文章编号:1674-7070(2010)04-0320-04

基于粒子群优化的投影寻踪聚类模型及其应用

王李进[1]　胡欣欣[1]　宁正元[1]

摘要

投影寻踪聚类分析是根据设计的投影指标函数,并在相关约束条件下进行问题优化分析的过程.给出了用于求解投影指标函数的粒子群算法,并将构造的模型应用于森林承载力评价.仿真实验结果表明:与基于遗传算法优化的模型比较,基于粒子群优化的模型简单、容易实现并且没有许多参数需要调整;在应用上,基于粒子群优化的模型可获得更优的解,并可预计模型在森林承载力评价中具有重要的应用价值.

关键词

投影寻踪;粒子群优化;森林承载力

中图分类号 TP301.6

文献标志码 A

0　引言

Introduction

随着社会的发展,人们要求了解事物更多方面的性质,加之计算机技术日新月异的发展,使得高维数据的统计分析越来越重要.在许多实际问题中数据的维数很高,应用统计分析时,属于高维问题,这会降低参数估计的稳健性.在近代统计学中,出现了一种解决高维问题的统计方法——投影寻踪.

投影寻踪是一种新兴的、有价值的高新技术,是统计学、应用数学和计算机技术的交叉科学,是用来分析和处理高维观测数据,特别是非线性和非正态高维数据的一种统计方法[1].因具有稳健性好、抗干扰性强和准确度高等优点,其在工业、农业、水利和遥感等领域被广泛应用于求解预测、模式识别和分类等问题[1-2].

投影寻踪聚类分析是投影寻踪在应用上主要涉及的内容之一,实质上就是根据设计的投影指标,并在相关约束条件下进行的优化问题.Friedman 等[3]应用高斯-牛顿法求该优化问题;Zhao 等[4]应用梯度下降法寻优.上述传统的优化方法在处理多变量寻优时往往易陷入局部最优、早熟或提前收敛,寻求不到真正的最优解.文献[5-7]运用遗传算法(Genetic Algorithms,GA)优化投影寻踪目标函数.对遗传算法的研究表明,尽管算法在一定条件下具有全局收敛特性,但算法的交叉、变异和选择算子都是在概率意义下随机进行的,虽然保证了种群的群体进化性,但在一定程度上不可避免退化现象的出现,同时参数的设置也较为复杂.

粒子群优化算法(Particle Swarm Optimization,PSO)与遗传算法类似,是一种基于进化的优化工具.同遗传算法比较,PSO 的优势在于简单容易实现并且没有许多参数需要调整.目前已广泛应用于函数优化、神经网络训练、模糊系统控制以及其他遗传算法的应用领域.

因此,本文采用粒子群算法优化投影寻踪聚类模型的目标函数,同时将优化后的模型应用于森林资源经营管理,为森林生态系统的建模提供一种新的思路,也拓宽了模型的应用领域.

1　投影指标函数的数学描述

Mathematical description of projection index function

投影寻踪聚类分析是一种降维处理分析,是通过投影寻踪方法

收稿日期 2009-11-02

资助项目 福建省自然科学基金(2009J05043);福建省教育厅项目(JA08063)

作者简介

王李进,男,博士,讲师,主要从事计算机在林业中应用研究. style_wang@163.com

1　福建农林大学 计算机与信息学院,福州,350002

将多维分析问题转化为一维问题进行分析研究,其内容主要涉及投影指标的构造和最优投影方向的求解.在实际应用中,构建投影寻踪聚类模型,首先是对样本数据集进行归一化处理,其目的是为了消除各指标的量纲和统一各指标值的变化范围;其次,构造投影指标函数,以便获得最佳投影指标方向;最后,根据由最佳投影方向计算出各样本数据的投影值,进行分类或评价分析.

由于投影指标函数只随着投影方向的变化而变化,不同的投影方向反映不同的数据结构特征,而最佳投影方向是最大可能的反映高维数据某类结构特征.因此,最佳投影方向的求解是通过求解投影指标函数的最大化问题来估计.其数学描述[1]为

$$\text{投影指标函数最大值:}\ Q(a) = S_z \cdot D_z, \tag{1}$$

$$\text{约束条件满足:}\ \sum_{j=1}^{p} a_j^2 = 1. \tag{2}$$

其中,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:

$$z_i = \sum_{j=1}^{p} a_j x_{i,j}, \quad i = 1,2,\cdots,n; \tag{3}$$

$$S_z = \sqrt{\frac{\sum_{i=1}^{n} (z_i - E_z)^2}{n-1}}; \tag{4}$$

$$D_z = \sum_{i=1}^{n} \sum_{j=1}^{n} \left((R - r_{i,j}) \cdot u(R - r_{i,j}) \right). \tag{5}$$

其中:E_z 为投影值 Z_i 的平均值;R 为局部密度的窗口半径,可以根据经验值来确定;$r_{i,j}$表示样本之间的距离;$u(t)$为一单位阶跃函数,当 $t \geqslant 0$ 时,其值为 1,否则为 0;$x_{i,j}$是归一化后的样本数据.

