

基于SVR模型的重庆市生态安全指标预测模型研究

冯海亮 夏磊 黄鸿

(重庆大学光电技术及系统教育部重点实验室 重庆 400044)

摘要 鉴于实际工作中生态安全指标数据统计的滞后性,通过采集并整理重庆市1988年至2007年与生态安全指数极度相关的28项指标构建了重庆市生态安全指标体系,并以此作为训练样本数据,以2008、2009年对应指标作为输出结果,分析比较多变量灰色预测模型、径向基神经网络、支持向量回归方法,通过对比真实统计数据做出误差分析,得到支持向量回归重庆市生态安全指标预测模型。该模型对重庆市主要生态安全指标进行预测的结果具有明显的优势,可以用于实际预测。

关键词 生态安全预测,多变量灰色模型,径向基神经网络,支持向量回归

中图分类号 X821 **文献标识码** A

Study on Prediction Model of Ecological Security Index in Chongqing City Based on SVR Model

FENG Hai-liang XIA Lei HUANG Hong

(Key Lab. on Opto-electronic Technique and Systems, Ministry of Education, Chongqing University, Chongqing 400044, China)

Abstract In actual practice, owing to hysteresis of the conventional statistical analysis on ecological safety index, this article implemented Multivariable Grey model, Radical Basis Function Network and Support Vector Regression to input extremely relevant samples of ecological safety index from 1988 to 2007 in Chongqing. The outputs generated by the three models were evaluated and compared with ecological safety index gathered in 2008 and 2009. According to the error analysis between the outputs and the actual index, more accurate predictions were produced by the Support Vector Regression model. In conclusion, the Support Vector Regression model is applicable to actual practice and has higher accuracy than the other two models.

Keywords Ecological prediction, Multivariable grey model, Radical basis function, Support vector regression

1 引言

随着重庆市经济的跨越式发展以及城市化、工业化进程的加快,土地利用变化引起众多生态环境问题,这会在未来影响到普通市民的正常生活。为此,开展生态环境指标变化的动态预测,寻找在未来几年中可能会受到严重影响的地区^[1],为职能部门做出快速、正确的决策提供科学依据,具有极其重要的现实意义。

从1972年在斯德哥尔摩举行的联合国人类环境会议开始,建立生态安全指标的预测模型就成为各国研究学者关注的焦点之一。在时间序列预测中,线性的概率统计模型曾得到广泛的应用。如今,最常用的方法属灰色预测、神经网络、支持向量回归等方法。

伴随非线性理论和人工智能技术的高速进步,近年来以SVR等为代表的计算方法已经成为有力的分析和预测工具^[2]。本文分析讨论了多变量灰色预测模型、径向基神经网络、支持向量回归的方法各自的特点,编写了3种方法的预测程序,最后将这3种方法分别运用到重庆市生态安全指数预

测系统中并进行对比。实验结果表明,采用SVR理论方法的预测系统较其余两种方法而言更加适用于重庆市生态安全指标预测系统。

2 MGM、RBF原理介绍

2.1 MGM多变量灰色模型预测理论^[3]

多变量灰色系统是指既含有已知信息又含有未知信息的系统。那些难以精确地建立数学模型的系统都属于灰色系统的范畴。在灰色系统理论中,称抽象系统的逆过程为灰色模型。它是根据关联度、生成数的灰倒数、灰微分等观点和一系列数学方法建立起来的连续性的微分方程。灰色预测是灰色系统理论的一个重要方面,它利用这些信息建立灰色预测模型,从而确定系统未来的变化趋势。灰色预测模型能够根据现有少量信息进行计算和推测,因而在人口、经济、生态、农业、医学、工程、气象、水文等许多部门得到了广泛应用。

