投影寻踪模型用于水质等级评价的研究

刘建,刘丹　(西南交通大学环境科学与工程学院,四川成都 610031)

摘要　投影寻踪模型具有可处理高维、非正态数据分类问题的优点,适合于在水质等级评价中应用。该文在介绍该模型基本计算步骤的基础上,经建模分析,得出:各评价等级取值范围内样本个数不小于20 即可具有较好的稳定性,投影向量分量的取值范围为[0 ,1] 时更有利于提高计算速度,选用加速遗传算法寻优时间更短,使用图解法或直接比较法更为方便和快捷。这些探索能为该模型在今后的实际应用提供有益参考。

关键词　投影寻踪;水质评价;遗传算法;样本数量
中图分类号　X52　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)01 - 00261 - 02

Study on Projection Pursuit Model Applied in Water Quality Assessment

LIU Jian et al　(School of Environmental Science and Engineering , Southwest Jiactong University , Chengdu , Sichuan 610031)

Abstract　Projection pursuit model(PP) , being good at dealing with high dimension and non-normal data , was suitable for water quality assessment . This paper built water quality assessment model based on some basic introduction to PP . It made several conclusions as fdllows : PP model was stable when sample size was no less than 20 . It run much faster when the range of projection vector was constrained between 0 and 1 . Accelerated genetic algorithm should be recommended because of its unique advantages . Using scatter diagram of making comparison directly was more convenient than using regression model . All these reflections were expected to provide some useful references for the use of PP in the future .

Key words　Projection pursuit ; Water quality assessment ; Genetic algorithm; Number of samples

水质评价是环境管理和决策的重要组成部分,其实施过程就是依据已知的水质标准建立分类模型,然后根据该模型对计算样本的水质进行评价[1] 。由于实际水体各单项指标的评价结果常常是不相容的,故直接利用水质评价标准表进行水环境质量等级评判缺乏实用性[2] 。鉴于投影寻踪模型能在一定程度上解决高维、非正态数据分类问题[3] ,近年来,相关学者将该模型用于水质综合评价,并取得了较为满意的效果[1 - 8] 。尽管如此,就投影寻踪模型中各等级取值范围内随机样本生成数量、单位投影向量分量的取值范围、回归模型选择、优化方法选取等问题,各学者观点不尽统一。为此,笔者也进行了一些相关探索。

1　水质等级评价的投影寻踪模型简介

投影寻踪法[9] (Projection pursuit , PP) 最先是由 Friedman 于1974 年提出的一种新型数理统计分析方法。其基本思想是:利用计算机技术,将高维数据(尤其高维非正态数据) 通过某种组合,投影到低维(1～3 维) 子空间上。通过优化投影指标函数,寻找出能反映原高维数据结果或特征的投影向量,并在低维空间上对数据结构进行分析,以达到研究和分析高维数据的目的。建立水质等级评价的投影寻踪模型一般包括以下5 个步骤[2,4 - 6] 。

(1) 建立投影样本数据。根据水质评价标准产生用于水质评价的样本数据。它包括水质指标 $x^*(i,j)$ 及对应阶段的等级 $y(i)(i=1,\cdots,n,j=1,\cdots,p)$,其中,n ,p 分别为样品的个数和水质评价的指标数。为消除各指标的量纲效应和统一各指标的变化范围,需对 $x^*(i,j)$ 进行规范化处理,即:

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

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

式中,$x(i,j)(i=1,\cdots,n,j=1,\cdots,p)$ 为经过规范化处理后的评价指标集;$x^*(i,j)$ 为第 i 个样本的第 j 项指标值;$x_{min}(j)$ 、$x_{max}(j)$ 分别为第 j 项指标的最小、最大值。当评价指标属越小越优型时,应采用(1) 式进行处理,反之,则应用(2) 式进行处理。