2 求解投影指标函数的粒子群算法
PSO for projection index function

2.1 粒子群算法

PSO 算法求解优化问题时,问题的解对应于搜索空间中一只鸟,它被抽象为没有质量和体积的"粒子",并将其延伸到 N 维空间.每个粒子都有自己的位置和速度(决定飞行的方向和距离),还有一个由被优化问题决定的适应值.每个粒子知道自己到目前为止发现的最好位置和现在的位置,除此之外,每个粒子还知道到目前为止整个群体中所有粒子发现的最好位置,可看做是粒子同伴的经验.各个粒子记忆并追随当前的最优粒子在解空间中搜索,这样,每次迭代的过程不是完全随机的,如果找到较好解,将会以此为依据来寻找下一个解.

PSO 算法通常采用随机化的方式为粒子产生初始位置和速度(随机初始解).假设 d 维搜索空间的各个粒子的位置和速度分别为 $X = (x_1, x_2, \cdots, x_d)$ 和 $V = (v_1, v_2, \cdots, v_d)$,在此之后的每一次迭代中,粒子通过跟踪两个最优解来更新自己,第一个就是粒子本身所找到的最好解,即个体极值点 p_b,另一个是整个粒子群目前找到的最好解,称为全局极值点 g_b.在找到这两个最优值时,粒子根据式(6)和式(7)来更新自己的速度和新的位置.

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_{bi}^t - x_i^t) + c_2 r_2 (g_{bi}^t - x_i^t), \tag{6}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1}. \tag{7}$$

其中:ω 是惯性系数,其主要作用是产生扰动,以防止算法的早熟收敛;c_1 和 c_2 是加速系数(或称学习因子),分别调节向个体最好粒子和全局最好粒子方向飞行的最大步长,若太小,则粒子可能远离目标区域,若太大,则会导致突然向目标区域飞去,或飞过目标区域,合适的 c_1和 c_2,可在加快收敛速度的同时还能不易陷入局部最优,通常令 $c_1 = c_2 = 2$;r_1 和 r_2 为 0 和 1 之间均匀分布的随机数.

粒子群算法发展到现在有很多种变形及其改进算法.为了有效地控制粒子的飞行速度使算法达到全局探测与局部开采两者的有效平衡,本文采用带压缩因子的粒子群算法求解投影指标函数,即式(6)中的 $\omega = 0.729$,$c_1 = c_2 = 1.494\ 45$.

2.2 算法描述

投影指标函数优化问题是属于具有约束优化问题,而求解此类问题的关键在于如何处理约束.处理约束的主要方法有保证解的合理性、惩罚函数法、区分可行和不可行域方法和其他混合方法[8].本文采用根据约束条件方程选择保证解的合理性方法,即只有当解在可行域的前提下,粒子才能停止初始化和进行经验更新.用于优化求解投影指标函数的粒子群算法描述如下:

1) 选定粒子群种群规模 n;

2) 设 $\boldsymbol{x}_i$ 为种群中第 i 个粒子的位置向量;

3) 设 $F_{\text{fitness}}(\boldsymbol{x}_i)$为求第 i 个粒子的适值函数,如式(3)所示;

4) 设 $F_{\text{voilent}}(\boldsymbol{x}_i)$为第 i 个粒子的约束函数,如式(2)所示;

5) 设 $\boldsymbol{v}_i$ 为第 i 个粒子的速度向量;

6) 设 $\boldsymbol{m}_i$ 为第 i 个粒子更新的中间代;

7) 设 $\boldsymbol{p}_g$ 为种群中适应度最高的位置向量;

8) 设 $\boldsymbol{p}_i$ 为第 i 个粒子自身搜索到的最优位置向量.

第 1 步. (初始化)对于每一个种群中的粒子 i, $i=1,2,\cdots,n$.

1) 随机初始化 $\boldsymbol{x}_i$,使满足 $F_{\text{voilent}}(\boldsymbol{x}_i)$约束函数;

2) 随机初始化 $\boldsymbol{v}_i$;

3) 计算 $F_{\text{fitness}}(\boldsymbol{x}_i)$,并令 $\boldsymbol{p}_i=\boldsymbol{x}_i$;

4) 以种群中适应值最优的粒子的位置向量初始化 $\boldsymbol{p}_g$;

5) 以 $\boldsymbol{x}_i$ 初始化 $\boldsymbol{p}_i$.

第 2 步. 循环迭代,直到满足 PSO 终止条件为止.