由于灰色模型中的一阶微分方程都是已知信息的线性组合,因此仅可以得到预测数据的一组采样值。而由拉格朗日中值定理可知,一阶微分方程都存在较大误差,从而引起辨别

到稿日期:2012-10-30 返修日期:2013-01-26 本文受重庆市科技攻关重点项目(CSTC2009AB2231)资助。

冯海亮(1962—),男,博士,教授,主要研究方向为应用数学、流形学习、人脸识别, E-mail: fhliang@cqu.edu.cn; 夏磊(1988—),男,硕士,主要研究方向为模式识别、图像处理; 黄鸿(1980—),男,博士,讲师,主要研究方向为遥感图像处理、模式识别、流形学习。

矩阵有较大误差,所以最后所得其辨识值以及最后模型的时间响应函数的误差将无法估量。

2.2 RBF 径向基神经网络预测原理

由于多变量灰色预测模型中出现的中值误差问题,引入前馈型的、具有很强非线性拟合能力的 RBF 神经网络进行计算。

1988 年, Broomhead 以及 Moody 和 Darken 首先将 RBF (Radial Basis Function, 径向基神经网络) 模型应用于人工神经网络的设计, 并成功把 RBF 神经网络应用于模式识别, 构成了 RBF 神经网络^[4]。RBF 神经网络是一种性能优良的前馈型神经网络, 能以任意人为设定的精度逼近非线性函数, 具有全局逼近能力。Poggio 和 Girosi 已经证明, RBF 网络是连续函数的最佳逼近^[5]。只有当输入接近 RBF 网络的接受域时, 网络才会做出响应, 避免了局部极小的问题以及输入层与隐含层之间的反向传播繁琐冗长的计算。

RBF 神经网络的构成可以简要概括为 3 层: 第一层为输入层, 第二层为隐含层, 第三层为输出层。从输入层到隐含层的变换是非线性的, 而从隐含层到输出层的变换是线性的。其学习方法分为两个步骤, 第一步是通过聚类算法得到初始的权值, 第二步是根据训练数据训练网络的权值。需要通过网络学习得到参数为第 i 个隐节点的中心 C_i 、基函数的宽度 σ_i 以及隐含层到输出层的权重 ω_{jk} 。RBF 神经网络学习算法中, 常用的聚类分析算法是 K-均值算法, 其用来确定隐节点的中心 C_i ; C_i 确定后, 可以求得基函数的宽度 σ_i ; 最后利用最小二乘法确定权重 ω_{jk} ^[6]。

RBF 神经网络运用在数据预测领域中也存在着一些不足和难以解决的问题, 由此产生的系统误差也是不能控制的^[7]。具体问题总结为如下 4 点:

- (1) 不能依据推理过程给出明确的推理依据;
- (2) 对问题的特征进行了简化处理, 将其变换为数字, 这样势必导致信息的丢失;
- (3) 从数据点中任意选取中心构造出的网络, 其性能显然是不能完全令人满意的;
- (4) 训练网络建立初始, 模型的中心点选择过多, 导致在优选过程中经常出现数据病态的现象。

3 SVR 支持向量回归预测方法

针对 RBF 神经网络存在的这些问题, 引入支持向量机 (Support Vector Machine, SVM) 的方法, 延伸出支持向量回归的预测模型, 从而有效地回避了神经网络的部分缺陷。

支持向量机的方法最早是 1995 年 Vapnik 针对模式识别问题提出的, 随着对 ϵ 不敏感损失函数的引入, 已将其推广应用到非线性回归和曲线拟合中, 得到了用于曲线拟合的回归型支持向量机方法 (Support Vector machine for Regression, SVR), 其在非线性系统辨识、预测预报、建模与控制等领域都有广泛应用, 因此将支持向量回归的方法用于城市生态安全指标的预测成为可能。支持向量机是统计学习理论的重要组