(2) 计算投影值。$a=(a_1,a_2,\cdots,a_p)$ 设为投影方向的单位长度向量,则将 $x(i,j)(j=1,\cdots,p)$ 在该方向作线性投影得:

$$z(i)=\sum_{j=1}^{p}a_jx_{ij} \tag{3}$$

式中,$z(i)(i=1,\cdots,n)$ 为投影值。

(3) 建立投影目标函数。在综合投影值时,要求投影值 $z(i)$ 的散布特征为:局部投影点尽可能密集,最好聚成若干个团,而在整体上投影团之间尽可能散开。基于此,投影指标函数可构造为:

$$Q(a)=S_zD_z \tag{4}$$

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:

$$S_z=\sqrt{\sum_{i=1}^{n}[z(i)-E_z]^2/(n-1)} \tag{5}$$

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

式中,E_z 为各水样投影特征值 $z(i)(i=1,\cdots,n)$ 的均值;R 为局部密度窗口半径,其取值范围为 $\max(r_{ij})+\frac{m}{2}\leq R\leq 2m$;$r_{ij}=|z(i)-z(j)|(i=1,\cdots,n;j=1,\cdots,n)$;$u(t)$ 为单位跃升函数,当 $t<0$ 时,$u(t)=0$,否则 $u(t)=1$。

(4) 优化投影目标函数。当给定水样监测数据时,投影指标函数 $Q(a)$ 只随投影方向 $a=(a_1,a_2,\cdots,a_p)$ 的变化而变化,最佳投影方向可以最大限度地暴露出给定高维数据的某类特征结构。通过求解投影指标函数最大问题可估计最佳投影方向,即:

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

$$s.t.\sum_{i=1}^{m}a_j^2=1 \tag{8}$$

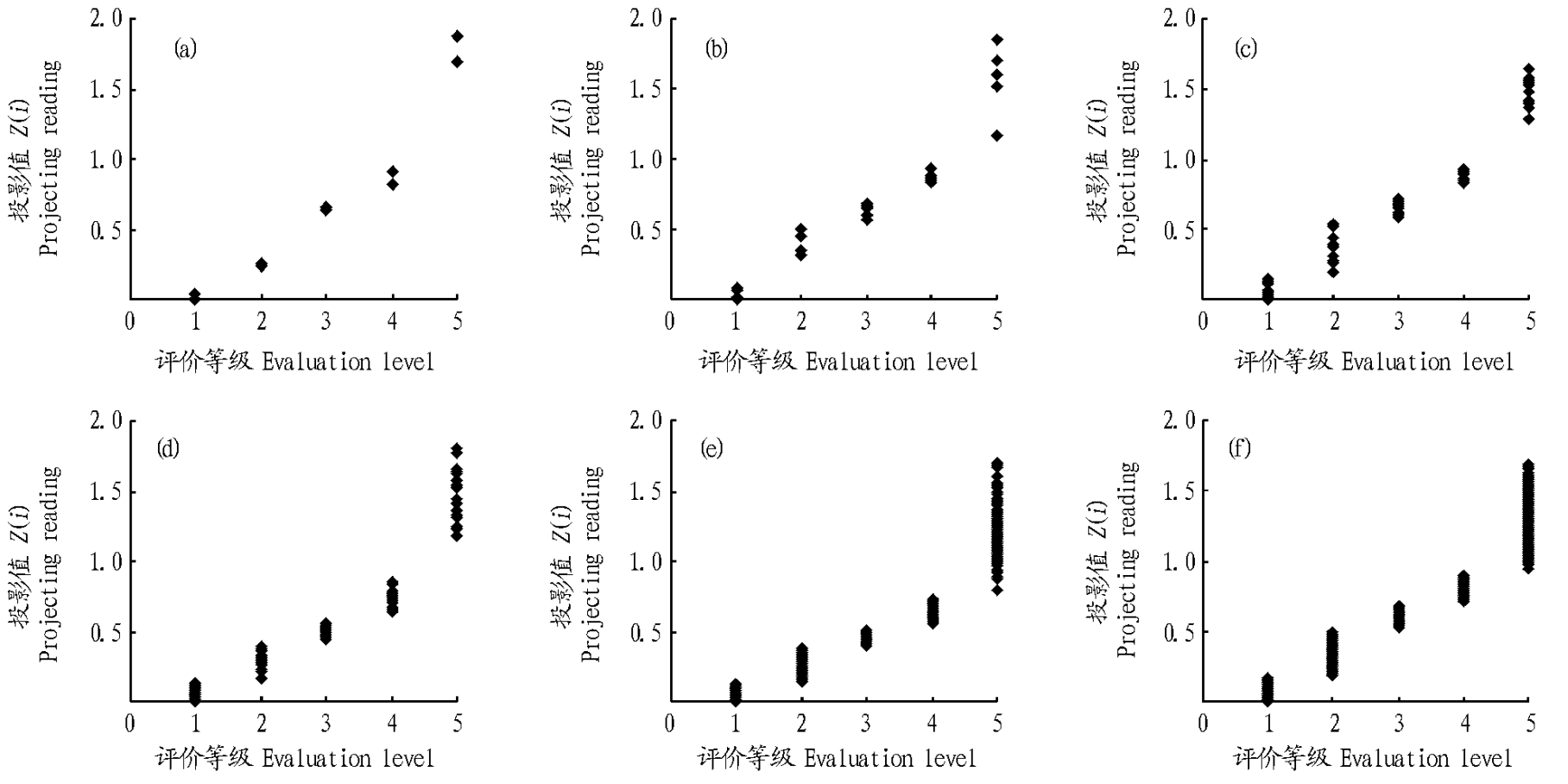
(5) 建立水质等级评价模型。作 $z(i)$ 和 $y(i)$ 的散点分布图,根据该图确定水质等级评价模型参数,并应用于其他

