参数投影寻踪模型在坡耕地水土保持牧草优选中的应用

张 斌
(塔里木大学农业工程学院, 新疆 阿拉尔 843300)

摘 要: 针对坡耕地水土保持牧草选择问题, 采用高维降维技术——投影寻踪分类模型(PPC), 利用基于实数编码的加速遗传算法(RAGA)优化其投影方向, 将多维数据指标(样本评价指标)转换到低维子空间, 根据投影函数值的大小评价出样本的优劣, 从而做出决策, 最大限度避免了灰色关联法评判中的人为干扰, 取得了满意的效果, 为坡耕地水土保持牧草选择及其它评判决策问题提供一条新的方法与思路。

关键词: RAGA; PPC; 坡耕地; 水土保持
中图分类号: S157 文献标识码: A 文章编号: 1005-3409(2007)02-0299-03

Application of PPC Model to Grasses of Sloping Field for Soil and Water Conservation

ZHANG Bin
(*College of Agricultural Engineering, Tarim University, Alar, Xinjiang* 843300, *China*)

Abstract: Through applying PPC model base on RAGA in the grasses of sloping field, the author turns multi-dimension data into low dimension space. So the optimum projection direction can stand for the best influence to the collectivity. Thus, the value of projection function can evaluate each item good or not. The PPC model can avoid jamming of weight matrix in the method of grey relation, and obtain better result. A new method and thought is provided for readers who engaged in grasses of sloping field, and other relation study.

Key words: RAGA; PPC; sloping field; soil and water conservation

1 引 言

随着坡耕地可持续发展农业的发展和农业产业结构的调整, 种植牧草越来越得到人们的重视和利用。它不但可以防治坡地水土流失、提高坡耕地持续利用年限, 而且是农牧结合的纽带。牧草对于坡耕地水土保持起着重要的作用。为了牧草能有效的保持水土, 对牧草进行综合评价是进行牧草坡耕地水土保持的重要环节。目前常用的评价方法有模糊综合评判、层次分析法, 灰色关联等, 综合评判的实质是对高维数据(多个评价指标值)的处理, 即降低高维数据的维数, 通过专家给出的权重矩阵, 对应于每个指标的低维投影值, 在低维子空间实现其降维评价过程。但专家的权重矩阵是否属于各项指标(多维)在低维子空间的最佳投影, 还无法确定。为此, 我们引进一种新兴而又有效的降维技术——投影寻踪方法(Projection Pursuit, 简称 PP), 来实现其高维数据的降维过程, 并将新兴的适合于多维全局优化的算法——遗传算法与 PP 模型结合, 共同实现对坡耕地水土保持牧草的综合评价。

2 投影寻踪模型(PP)

投影寻踪是一种可用于高维数据分析, 既可作探索性分析, 又可作确定性分析的方法。Friedman 和 Tukey (1974)[1-4] 模仿有经验的数据分析工作者的做法, 提出了一种把整体上散布程度和局部的凝聚程度结合起来的新指标来作聚类和分类分析。

PPC(Projection Pursuit Classification Model, 简称 PPC 模型)的建模过程包括如下几步[5-10]:

步骤一: 样本评价指标集的归一化处理。设各指标值的样本集为$\{x^*(i,j)\mid i=1\sim n, j=1\sim p\}$, 其中 $x^*(i,j)$ 为第 i 个样本第 j 个指标值, n、p 分别为样本的个数(样本容量)和指标的数目。为消除各指标值的量纲和统一各指标值的变化范围, 可采用下式进行极值归一化处理:

对于越大越优的指标:

$$x(i,j)=\frac{x^*(i,j)-x_{\min}(j)}{x_{\max}(j)-x_{\min}(j)} \quad (1)$$

对于越小越优的指标:

$$x(i,j)=\frac{x_{\max}(j)-x^*(i,j)}{x_{\max}(j)-x_{\min}(j)} \quad (2)$$

式中: $x_{\max}(j)$、$x_{\min}(j)$——第 j 个指标值的最大值和最小值; $x(i,j)$——指标特征值归一化的序列。

步骤二: 构造投影指标函数 $Q(a)$。PP 方法就是把 p 维数据$\{x^*(i,j)\mid j=1-p\}$综合成以 $a=\{a(1), a(2), a(3), \cdots, a(p)\}$ 为投影方向的一维投影值 $z(i)$。

$$z(i)=\sum_{j=1}^{p}a(j)x(i,j) \quad (i=1\sim n) \quad (3)$$

然后根据$\{z(i)\mid i=1\sim n\}$的一维散布图进行分类。式(3)中 a 为单位长度向量。综合投影指标值时, 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。因此。投影指标函数可以表达成:

