利用 GQC-RAGAPP 模型对该样本集进行分类。表 1 中,羊单位需草地面积为逆向指标,按照式(2)进行标准化处理,其余指标为正向指标,按照式(1)进行标准化处理。

把标准化后的样本数据依次代入式(3)、式(5)、式(6)和式(4),即得此例的投影指标函数,再用 RAGA 解由式(7)和式(8)所确定的优化问题,得最大投影指标函数值为 0.390,最佳投影方向 $a^*=(0.2511, 0.5058, 0.1865, 0.0207, 0.0103, 0.0256)$。把 a^* 代入式(3)后即得各天然草地的投影值 $z^*(i)$,结果见表 1。天然草地的投影值 $z^*(i)$ 越大,表示该天然草地的综合质量越高,各 $z^*(i)$ 值的散布情况见图 1。为便于对比,表 1 和图 1 同时列出了文献[1]用灰色关联度分析的相应分类结果。

表 1　某地区各天然草地的分类指标样本数据及其投影值

Table 1　The indexes values of the natural grassland samples data and their projection values in a zone

草地序号(i) No.	植被覆盖度 (%)	可食风干牧草产量 (t/hm²)	牧草利用率 (%)	草群中优良牧草比率 (%)	草地可利用面积系数 (%)	羊单位需草地面积 (hm²/只)	投影值 $z^*(i)$	加权灰色关联度[1]
1	8.0	0.3615	50.00	1.80	86.0	3.64	0.074	0.3795
2	6.0	0.5115	48.00	0.00	82.0	2.67	0.076	0.3848
3	10.0	0.2340	47.50	32.80	90.0	5.91	0.062	0.3850
4	15.0	0.3930	45.00	49.60	95.0	3.71	0.097	0.4372
5	15.0	0.3435	42.00	29.80	85.0	4.55	0.070	0.3842
6	25.0	0.1695	45.00	100.00	98.0	8.60	0.103	0.5261
7	20.0	0.2970	43.00	71.20	95.0	5.14	0.096	0.4523
8	25.0	0.3089	43.00	75.90	93.0	4.97	0.113	0.4563
9	30.0	0.6089	42.00	87.10	90.0	2.57	0.160	0.5031
10	25.0	0.5190	57.98	60.20	92.0	2.18	0.185	0.4707
11	25.0	0.3362	40.00	68.20	85.0	4.87	0.100	0.4248
12	25.0	0.4425	40.18	78.77	94.0	3.69	0.121	0.4776
13	25.0	0.4800	45.00	73.90	100.0	3.06	0.144	0.5182
14	45.0	1.0145	45.00	86.80	94.0	1.44	0.259	0.5521
15	50.0	0.6015	64.61	95.30	98.0	1.68	0.301	0.6033
16	45.0	0.7809	43.00	96.00	92.0	1.95	0.230	0.5581
17	40.0	0.6135	47.30	80.70	92.0	2.27	0.209	0.5075
18	55.0	0.5700	65.00	100.00	98.0	1.77	0.315	0.6236
19	50.0	0.6495	64.50	100.00	95.0	1.57	0.305	0.6056
20	50.0	0.4230	46.00	80.10	94.0	3.38	0.214	0.5037
21	50.0	0.6884	47.00	100.00	95.0	2.03	0.251	0.5877
22	45.0	0.9203	56.16	81.10	95.0	1.45	0.285	0.5472
23	70.0	10.0517	65.00	84.20	100.0	0.91	0.406	0.6391
24	80.0	1.6035	64.73	26.70	100.0	0.67	0.476	0.6033
25	90.0	2.0586	62.04	82.40	98.0	0.34	0.551	0.6977
26	90.0	5.5350	98.00	87.70	98.0	0.12	0.996	0.9423
27	55.0	3.0450	55.00	0.00	98.0	0.39	0.500	0.5672
28	70.0	2.8740	55.00	0.00	98.0	0.28	0.529	0.5855
29	10.0	0.5595	50.00	0.00	90.0	2.37	0.104	0.4064
30	90.0	4.3635	95.00	15.15	99.0	0.16	0.862	0.7622
31	80.0	1.1565	95.00	5.50	98.0	0.62	0.526	0.6110

表 1 和图 1 说明:①草地序号为 13 与 14、22 与 23 的投影值 $z^*(i)$ 分别为 0.144 与 0.259、0.285 与 0.406,这些投影值的变化幅度较大,而草地序号为{1～13,29}、{14～22}、{23～28,30,31}这 3 组草地序号子集的投影值的变化幅度不大,可以分别单独聚成 3 类,这 3 类的投影值的变化范围分别为[0.0,0.2)、[0.2,0.4)和[0.4,1.0],这 3 类草地分别对应为温性荒漠类、温性草原类和草甸类,这就是该地区天然草地样本的分类结果,这与文献[1]表 1 的实地调查结果是一致的。②灰色关联度计算值的最大弱点是离散性