基金项目　四川省交通厅科学技术研究项目(05009) 。
作者简介　刘建(1982 -) ,男,四川威远人,博士研究生,从事工程环境控制技术方面的研究。
收稿日期　2008-10-27

样本的水质评价。

2 关于水质等级评价投影寻踪模型的思考

2.1 关于投影样本数量的思考 文献[3]和文献[8]仅采用评价标准自身构成的小样本数据(样本数分别为5和3)进行计算;文献[6]采用实测数据加虚拟数据(实测数据20组,虚拟数据5组)进行计算;文献[4]采用各等级取值范围内随机生成5个样本(共25个样本)进行计算;文献[1]采用各等级取值范围内随机生成200个样本(共1 000个样本)进行计算;文献[2]则采用在各等级取值范围内随机生成1 000个样本(共5 000个样本)进行计算。从上可以看出,不同学者构造的样本数量差异极大。理论上,样本数据量越大,计算结果越精确,但事实上,由于投影寻踪模型为非线性优化模型,当样本数据量过大后会导致后续优化十分缓慢甚至难以进行,故笔者不推荐为过于追求精度而采用庞大样本数据。为探讨合适的样本数量,该文以文献[4]使用的5等级4指标湖泊评价标准进行不同样本数量的稳定性分析,见图1。



注:(a)、(b)、(c)、(d)、(e)、(f)的样本数量分别为10、25、50、100、500、1 000。

Note: Sample quantities in (a), (b), (c), (d), (e) and (f) are 10, 25, 50, 100, 500 and 1 000 resp.

图1 不同样本数量下z(i)~y(i)散点图

Fig.1 Scatter diagram of z(i) - y(i) of different sample quantities

从图1可以看出,由于构造样本数据时随机因素的影响,并不是样本量越大,z(i)分别就越集中,当然样本数量也不宜过小,否则分辨能力会较差。通过对不同样本数量的z(*i*)~y(i)散点分布图分析,笔者认为各等级取值范围内随机产生样本数量不少于20个(即5个等级共100个)即可具有较高的分辨能力和计算精度。

2.2 关于投影向量分量取值范围的思考 多数学者计算出的投影向量分量的取值范围为[0,1],但文献[2]和文献[3]则为[-1,1],这是因为在进行投影样本规范化处理时只使用了(1)式,致使越大越优型指标的投影方向为负。笔者认为,在进行样本规范化处理时,(1)、(2)式应联合使用,其原因如下:①联合(1)式和(2)式处理后的数据与人们常规思维一致,指标的值越小,则水质越优;②这样处理后,可将单位投影向量各分量的取值范围限定为[0,1],有利于减少取值区间,加快计算速度和提高计算结果精度;③由于计算出的单位投影向量的各分量均为正,因此可直接用于评判各指标对评价等级贡献大小,避免他人误认为取负值的指标贡献小。