1) 选择算法惯性因子 ω;

2) 对每个粒子,计算其适应值 $F_{\text{fitness}}(\boldsymbol{x}_i)$. 若 $F_{\text{fitness}}(\boldsymbol{x}_i)>F_{\text{fitness}}(\boldsymbol{p}_i)$,则 $\boldsymbol{p}_i=\boldsymbol{x}_i$;

3) 搜索 $\boldsymbol{p}_g$ 值:若 $F_{\text{fitness}}(\boldsymbol{p}_i)>F_{\text{fitness}}(\boldsymbol{p}_g)$且满足约束函数 $F_{\text{voilent}}(\boldsymbol{p}_i)$,则 $\boldsymbol{p}_g=\boldsymbol{p}_i$;

4) 对每个粒子,$\boldsymbol{v}_i$ 按公式(6)更新,$\boldsymbol{m}_i=\boldsymbol{x}_i+\boldsymbol{v}_i$;

5) 更新 $\boldsymbol{x}_i$:若满足 $F_{\text{voilent}}(\boldsymbol{x}_i)$ 且满足 $F_{\text{voilent}}(\boldsymbol{m}_i)$,$\boldsymbol{x}_i=\left(F_{\text{fitness}}(\boldsymbol{x}_i)>F_{\text{fitness}}(\boldsymbol{m}_i)\right)?\ \boldsymbol{x}_i:\boldsymbol{m}_i$;若满足 $F_{\text{voilent}}(\boldsymbol{x}_i)$ 且不满足 $F_{\text{voilent}}(\boldsymbol{m}_i)$,$\boldsymbol{x}_i=\boldsymbol{x}_i$;若满足 $F_{\text{voilent}}(\boldsymbol{m}_i)$ 且不满足 $F_{\text{voilent}}(\boldsymbol{x}_i)$,$\boldsymbol{x}_i=\boldsymbol{m}_i$;若不满足 $F_{\text{voilent}}(\boldsymbol{x}_i)$ 且不满足 $F_{\text{voilent}}(\boldsymbol{m}_i)$,$\boldsymbol{x}_i$ 取违反约束函数值小的粒子.

3 在森林承载力评价中的应用

Application of projection pursuit cluster model in forest carrying capacity evaluation

投影寻踪聚类模型在应用上可归纳为两个方面,一是运用投影特征值对样本进行合理分类;二是根据给定的判别标准利用投影值对评价样本进行等级水平评价.

本文以 1998 年连续清查数据和社会经济发展数据(表 1[9]),运用投影寻踪聚类模型对闽江流域森林承载力进行评价分析. 其核心思想是将森林承载力的各等级阈值作为样本数据,运用投影寻踪聚类模型,求得各评价指标的最佳投影方向,再根据最佳投影方向计算各样本数据的特征值以及评价样本的特征值,最后根据待评价样本的特征值与各样本数据的特征值的关系给予评价.

表 1 评价样本数据

Table 1 Evaluation sample data

指标	Ⅰ级	Ⅱ级	Ⅲ级	Ⅳ级	评价样本
A1	70.000	100.000	130.000	160.000	82.590
A2	10.000	30.000	50.000	60.000	64.960
A3	100.000	90.000	80.000	70.000	91.800
A4	10.000	20.000	30.000	50.000	65.000
A5	10.000	20.000	30.000	50.000	87.000
A6	10.000	20.000	30.000	50.000	65.000
A7	5.000	10.000	15.000	29.000	8.300
A8	5.000	10.000	15.000	19.890	6.080
A9	1.000	0.800	0.600	0.400	1.390
A10	1.000	0.800	0.600	0.400	1.910
A11	0.500	0.800	1.000	2.400	1.080
A12	0.355	0.300	0.200	0.100	0.310
A13	5.000	8.000	12.000	15.000	13.350
A14	2.000	4.000	6.000	7.200	12.500
A15	300.000	100.000	60.000	20.000	159.000
A16	20.000	40.000	60.000	70.000	78.300
A17	10.000	30.000	50.000	60.560	28.190
A18	0.400	0.600	0.800	1.210	0.813

注:A1、A2、…、A18 见文献[9].

利用投影寻踪聚类模型对闽江流域森林承载力进行评价,经在 Matlab 环境下仿真,求得投影指标函数最大值是 0.584 3,最佳投影方向为(0.402 8, 0.289 2, 0.047 3, 0.058 3, 0.290 0, 0.349 7, 0.126 9, 0.364 4, 0.038 7, 0.152 8, 0.369 5, 0.040 4, 0.314 7, 0.085 0, 0.053 9, 0.071 6, 0.123 0, 0.319 0),其中粒子群种群大小为 40,最大迭代次数为 100. 将最佳投影方向带入式(3)求得样本数据的投影值,根据文献[9]的判断方法,可判断闽江流域森林承载力等级为Ⅲ级. 与文献[9]的基于遗传算法优化的研究方法比较,结果列于表 2.

