参数投影寻踪模型在灌区运行状况综合评价中的应用

蒋国勇[1], 操华良[2]
(1. 绿洲工程造价有限服务公司, 新疆 阿拉尔 843300; 2. 绿洲水利工程建设监理站, 新疆 阿克苏 843000)

摘 要: 针对灌区运行状况, 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, 从而做出评价, 最大限度避免了灰色关联法评判中权重矩阵取值的人为干扰, 取得了满意的效果, 为灌区运行状况评价及其它评判问题提供一条新的方法与思路。

关键词: RAGA; PPC; 灌区; 综合评价
中图分类号: S274.3 文献标识码: A 文章编号: 1005-3409(2006)05-0087-03

Application of PPC Model for Comprehensive Evaluation of Irrigation Scheme

JIANG Guo-yong[1], CAO Hua-liang[2]
(1. *Limited Service Company of Fabrication Cost of the Oasis Project, Alar, Xinjiang* 843300, *China*;
2. *Oasis Water Conservancy Engineering Construction Surpervision Station, Aksu, Xinjiang* 843000, *China*)

Abstract: Through applying PPC model based on RAGA in the irrigation schemes, the authors turn 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. The PPC model can avoid jamming of weight matrix in the method of grey relation, and obtain better result. It provides a new method and thought for comprehensive evaluation of irrigation scheme and other relative study.

Key words: RAGA; PPC; irrigation area; comprehensive evaluateion

1 引 言

大型灌区在我国国民经济中具有重要的意义, 它是稳固我国经济发展与保障粮食安全的重要支撑。对大型灌区进行客观、合理的评价可为灌区发现差距, 改进运行管理水平提供方向, 为大型灌区进行续建配套改造的先后顺序及资金分配提供科学的依据。

以往大型灌区运行状况的评价多采用模糊综合评判、灰色关联投影法, 通过一系列评价指标, 来评价哪个方案为最优。在评判过程中, 都涉及权重矩阵。对于权重矩阵, 多由专家凭经验给出, 因此存在较大的主观性和人为干扰因素。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, 共同实现对大型灌区运行状况的综合评价。

2 投影寻踪模型

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:

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

对于越小越优的指标:

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

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

步骤二: 构造投影指标函数 $Q(\alpha)$。*PP* 方法就是把 p 维数据 $\{x^*(i,j) \mid j=1\sim p\}$ 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 $z(i)$

* 收稿日期: 2006-03-16
作者简介: 蒋国勇(1970-), 男, 毕业于塔里木农垦大学(现塔里木大学) 水利系农田水利专业, 主要从事水利工程预算和造价工作, 现任新疆阿拉尔绿洲工程造价有限服务公司工程师。

$$z(i)=\sum_{j=1}^{p}a(j)x(i,j) \qquad (i=1\sim n) \qquad (2)$$

然后根据$\{z(i)\mid i=1\sim n\}$的一维散布图进行分类。式(2)中a为单位长度向量。综合投影指标值时,要求投影值的散布特征应为: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:

$$Q(a)=S_zD_z \qquad (3)$$

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:

$$S_z=\sqrt{\frac{\sum_{i=1}^{n}(z(i)-E(z))^2}{n-1}} \qquad (4)$$

$$D_z=\sum_{i=1}^{n}\sum_{j=1}^{n}(R-r(i,j))\cdot u(R-r(i,j)) \qquad (5)$$

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 的增大而增加太高, 可以根据试验来确定,一般可取值为0.1S_z;$r(i,j)$表示样本之间的距离,$r(i,j)=\mid z(i)-z(j)\mid$;$u(i)$为一单位阶跃函数,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:

$$Max:Q(A)=S_z\cdot D_z \qquad (6)$$

约束条件:

$$s.t.:\sum_{j=1}^{p}a^2(j)=1 \qquad (7)$$

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

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:$\begin{matrix}Max:f(X)\\ s.t.:a_j\leqslant x_j\leqslant b_j\end{matrix}$

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 求最大作为目标函数,各个指标的投影 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 及相应的投影值,将 按其值大小进行比较,从而求得评价结果。

4 应用实例

本文以湖北省大型灌区运行状况综合评价为例,说明投影寻踪在灌区运行状况综合评价中的应用。待评价方案的资料取自参考文献[14],其评价指标见表1。选取了15个灌区作为待评价对象。考虑的评价指标有9个,在这9个评价指标中,除了单位灌溉面积灌溉用水量是越小越优型指标外,其余指标均为越大越优型。

表1 湖北省各灌区评价指标值(1998 年)