2.3 关于优化方法的思考 目前,用于优化投影目标函数的方法主要有遗传算法(GA)和序列二次规划算法(SQP),其中又以遗传算法使用最为普遍。LINGO是美国Chicago大学Linus Schrage教授在1980年开发的一种专门用于求解数学线性、非线性和整数优化的工具软件,它具有简洁直观、执行速度快和易于与Excel、数据库等其他软件交换数据的特点[10]。但笔者在试图使用该软件求解含1 000个样本数据的投影目标函数时却遇到了困难——系统资源消耗量大,执行速度十分缓慢。用Excel VBA编程检验,笔者认为基于实数编码的加速遗传算法[11]不失为求解该类问题的一种好方法,样本数量在10~1 000时,均只用不到1 min时间即可得到最优计算结果。

2.4 关于回归模型的思考 文献[4]采用了连续型逻辑斯谛曲线(Logistic curve)进行y(i)~z(i)回归分析,文献[2]则使用了分段线性插值模型(Linear interpolation)。根据该文图1计算结果,z(i)值呈十分明显的阶梯状分布,因此,笔者认为使用连续回归曲线难以刻画其本身的分散特点,致使最终计算误差可能较大;而分段模型显然更具合理性,不过笔者认为可省去插值步骤,直接将计算出的z^*与y(i)~z(i)散点图作比较或看其是否落在某等级i的上下限范围内[z^i_{sub}, z^i_{sup}]即可。

3 结语

投影寻踪模型在水质等级评价中的应用已得到众多学者研究,笔者就该模型中各等级取值范围内随机样本生成数量、投影向量分量的取值范围、优化方法选取、回归模型选择等问题阐述了自己的观点,认为:①各等级取值范围内随机

(下转第265页)

度已成为发展节水农业技术的必然选择。

3.6　加强节水农业推广中的组织管理　现代化的节水农业必须做到工程节水、农艺节水和管理节水的有机结合,才能实现真正意义上的节水农业[12]。在节水农业的组织管理中,要做到输水阶段的"定额供水,超额加价"的管理体制,利用经济杠杆的作用,实现农民利益与节水效果的最优配置。在工程管理中实行责任到人制,充分保证节水农业工程的安全与顺利实施,防治水资源浪费。同时,鼓励更多的农民亲自参与管理与决策,为组织管理提供更广泛的保障。

3.7　改变灌溉模式　单一的灌溉模式不利于节水农业的发展,当前应该结合该地区经济水平和资金投入情况,逐步改变以工程性灌溉过多,非工程性灌溉较少的局面,因地制宜的发展多种灌溉模式,广泛推广投资少、见效快、无污染、且易于被广大农民所接受的节水农业技术。

3.8　开展污水处理在节水农业中的二次利用　据估计所有农作物中大约有10%是用污水来灌溉的,通过对污水的处理,既可以保障生态环境的健康与安全,又可以实现污水的资源化及水资源的优化配置,解决缺水困难。但在利用污水的同时,要注意污水中的重金属、有机溶剂以及药品生产所残留的激素等物质可能带来的民生和环境问题[13]。因此,可以通过污水的严格有效处理,消除污水利用可能带来的威胁,真正的实现无害污水在农业中的二次利用,为发展节水农业开辟更多的水源,促进节水农业的发展。

3.9　调整作物种植结构,开发节水型作物　当前,在一些地区,干旱是进一步扩大作物种植量的最大限制因素,随着人们对全球气候变化可能带来的干旱认识的加深,人们也更加重视通过基因等手段来改变作物的耐旱特性,通过少消耗水实现作物的节水。此外,通过各种措施增加作物水分生产率,开发出水分生产率高的作物也是实现作物节水的良好途径[14]。另外,通过调整农作物的种植结构,如:减少玉米、小麦等水资源消耗大的作物的种植量,多种植花生等水量消耗相对较低的作物,也可以有效的节约农业水资源[15]。

3.10　将种植与畜牧相结合,提高水资源的重复利用率　据统计,每生产1 L 牛奶就要消耗大约700 L 的水,而每生产1 kg 的苜蓿干草大约需要600 L 的水,因此,在淮北的一些畜牧业较为发达的地区,可以通过将种植与畜牧相结合,来提高水资源的重复利用率,在实现更多的经济效益的同时,节约大量的农业用水资源[16]。

4　小结

淮北地区面临着资源型缺水,水资源时空分布不均,作为重要的粮食主产区,发展节水农业显得尤为重要。笔者就当前淮北地区节水农业发展中面临的一些突出问题,提出了相应的解决方案,为更好的发展淮北地区的节水农业提供借鉴与理论指导。

参考文献

[1] 韩波.淮北市水资源状况的调查与思考[J].淮北职业技术学院学报,2006,5(3):28-32.