* 收稿日期: 2006-05-09
基金项目: 塔里木大学校长基金资助项目(TDZKSS05018)
作者简介: 张 斌(1976-), 女, 四川安岳人, 讲师, 主要从事水土保持研究。

$$Q(a) = S_z D_z \quad (4)$$

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

$$S_z = \sqrt{\frac{\sum_{i=1}^{n}[z(i) - E(z)]^2}{n-1}} \quad (5)$$

$$D_z = \sum_{i=1}^{n}\sum_{j=1}^{n}(R - r(i,j)) \cdot u(R - r(i,j)) \quad (6)$$

式中: $E(z)$ ——序列 $\{z(i) \mid i = 1 \sim n\}$ 的平均值; R ——局部密度的窗口半径, 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 的增大而增加太高, R 可以根据试验来确定, 一般可取值为 $0.1S_z$; $r(i,j)$ ——样本之间的距离, $r(i,j) = |z(i) - z(j)|$; $u(t)$ ——一单位阶跃函数, 当 $t \geqslant 0$ 时, 其值为 1, 当 $t < 0$ 时其函数值为 0。

步骤三: 优化投影指标函数。当各指标值的样本集给定时, 投影指标函数 $Q(a)$ 只随着投影方向 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, 最佳投影方向就是最大可能暴露高维数据某类特征结构的投影方向, 因此可以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:

最大化目标函数:

$$\text{Max: } Q(a) = S_z \cdot D_z \quad (7)$$

约束条件:

$$s.t.: \sum_{j=1}^{p} a^2(j) = 1 \quad (8)$$

这是一个以 $\{a(j) \mid j = 1 \sim p\}$ 为优化变量的复杂非线性优化问题, 用传统的优化方法处理较难。因此, 本文应用模拟生物优胜劣汰与群体内部染色体信息交换机制的基于实数编码的加速遗传算法(RAGA)来解决其高维全局寻优问题。

步骤四: 分类(优序排列)。把由步骤三求得的最佳投影方向 a^* 代入式(2)后可得各样本点的投影值 $z^*(i)$。将 $z^*(i)$ 与 $z^*(j)$ 进行比较, 二者越接近, 表示样本 i 与 j 越倾向于分为同一类。若按 $z^*(i)$ 值从大到小排序, 则可以将样本从优到劣进行排序。

3 基于实数编码的加速遗传算法(RAGA)

遗传算法由美国密执安大学的 Holland 教授提出的[11, 12], 是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法。主要包括选择(selection)、交叉(crossover)和变异(mutation)等操作。基于实数编码的加速遗传算法(Real coding based Accelerating Genetic Algorithm, 简称 RAGA)包括以下几个步骤:[5-10,13]

例如求解如下最优化问题: $\max: f(x)$, $s.t.: aj \leqslant xj \leqslant bj$

步骤 1: 在各个决策变量的取值变化区间随机生成 N 组均匀分布的随机变量(实数); 步骤 2: 计算目标函数值, 从大到小排列; 步骤 3: 计算基于序的评价函数(用 eval(v) 表示)。步骤 4: 进行选择操作, 产生新的种群; 步骤 5: 对步骤 4 产生的新种群进行交叉操作; 步骤 6: 对步骤 5 产生的新种群进行变异操作; 步骤 7: 进化迭代; 步骤 8: 上述 7 个步骤构成标准遗传算法(Standard Genetic Algorithm, 简称 SGA)。由于 SGA 不能保证全局收敛性, 在实际应用中常出现在远离全局最优点的地方 SGA 即停滞寻优工作。为此, 可以采用第一次、第二次进化迭代所产生的优秀个体的变量变化区间作为变量新的初始变化区间[13], 算法进入步骤 1, 重新运行 SGA, 形成加速运行, 则优秀个体区间将逐渐缩小, 与最优点的距离越来越近。直到最优个体的优化准则函数值小于某一设定值或算法运行达到预定加速次数, 结束整个算法运行。此时, 将当前群体中最佳个体指定为 RAGA 的结果。上述 8 个步骤构成基于实码的加速遗传算法(Real coding based Accelerating Genetic Algorithm, 简称 RAGA)。

将 PPC 模型中投影指标函数 $Q(a)$ 求最大作为目标函数, 各个指标的投影 $a(j)$ 作为优化变量, 运行 RAGA 上述 8 个步骤, 即可求得最佳投影方向 $a^*(j)$ 及相应的投影值 $z(i)$, 将 $z(i)$ 按其值大小进行比较, 从而求得评价结果。