成部分, 它主要采用 SVC 分类算法和 SVR 回归算法两类。其中 SVR 算法已被成功运用于时间序列预测研究, 并表现出了很好的性能^[8]。

由于重庆市各项生态安全指标按年份统计, 数据呈现离散状态, 因此采用非线性回归的方法对指标数据进行预测。

首先假设给定训练数据为: $\{(x_i, y_i), i=1, 2, \dots, l\}$, 其中 $x_i \in \mathbf{R}^d$ 是第 i 个学习样本的输入值, 且为 d 维向量 $x_i = [x_i^1, x_i^2, \dots, x_i^d]^T$, $y_i \in \mathbf{R}$ 为对应的目标值。先定义线性 ϵ 不敏感损失函数为:

$$|y - f(x)|_{\epsilon} = \begin{cases} 0, & |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon, & |y - f(x)| > \epsilon \end{cases}$$

对于非线性支持向量机回归, 其基本思路是将输入样本空间非线性变换到另一个高维空间, 在这个高维空间中构建回归估计函数, 而这种非线性变换是通过定义适当的核函数 $K(x_i, x_j)$ 来实现的^[9]。其中 $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$, $\varphi(x)$ 为某一非线性函数。所以可以假设非线性情况下的回归估计函数为:

$$f(x) = \omega^T \cdot \varphi(x) + b$$

在采用 SVR 算法处理数据之前, 首先要对 3 个参数进行确定: ϵ 、 C 与核函数 $K(x_i, x_j)$ 。将支持向量回归算法运用到城市生态安全指标预测中, 支持向量机的核函数决定了回归性能。实验表明, 基于径向基核函数的支持向量机在回归估计中, 其性能优于其他核函数的支持向量机^[10]。确定了核函数, 其相关参数就决定了支持向量回归机的回归精度^[11]。在确定 ϵ 和 C 时, 采用的方法类似穷举, 即限制迭代次数使误差小于某一绝对值, 然后将此时的 ϵ 和 C 值代入模型进行数据预测。

4 重庆市生态安全指标预测模型的建立和实验分析比较

4.1 数据预处理及指标体系的建立

为了验证本文模型的有效性, 预测模型全部数据来自各年份的重庆市统计年鉴、环境公报及统计局文献^[12]。本文所建立的模型采用重庆市 1988 年至 2007 年的生态安全指标数据, 2008、2009 年的数据作为目标进行预测, 并对比 2008 年、2009 年的真实数据得出模型的误差。模型中所采用的各项指标数据包括生态环境压力、生态环境状态、人文环境响应 3 个方面, 具体为: 人口密度 (D_1)、人口自然增长率 (D_2)、城市化率 (D_3)、人均住房面积 (D_4)、人均水资源量 (D_5)、人均煤炭消费量 (D_6)、人均汽油消费量 (D_7)、人均年供气量 (D_8)、人均 GDP (D_9)、每万人拥有医生数 (D_{10})、每万人拥有公交车辆 (D_{11})、每万人拥有病床数 (D_{12})、森林覆盖率 (D_{13})、城市绿化覆盖率 (D_{14})、人均耕地面积 (D_{15})、人均公共绿地 (D_{16})、固废无害化处理率 (D_{17})、废水处理率 (D_{18})、酸雨频率 (D_{19})、悬浮微粒 (D_{20})、单位面积化肥使用量 (D_{21})、人均 GDP 增长率 (D_{22})、环保投资占 GDP 比重 (D_{23})、科技投资占 GDP 的比重 (D_{24})、第三产业占 GDP 的比重 (D_{25})、万人拥有大学以上文

化程度人数(D₂₆)、每万人藏书量(D₂₇)、万人均区内各类学校(D₂₈)。

为简化计算,引入了逐步回归模型^[13]构建城市生态安全指数的预测模型。对重庆市主城九区的生态安全指标数据进行分析,以北碚区为例进行数据分析。

根据逐步线性回归模型原理,各个指标变量进入模型的情况如表 1 所列。

表 1 输入的变量

模型	输入的变量	方法
1	悬浮微粒 D ₁₆	步进(准则:F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)
2	人均用水 D ₅	步进(准则:F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)
3	人口增长率 D ₂	步进(准则:F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)
4	绿化率 D ₁₉	步进(准则:F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)
5	人均汽油 D ₇	步进(准则:F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)