灌区编号	灌区名称	单位灌溉面积灌溉用水量/($m^3\cdot hm^{-2}$)	取系水利用系数	收入支出比/%	水费实收度/%	水价到位程度/%	人均管理灌溉面积/($hm^2\cdot 人^{-1}$)	单位灌溉用水量收益/(元$\cdot m^{-3}$)	单位控制面积产值/(元$\cdot hm^{-2}$)	单位灌溉用水量产值/(元$\cdot m^{-3}$)
X1	隔堤北	6643	0.45	85	61	43	202	0.020	13682	2.77
X2	蕲 水	6473	0.5	101	87	36	151	0.015	8555	3.55
X3	随 中	6552	0.521	99	65	63	106	0.032	11116	3.05
X4	徐家河	3495	0.422	85	65	51	291	0.021	9023	4.63
X5	郑家河	3641	0.45	65	46	46	178	0.010	12800	4.34
X6	环 东	3385	0.45	75	60	25	178	0.038	17859	7.57
X7	引 丹	4734	0.46	50	40	42	394	0.011	4785	2.23
X8	熊河水	4837	0.34	96	50	49	268	0.021	11158	3.59
X9	何王庙	7110	0.55	86	76	38	500	0.010	12308	2.81
X10	东西湖	5371	0.52	69	70	43	145	0.033	34168	6.68
X11	颜家台	4576	0.39	93	85	43	234	0.026	12651	3.03
X13	兴 隆	5382	0.47	64	60	46	200	0.032	10441	2.60
X13	泽 口	4511	0.47	95	90	37	674	0.027	18985	4.21
X14	天 门	8970	0.40	100	86	24	337	0.010	6512	1.10
X15	荆 门	4755	0.45	74	52	42	340	0.009	14064	4.09

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 P_c= 0.80, 变异概率 P_m= 0.80, 优秀个体数目选定为 20 个, α= 0.05, 加速次数为 20, 得出最大投影指标值为: 0.797 5, 各个状态变量的最佳投影方向 α= (0.259 1, 0.246 8, 0.062 5, 0.133 7, 0.085 9, 0.141 3, 0.514 5, 0.564 2, 0.489 7), 将 α^* 代入式(2)后即得各个规划站点综合评价的投影值 $Z^*(j)$= (0.894 4, 0.894 0, 1.214 9, 1.127 1, 0.894 2, 1.747 8, 0.569 8, 0.893 9, 0.893 8, 1.945 9, 1.108 0, 1.091 9, 1.545 8, 0.363 2, 0.888 7)。将 $Z^*(j)$ 从大到小排列, 可得灌区运行状况的优劣排序, 即东西湖> 环东> 泽口> 随中> 徐家河> 颜家台> 兴隆> 隔堤北> 郑家河> 蕲水> 熊河水> 何王庙> 荆门> 引丹> 天门, 与文献[14]灰色决策模型的评价结果相比, 参数投影寻踪的评价结果更为符合实际情况。

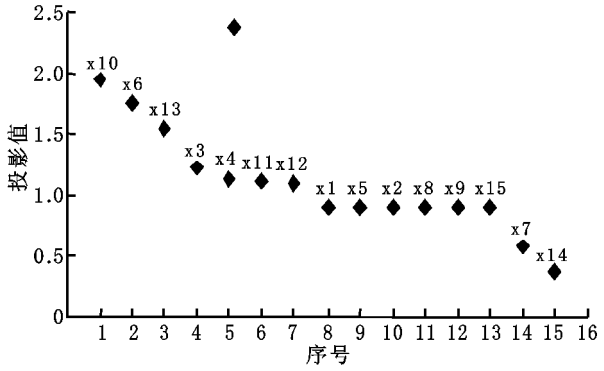

图 1　灌区运行状况综合评判投影值散布图(排序后)

根据最佳投影方向, 可以进一步分析各个评价指标对综合评价结果的影响程度。将 α^* 值从大到小排列, 得到各个指标的贡献大小顺序, 即: 单位控制面积产量, 单位灌溉用水量收益, 单位灌溉用水量产量, 单位灌溉面积灌溉用水量, 渠系水利用系数, 人均管理灌溉面积, 水费实收程度, 水价到位程度, 收入支出比。可见基本上反映了实际情况。样本优序关系及投影方向见图 1 与图 2。

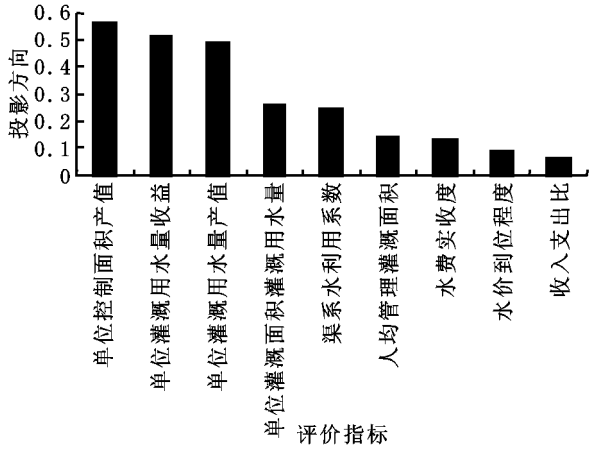

图 2　各个评价指标投影方向排序图

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。避免了灰色关联法、层次分析法等方法专家赋权的人为干扰, 克服了传统方法的不足。

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] 朱秀珍, 李远华, 崔远华, 等. 运用灰色关联法进行灌区运行状况综合评价[J]. 灌溉排水学报, 2004, 23(6): 44- 48.