不强[8],分类结果趋于均化,因此不容易分类[1],而 GQC-RAGAPP 模型的分类结果离散性较强。③最佳投影方向 $a^*=(0.2511,0.5058,0.1865,0.0207,0.0103,0.0256)$表示各分类指标的权重,这与文献[1]表 3 所给出的权重基本一致。④与生态学界其它分类排序方法相比,GQC-RAGAPP 模型的显著特点是可以根据分类对象的具体特性和分类要求构造相应的投影指标函数,而其余建模步骤是相同的,因此 GQC-RAGAPP 模型可适用于其它区域可持续发展评价中[9]。

3 结论

天然草地分类的实质就是如何把研究地区各天然草地的多维分类指标综合成一维或两维指标,然后根据相近原则进行聚类。为此,本文提出了用投影寻踪分类模型(GQC-RAGAPP)进行天然草地分类的新方法。利用该模型可把各草地指标样本综合成一维投影值,投影值越大,表示该草地的综合质量越高,根据投影值的大小就可对各草地样本集进行合理分类。给出了 GQC-RAGAPP 建模的详细步骤,采用实码加速遗传算法简化了投影寻踪技术的实现过程,克服了目前投影寻踪技术计算复杂、编程实现困难的缺点。实例研究的结果说明,GQC-RAGAPP 模型用于天然草地分类客观简便、有效,模型解析能力和适用性较强,为投影寻踪技术在区域可持续发展研究中的广泛应用提供了新途径。

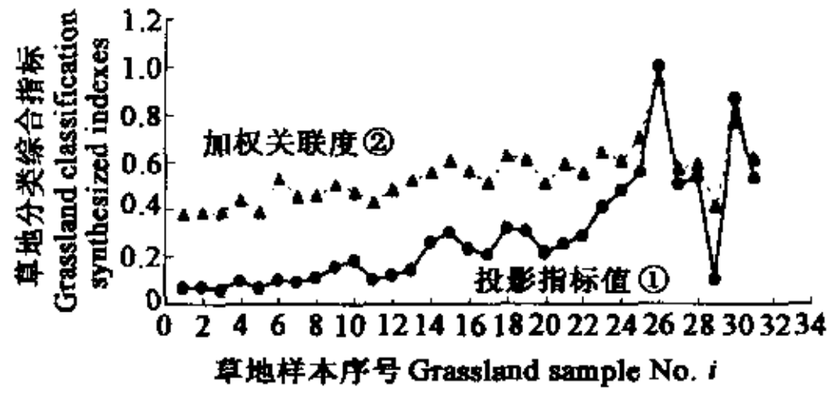

图 1 某地区天然草地分类指标样本数据的投影值 $z^*(i)$的散布图

Fig. 1 Scatter projection value pointers of the natural grassland samples data in a region

References:

[1] Wang Xin-Zhong, Lin Yi, Yu Lei. Data processing and grey correlation degree anysis in natural grassland types. *Journal of Theory and Application of Systems Engineering*, 2000, **20**(2):131～135,140.

[2] Sustainable development research group of Chinese Academy of Sciences. *Stratagem Report of Chinese Sustainable Development in* 2002. Beijing:Science Press,2002. 102～127.

[3] Niu Wenyuan, Mao Zhifeng. *System analysis of sustainable development theory*. Wuhan: Hubei Science and Technology Press,2002. 102～127.

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

[5] Li Zuoyong. Advances in projection pursuit technology and its applications. *Journal of Nature in China*, 1997, **19**(4):224～227.

[6] Li Zuoyong, Ding Jing, Zhang Xinli. Projection pursuit regression method of optimizing environmental monitoring site number. *Advance of Environment Science*,1999,**7**(6):127～130.

[7] Jin Juliang, Yang Xiaohua, Ding Jing. Real coding based accelerating genetic algorithm. *Journal of Sichuan University* (*Engineering Science Edition*) , 2000, **32**(4):20～24.

[8] Xiao Xinping. Theoretic study and review on the grey correlation degree quantitative models. *Journal of Theory and Application of Systems Engineering*, 1997,**17**(8):76～81.

[9] Min Qingwen, Li Wenhua. Assessment of regional sustainability and its application in Wulian county of Shandong Province. *Acta Ecologica Sinica*, 2002,**22**(1):1～9.

参考文献:

[1] 王新忠,林仪,于磊. 天然草地类型综合评价中的数据处理及灰色关联度分析. 系统工程理论与实践,2000,**20**(2):131～135,140.

[2] 中国科学院可持续发展研究组. 2002 年中国可持续发展战略报告. 科学出版社,2002. 102～127.

[3] 牛文元,毛志锋. 可持续发展理论的系统解析. 武汉:湖北科学技术出版社,1998. 67～163. 261～284.

[5] 李祚泳. 投影寻踪技术及其应用进展. 自然杂志. 1997,**19**(4):224～227.

[6] 李祚泳,丁晶,张欣莉. 环境监测优化布点的投影寻踪回归分析法. 环境科学进展,1999,**7**(6):127～130.

[7] 金菊良,杨晓华,丁晶. 基于实数编码的加速遗传算法. 四川大学学报(工程科学版),2000,**32**(4):20～24.

[8] 肖新平. 关于灰色关联度量化模型的理论研究和评论. 系统工程理论与实践,1997,**17**(8):76～81.

[9] 闵庆文,李文华. 区域可持续发展能力评价及其在山东五莲的应用. 生态学报,2002,**22**(1):1～9.