由表 1 可以得到由逐步引入(剔除法)依次引入各个指标变量,最终有 D₂、D₅、D₇、D₁₆、D₁₉ 这 5 个变量没有被剔除。表 2 列举了重庆市主城九区经过剔除法后所选定的指标以及由逐步回归模型生成的重庆市主城九区生态安全指数 Y 的计算表达式。

表 2 重庆市主城九区生态安全指数计算表达式

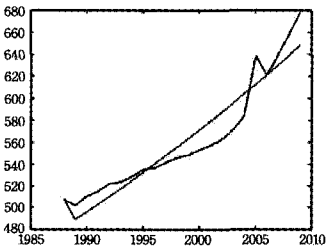
区域	所选指标	计算表达式
沙坪坝区	D ₁ D ₄ D ₂₂ D ₂₃	$Y=0.412-3.2 \times 10^{-15} D_1+6 \times 10^{-3} D_4+10^{-3} D_{22}+0.025 D_{23}$
南岸区	D ₁₆ D ₂₀ D ₂₃	$Y=0.733-0.858 D_{16}-0.402 D_{20}+0.012 D_{23}$
北碚区	D ₂ D ₅ D ₁₆ D ₁₉ D ₂₂	$Y=0.253-7 \times 10^{-3} D_2+10^{-3} D_5-0.508 D_{16}+7 \times 10^{-3} D_{19}+10^{-3} D_{22}$
渝中区	D ₂ D ₁₁ D ₂₆	$Y=0.2653-5 \times 10^{-3} D_2+7 \times 10^{-3} D_{11}-6.17 \times 10^{-3} D_{26}$
渝北区	D ₁₆ D ₁₈	$Y=0.43-0.26 D_{16}+6 \times 10^{-3} D_{18}$
巴南区	D ₄ D ₅ D ₁₅	$Y=0.585+5 \times 10^{-3} D_4-1.081 D_5-2 \times 10^{-3} D_{15}$
九龙坡区	D ₁ D ₄ D ₁₂ D ₁₃ D ₁₅ D ₂₈	$Y=0.399-3.5 \times 10^{-5}+4 \times 10^{-3} D_4+10^{-3} (D_{12}+D_{13}-D_{15})+5 \times 10^{-3} D_{28}$
江北区	D ₂ D ₆ D ₂₀ D ₂₁ D ₂₄	$Y=0.376-8 \times 10^{-3} D_2-0.77 D_6+0.122 D_{20}+0.011 D_{21}+0.03 D_{24}$
大渡口区	D ₁ D ₂ D ₇ D ₁₁ D ₂₀ D ₂₃	$Y=0.3-8.3 \times 10^{-5} D_1-0.03 D_2+0.064 D_7+3 \times 10^{-3} D_{11}+0.172 D_{20}+0.044 D_{23}$

4.2 预测算法分析及比较

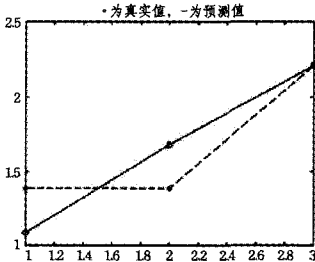
采用 Matlab 平台对 3 种算法分别进行了仿真,以分析每种算法在预测过程中不同的因素对预测结果和拟合过程的影响。在 SVR 算法中,针对重庆市不同区域与不同的生态指标数据,选择不同的 ε 和 C 参数,并且适时调整核函数 K(x_i,

x_j)自身参数,以保证拟合程度接近真实值变化的情况。同样,在 RBF 与 MGM 模型中也会适时调整参数,以保证拟合程度接近真实值。

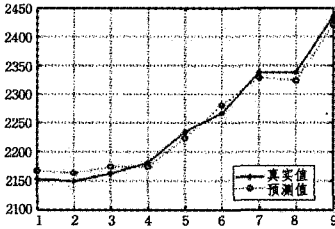
图 1 依次列举了多变量灰色模型用于渝北区人口密度的拟合曲线、径向基神经网络用于大渡口区环保投资占 GDP 比重的拟合曲线、支持向量回归用于大渡口区人口密度的拟合曲线。