4 应用实例

利用文献[14]的资料, 利用参数投影寻踪分类模型对坡耕地牧草进行综合评价。选定的指标有 12 个, 即总盖度(%)、草层高度(cm)、根入土深(cm)、越冬率(%)、抗病性(%)、青绿期、分蘖数(公顷/丛)、根长(cm)、根数(条/丛)、根干重(g/丛)、茎叶干重(g/丛)、适口性。具体见表 1。

表 1 各实验品种指标值

比较方案	总盖度/%	草层高度/cm	根入土深/cm	越冬率/%	抗病性/%	青绿期	分蘖数/(hm²·丛⁻¹)	根长/cm	根数/(条·丛⁻¹)	根干重/(g·丛⁻¹)	茎叶干重/(g·丛⁻¹)	适口性
迈洛克	82.2	28.5	26.4	59.6	90.0	258.0	232.5	28.9	66.7	1.3	3.2	6
翠 碧	81.7	26.6	25.1	83.5	95.0	262.0	244.5	27.1	81.5	1.2	3.1	6
雷得昆	50.8	19.5	27.6	65.2	82.0	224.0	199.5	30.9	56.4	0.8	1.9	10
IVORY	68.3	20.4	27.4	71.7	88.0	217.0	132	29.9	48.7	0.7	1.7	8
优 异	70.5	11.0	19.4	54.4	54.0	219.0	229.5	21.1	50.9	0.1	0.7	8
新哥来德	61.1	10.9	18.8	60.2	52.0	222.0	246	20.6	62.8	0.2	0.9	8
光脚丫	84.8	18.5	21.9	76.9	70.0	225.0	294	22.1	62.0	0.9	2.7	8
多福	83.8	27.5	24.4	56.9	75.0	216.0	244.5	26.9	56.3	1.2	3.3	8
弯叶画眉草	82.2	43.6	23.5	69.3	98.0	199.0	205.5	27.3	26.4	1.0	2.8	4

根据表 1 中数据建立投影寻踪综合评价模型。采用 MATLAB 6.5 编程处理, 选定父代初始种群规模为 $n = 400$, 交叉概率 $Pc = 0.80$, 变异概率 $Pm = 0.80$, 优秀个体数目选定为 20 个, $\alpha = 0.05$, 加速次数为 20, 得出最大投影指标值为: 1.135 1, 各个状态变量的最佳投影方向 $\alpha^* = (0.304\,4, 0.150\,3, 0.383\,7, 0.249\,7, 0.321\,3, 0.377\,5, 0.105\,1, 0.298\,1, 0.297\,4, 0.349\,3, 0.324\,1, 0.118\,5)$, 将代入式(2)后即得各个规划站点综合评价的投影值 $z^*(j) = (2.580\,3, 2.768\,4, 1.850\,9, 1.850\,9, 0.625\,8, 0.683\,3, 1.850\,9, 2.039\,9, 1.850\,8)$。将从大到小排列, 可得评级方案优劣顺序。

从评价结果可以看出, 综合评价投影值 $z^*(j)$ 为 2.768 4, 其次为 2.580 3, 即最优选择为翠碧, 其次是迈洛克。

5 结　论

(1) 投影寻踪模型直接采取各样本的原始数据进行分析, 信息量不会丢失。

(2) 投影寻踪模型将指标体系(高维数据) 投影到一维子空间上, 借助 RAGA 算法, 建立投影寻踪模型, 多次运算, 寻找最佳投影方向, 形成评价指标值, 按大小进行排序。模糊综合评判, 层次分析等方法专家赋权的人为干扰, 克服了传统方法的不足。

(3) 参数投影寻踪对坡耕地牧草选择评价得出最优选择为翠碧, 其次是迈洛克, 在实际生产中得到应用, 取得了良好的效果。此方法可在水土保持及相关领域进行应用。

参考文献:

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

[2] Friedman, J H Stuetzle W. Projection pursuit regression[J]. J. Amer. Statist. Assoc, 1981, (76): 817- 823.

[3] Friedman J H. Projection pursuit density estimation[J]. J. Amer. Statist. Assoc, 1984, (79): 599- 608.

[4] Diaconis, P, Freedman, D. Asymptotics of graphical projection pursuit[J]. The Annals of Statistics, 1984, 12: 793- 815.

[5] 付强. 农业水土资源系统分析与综合评价[M]. 北京: 中国水利水电出版社, 2005.