[2] IRSHAD M, INOUE M, ASHRAF M.et al. The management options of water for the development of agriculture in dry areas[J]. Journal of Applied Sciences, 2007,7(11):1551-1557.

[3] 山仑.中国节水农业[M].北京:中国农业出版社,2004.

[4] 崔曾团.美国的节水农业及其启示[J].水土保持通报,2006,26(3):141-142.

[5] 林性粹. 旱区农田节水灌溉技术[M].北京:农业出版社,1991:73.

[6] 许迪.田间节水灌溉新技术研究与应用[M].北京:中国农业出版社,2002.

[7] 李家年,魏荣萍.安徽省淮北地区水文特性[J].治淮,1999(4):21-22.

[8] SUN H M.Developing models on water-saving agriculture through rainwater harvesting for supplemental irrigation in northern China semi-aid reigion[J]. Ying Yong Sheng Tai Xue Bao,2005,16(6):1072-1076.

[9] 李银国.柑橘园土肥水管理及节水灌溉[M].北京:金盾出版社,1997.

[10] 郭培章.中外节水技术与政策案例研究[M].北京:中国计划出版社,2003.

[11] 王会肖,薛明娇.节水农业推广的若干问题及对策建议[J].节水灌溉,2008(5):38-41.

[12] 董玉秀.山西省节水农业现状与发展对策[J].山西水利科技,2006(3):94-96.

[13] SCOTT C, FARUQUI N, RASCHID L, Wastewater use in irrigated agriculture: Confronting the livelihood and environmental realities[M]. United Kingdom: CABI,2004.

[14] 雷志栋,杨诗秀,谢森传,等.落实十五届三中全会决议,明确推广节水灌溉的主攻方向[C]// 中国节水农业问题论文集.北京:中国水利水电出版社,1999:141-146.

[15] CHAVES M M,OLIVEI M M.Mechanisms underlying plant resilience to water deficits: prospects for water-saving agriculture[J]. Journal of Experimental Botany, 2004,55(407):2365-2384.

[16] PRINZ DIETER, MALIK AMIR H. More yield with less water. How efficient can be water conservation in agriculture[J]. Elektronisches Volltextarchiv (EVA),2002,S1:18-35.

(上接第262 页)

生成样本不小于20 个即可具有较好的分辨能力和计算精度;②投影向量分量的取值范围取[0,1]更为合适;③实数编码加速遗传算法是优化投影目标函数的优选方法;④使用图解法或直接比较法更简单快捷。这些探索可为该模型在今后的实际应用提供有益参考。

参考文献

[1] 周惠成,董四辉.基于投影寻踪的水质评价模型[J].水文,2005,25(8):15-18.

[2] 杨晓华,杨志峰,郦建强.水质综合评价的遗传投影寻踪插值模型[J].环境工程,2004,22(3):69-72.

[3] 张欣莉,丁晶,李祚泳,等.投影寻踪新算法在水质评价模型中的应用[J].中国环境科学,2000,20(2):187-189.

[4] 金菊良,魏一鸣,丁晶,等.水质综合评价的投影寻踪模型[J].环境科学学报,2001,21(4):431-434.

[5] 彭坤泉,张平.投影寻踪模型在水环境质量评价中的应用[J].江苏环境科技,2007,20(S1):59-62.

[6] 叶浩,钱家忠,黄夕川,等.投影寻踪模型在地下水水质评价中的应用[J].水文地质工程地质,2005,32(5):9-12.

[7] 付强,付红,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,2003,23(4):236-239.

[8] 王顺久,张欣莉,侯玉,等.投影寻踪聚类分析在环境质量综合评价中的应用[J].重庆环境科学,2002,24(3):74-76.

[9] FRIEDMAN J H,TUKEY J W.A Projection pursuit algorithm for exploratory data analysis[J].IEEETrans on Computer,1974,23(9):881-890.

[10] 张宏伟,牛志广.Lingo8.0 及其在环境系统优化中的应用[M].天津:天津大学出版社,2005.

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