(a) MGM(1, n)模型用于渝北区人口密度的拟合曲线



(b) RBF 模型用于大渡口区环保投资占 GDP 比重的拟合曲线



(c) SVR 模型用于大渡口区人口密度的拟合曲线

图 1

由图 1 可以看出,由 SVR 模型建立的生态指标预测模型在数据拟合的程度上,明显优于多变量灰色预测模型和 RBF 神经网络模型。鉴于系统实际的预测要求,选择基于 SVR 模型的生态指标预测模型为宜。

将采用逐步线性回归法剔除冗余指标后的数据分别送入 3 种算法中,并调整对应参数,计算出 3 种算法输出的重庆市主城九区 2008、2009 年分别对应的生态安全指数 Y,如表 3 所列,进而得到每种算法对重庆市生态安全指标预测的误差,如表 4 所列。

表 3 重庆市主城九区 2008、2009 年生态安全指数

年份	区域	MGM	RBF	SVR	真实值	年份	区域	MGM	RBF	SVR	真实值
2008	沙坪坝区	0.730687	0.585081	0.607983	0.625829	2009	沙坪坝区	0.571465	0.617366	0.619827	0.632005
2008	南岸区	0.521893	0.602472	0.598007	0.598015	2009	南岸区	0.630545	0.599302	0.594415	0.594409
2008	北碚区	0.764362	0.703694	0.658068	0.63761	2009	北碚区	0.763681	0.738527	0.691101	0.667292
2008	渝中区	0.604681	0.534518	0.535294	0.55597	2009	渝中区	0.694225	0.647754	0.622042	0.627299
2008	渝北区	0.656897	0.586601	0.599415	0.60386	2009	渝北区	0.659486	0.59638	0.624916	0.623033
2008	巴南区	0.734685	0.649506	0.65749	0.658313	2009	巴南区	0.703515	0.647639	0.65015	0.652437
2008	九龙坡区	0.579084	0.616844	0.64261	0.645415	2009	九龙坡区	0.680298	0.629914	0.632649	0.640324
2008	江北区	0.516843	0.566077	0.606708	0.584345	2009	江北区	0.581025	0.521072	0.516896	0.493404
2008	大渡口区	0.667425	0.596071	0.557028	0.551003	2009	大渡口区	0.592467	0.516945	0.561898	0.561079

表4 重庆市生态安全指标预测误差(%)

年份	区域	MGM	RBF	SVR	年份	区域	MGM	RBF	SVR
2008	沙坪坝区	16.76	6.511	2.852	2009	沙坪坝区	9.579	2.316	1.927
2008	南岸区	12.73	0.7453	0.001338	2009	南岸区	6.079	0.8231	0.001009
2008	北碚区	19.88	10.36	3.209	2009	北碚区	14.44	10.68	3.568
2008	渝中区	8.761	3.858	3.719	2009	渝中区	10.67	3.261	0.838
2008	渝北区	8.783	2.858	0.7361	2009	渝北区	5.851	4.278	0.3022
2008	巴南区	11.6	1.338	0.125	2009	巴南区	7.829	0.7353	0.3505
2008	九龙坡区	10.28	4.427	0.4346	2009	九龙坡区	6.243	1.626	1.199
2008	江北区	11.55	3.126	3.827	2009	江北区	17.76	5.608	4.761
2008	大渡口区	21.13	8.179	1.093	2009	大渡口区	5.594	7.866	0.146
均值		13.5	4.601	1.777	均值		9.339	4.132	1.455