表 2 POS 方法和 GA 方法结果比较

Table 2 Results comparison between PSO-based and GA-based model

指标	PSO	GA[9]
投影指标函数最大值	0.584 3	0.575 3
样本综合投影值	0.225 7	0.571 8
	0.955 8	1.347 0
	1.721 5	2.135 5
	2.834 3	3.381 8
	2.112 3	1.939 9
评价结果	Ⅲ级	Ⅲ级

分析表2可知,从评价结果上看,基于粒子群优化的评价模型与基于遗传算法优化的评价模型的结果是一致的,说明了基于粒子群优化评价模型是可行的.从投影指标函数的最大值看,基于PSO的优化效果优于基于GA的,其所求的最佳投影方向更最大可能地反映了高维数据某类结构特征.从各样本数据的综合投影值上看,基于PSO优化的分布更均匀.

4 结束语
Summary

投影寻踪聚类分析,实质上是通过求解最佳投影方向转化为一维问题进行分析研究.采用粒子群优化算法求解最佳投影方向较遗传算法简单、容易实现且不需要设置较多的参数.

投影寻踪聚类模型已广泛应用于工业、农业、水利和遥感等领域.本文应用于林业中的森林承载力评价,评价结果与期望的结果是一致的,表明评价方法是可行而且是有效的.可以预计基于粒子群优化的投影寻踪聚类模型在森林承载力评价中具有重要的应用价值,可为森林生态系统建模提供新的思路和建模手段,也为其区域可持续发展研究中的广泛应用奠定基础.

参考文献
References

[1] 付强,赵小勇.投影寻踪模型原理及其应用[M].北京:科学出版社,2006
FU Qiang,ZHAO Xiaoyong. Projection pursuit model and its applications[M]. Beijing:Science Press,2006

[2] 田铮,林伟.投影寻踪方法与应用[M].西安:西北工业大学出版社,2008
TIAN Zheng,LIN Wei. Projection pursuit method and its applications[M]. Xi'an:Northwestern Polytechnical University Press,2008

[3] Friedman J H,Stuetzle W. Projection pursuit regression[J]. Journal of the American Statistical Association,1981,76(376):817-823

[4] Zhao Y,Atkeson C G. Implementing projection pursuit learning[J]. Neural Networks,IEEE Transactions on Neural Network,1996,7(2):362-373

[5] 张欣莉,丁晶,郑祖国.投影寻踪回归在紫坪埔洪水预报中的应用[J].四川大学学报:工程科学版,2000,32(2):22-24
ZHANG Xinli,DING Jing,ZHENG Zuguo. Application of projection pursuit regression in forecasting Zipingpu's flood[J]. Journal of Sichuan University:Engineering Science Edition,2000,32(2):22-24

[6] 金菊良,刘永芳,丁晶,等.投影寻踪模型在水资源工程方案优选中的应用[J].系统工程理论方法应用,2004,13(1):81-84
JIN Juliang,LIU Yongfang,DING Jing,et al. Application of projection pursuit model to optimal choice of water resources engineering schemes[J]. Systems Engineering-Theory Methodology Application,2004,13(1):81-84

[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(2):236-239
FU Qiang,FU Hong,WANG Likun. Study on the PPE Model based on RAGA to evaluating the water quality[J]. Scientia Geographica Sinica,2003,23(2):236-239

[8] 段晓东,王存睿,刘向东.粒子群算法及其应用[M].沈阳:辽宁大学出版社,2007
DUAN Xiaodong,WANG Cunrui,LIU Xiangdong. Particle swarm optimization and application[M]. Shenyang:Liaoning University Press,2007

[9] 王李进.闽江流域森林承载力评价与预测研究[D].北京:北京林业大学信息学院,2008
WANG Lijin. Study on evaluation and prediction on forest carrying capacity of Minjiang River Basin[D]. Beijing:School of Information Science & Technology,Beijing Forestry University,2008

Projection pursuit cluster model based on particle swarm optimization and its application

WANG Lijin[1] HU Xinxin[1] NING Zhengyuan[1]
1 College of Computer & Information Science,Fujian Agriculture and Forestry University,Fuzhou 350002

Abstract Based on projection index function,projection pursuit cluster analyis is a process to analyze optimization problems under certain constrain conditions. This paper employs Particle Swarm Optimization(PSO) to solve projection index function,and constructs a projection pursuit cluster model to evaluate forest carrying capacity. Simulation results show that this PSO-based model is simpler and easier to realize with less parameters,compared with GA-based model. Moreover,the proposed model can give optimal solutions in application,which verify its practical significance in forest carrying capacity evaluation and other regional sustainable development research.

Key words projection pursuit;particle swarm optimization;forest carrying capacity