[6] 付强, 金菊良, 梁川. 基于实数加速遗传算法的投影寻踪分类模型在水稻灌溉制度优化中的应用[J]. 水利学报, 2002, (10): 39- 45.

[7] 付强, 王立坤. 基于加速遗传算法的投影寻踪模型在水质评价中的应用研究[J]. 地理科学, 2003, (3): 236- 239.

[8] F Qiang, X Y Gang, W Z Min. Application of Projection Pursuit Evaluation Model Based on Real-Coded Accelerating Genetic Algorithm in Evaluating Wetland Soil Quality Variations in the Sanjiang Plain, China[J]. Pedosphere, 2003, 22(2): 65- 68.

[9] F. Qiang, L T Gung. Applying PPE Model Based on RAGA TO Classify and Evaluate Soil Grade[J]. Chinese Geographical Science, 2002, 12(2): 136- 141.

[10] Qiang Fu, Hong Fu. Applying PPE Model Based on RAGA in the Investment Decision-Making of Water Saving Irrigation Project[J]. Nature and Science, 2003, 1(1): 57- 58.

[11] Holland, J H. Genetic algorithms and the optimal allocations of trials[J]. SIAM Journal of Computing, 1973, 2: 88- 105.

[12] Holland, J H. Genetic algorithms[J]. Scientific American, 1992, 4: 44- 50.

[13] 金菊良, 丁晶. 遗传算法及其在水科学中的应用[M]. 成都: 四川大学出版社, 2000. 42- 45.

[14] 字淑慧, 段青松, 吴伯志, 等. 应用灰色关联选择坡耕地水土保持牧草草种的研究[J]. 水土保持研究, 2006, 13(2): 61- 63.

(上接第 298 页)

表 1　气候变化和人类活动对伊洛河流域径流量的影响

起止年份	实测值 /mm	计算值 /mm	总减少量/mm	气候因素		人类因素	
				mm	%	mm	%
背景值	207.6	209.5					
1970- 1979	108.6	151.9	99.0	55.7	56.3	43.3	43.7
1980- 1989	154.5	180.2	53.1	27.4	51.6	25.7	48.4
1990- 1995	72.0	120.4	135.6	87.2	64.3	48.4	35.7
1970- 1995	117.8	155.5	89.8	52.1	58.0	37.7	42.0

由表 1 可以看出: (1) 自 20 世纪 70 年代以来, 受环境变化影响, 伊洛河流域实测径流量较背景值有明显的减少, 其中 1990 年以来减少量最大, 约为 135.6 mm。(2) 20 世纪 80 年代, 由于人类活动和气候变化引起的径流减少量均相对较低, 分别为 25.7 mm 和 27.4 mm。(3) 不同年代人类活动和气候变化对径流的相对影响程度不同, 但各年代由于气候变化引起的径流减少量均占径流减少总量的 50% 以上, 因此, 气候变化是伊洛河流域径流减少的主要因素, 就平均情况而言, 人类活动对径流的影响量只占径流减少总量的 42%。

5 结　语

基于对黄河中游伊洛河天然径流过程的模拟, 采用水文模拟途径分析了环境变化对该流域径流量的影响。结果表明, 澳大利亚水量平衡模型对伊洛河流域径流具有良好的模拟效果, 降水减少是该流域径流锐减的主要原因, 人类活动对径流的影响占径流减少总量的 42%。

伊洛河是黄河中游下段的最大支流, 也是黄河下游重要的水源和暴雨洪水来源区, 该流域径流洪水的变化直接关系到当地水资源的开发利用和下游的防洪安全。因此, 在进行该流域水资源变化规律研究的同时, 应进一步加强环境变化对伊洛河暴雨洪水的影响研究, 为黄河下游防洪提供科学依据。

参考文献:

[1] 陈江南, 王云璋, 徐建华. 黄土高原水土保持对水资源和泥沙影响评价方法研究[M]. 郑州: 人民黄河出版社, 2005.

[2] 水利部水文信息中心. 国家"九五"重中之重科技攻关专题(96- 908- 03- 02)"气候异常对中国水资源及水分循环影响评估模型研究"[R]. 2000.

[3] Boughton, W C. A simple model for estimating the water yield of ungauged catchments[J]. Inst. Engs. Australia, Civil Engg. Trans., 1984, 26(2): 83- 88.

[4] Boughton, W C. An Australian water balance model for semiarid watersheds[J]. Jour. Soil and Water Cons., 1995, 50(5): 454- 457.

[5] Nash J E, Sutcliffe J. River flow forecasting through conceptual models, Part 1, A discussion of principles[J]. Journal of Hydrology, 1970, (10): 282- 290.