从表3和表4的结果可以看出:采用MGM模型构建的重庆市生态安全指标预测系统存在着较大的误差,2008年主城九区的误差均值更是达到了13.5%,可以认为这种方法不适用于重庆市生态安全指标的预测;采用RBF模型构建的重庆市生态安全指标预测系统的误差均值在2008、2009年都小于5%,但是由于其局部预测误差过高(2008年:北碚区、大渡口区;2009年:大渡口区),可以认为径向基神经网络对于重庆市生态安全指标的预测不具备普适性;采用SVR模型构建的重庆市生态安全指标预测系统的误差均值为1.777%和1.455%,2008、2009年主城九区预测误差均小于5%,可以认为支持向量回归模型构建的系统具有较高的可信度。从而得到采用SVR模型构建的预测系统较采用MGM模型或RBF模型构建的预测系统具有更高精度的结论。所以,SVR模型能够用于重庆市生态安全指标预测系统。

结束语 本文针对重庆市生态安全指标预测系统的构建,对现有的多变量灰色模型、径向基神经网络以及支持向量回归的方法进行仿真对比,其中采用逐步回归模型简化计算,最后求出了2008、2009年重庆市主城九区生态安全指标的预测误差及其均值,得出了采用支持向量回归模型构建的重庆市生态安全指标预测系统具有较高的精度、可以用于实际系统模型的构建的结论。本文并未对传统的3种预测方法进行修改,在防止“过拟合、过训练”的情况下,使模型的拟合程度较优,减小了模型误差,提高了预测精度。

但是本文中所含算法需要不断修改各个模型的参数以使模型拟合程度较好,计算复杂度偏大;而且当已知样本数量过少时,这种支持向量回归的方法的误差也有增大的趋势,如何解决这些问题,还需进一步研究和实践。

在实际的工业过程中,系统的信息往往只能从系统运行的操作数据中分析得到,因此如何从系统运行的操作数据中提取系统运行状况的信息,用以降低网络对训练样本的依赖,在实际应用中将具有非常大的价值。

参 考 文 献

[1] 黄鸿,杨媚,冯海亮,等. PSR模型支持下重庆市主城区生态安全评价[J]. 微计算机信息,2012,28:1-3

[2] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer, 1995: 12-38

[3] 熊和金,徐华中. 灰色控制 [M]. 北京: 国防工业出版社, 2005: 47-50

[4] Nekoukar, Beheshti H, Taghi M. A local linear radial basis function neural network for financial time-series forecasting [J]. 2010, 33(3): 352-356

[5] Poggio T, Girosi F. Theory of networks for approximation and learning [R]. A I Merno No 1140. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge Mass, 1989

[6] Liu Hong, Cui Wen-hua, Zhang Qing-ling. Nonlinear combination forecasting model and application based on radial basis function neural networks [C] // Control, Automation and System Engineering, 2009. CASE 2009. IITA International Conference. 2009: 387-390

[7] Beatson R K, Levesley J, Mouat C T. Better bases for radial basis function interpolation problems [J]. Journal of Computational and Applied Mathematics, 2011, 236(4): 434-446

[8] Pai Ping-feng, Kuo-Ping Lin, Chi-Shen Lin. Time series forecasting by a seasonal support vector regression model [J]. Expert Systems with Applications, 2010, 37(6): 4261-4265

[9] Al-Anazi A F, Gates I D. Support vector regression for porosity prediction in a heterogeneous reservoir: A comparative study [J]. Computers & Geosciences, 2010, 36(12): 1494-1503

[10] Stoean R. Evolutionary support vector regression machines [C] // SYNASC'06 Proceedings of the Eighth International Symbolic and Numeric Algorithms for Scientific Computing. 2006: 330-335

[11] Alenezi, Abdulrahman. Real-time prediction of order flow times using support vector regression [J]. Computers & Operations Research, 2008, 35(11): 3489-3503

[12] 重庆市统计局, 国家统计局重庆调查总队. 重庆统计年鉴 [M]. 北京: 中国统计出版社, 1988-2009

[13] Lan Yu-qing, Guo Shu-hang. Multiple stepwise regression analysis on knowledge evaluation [C] // 2008 International Conference on Management of e-Commerce and e-Government, 2008