

近岸海域生态环境模拟中水信息学的应用

向先全¹ 陶建华^{1 2}

(1. 天津大学环境科学与工程学院, 天津 300072; 2. 天津大学机械工程学院, 天津 300072)

摘要 介绍水信息学的产生背景。随着海量数据的增长和信息技术的发展, 越来越多的水信息学方法应用于近岸海域生态环境模拟研究中。以渤海湾生态环境模拟为例, 重点讨论遗传算法优化支持向量机和模糊模式识别优化遗传神经网络在生态环境模拟中的应用。以叶绿素 a 的质量浓度表征浮游植物生物量作为模型的输出, 构建 SVM 生态模型, 并将 BP 神经网络与遗传算法结合起来对训练样本进行遴选处理。模拟结果表明, 所建模型具有较高的认知、泛化能力。

关键词 近岸海域; 生态环境模拟; 水信息学; 支持向量机; 遗传算法; 神经网络

中图分类号 X171; TP391

文献标识码 A

文章编号 1006-764X(2011)S1-0099-04

近岸海域是海洋和陆地的交汇处, 这里的海洋动力条件在潮流、波浪和地形等作用下演变复杂, 由于海洋水动力条件及化学和生物过程的综合作用, 使得近岸海域生态环境非常复杂。生态水动力学模型是模拟和预测近岸海域生态系统变化的常用方法, 建立一个可靠的基于过程的生态水动力学模型需要对系统内部过程的透彻认识。由于海洋生态系统的高度复杂性和非线性, 系统内各因素之间的相互作用及其动态变化过程还未被完全知晓; 另一方面, 这种模型通常建立在空间一致性的假定之上, 很少考虑个体差异、空间特异性和局部的相互作用, 这些都制约了模型的发展。

随着包括遥感在内的现代化监测技术的发展和监测手段的多样化, 空间数据的数量快速增长、覆盖范围增大、复杂性随之增强。生态水动力学模型可利用的信息越来越多, 如何有效地利用和处理海量数据成为生态环境模拟中的重要问题。从信息方法的角度出发, 研究水信息的输入、存储、处理、反馈和输出成为发展趋势, 水信息学(Hydroinformatics)应运而生。水信息学是传统水科学与信息技术相结合的产物^[1], 它是采用计算机技术、通信网络技术、3S 技术、可视化技术等一系列高新技术, 并结合传统水科学和工程学的方法, 研究与水环境相关数据的收集、存储、处理、模拟、预测和结果显示等问题, 来揭示复杂水环境系统的规律, 解决水环境科技难题^[2-3]。

水信息学于 20 世纪 90 年代初在荷兰 Abbott 教授^[4]的倡导下正式形成, 此后水信息学得到了快速发展^[5-6]。水信息学在水环境中应用的 2 条主线是“知识发现”和“知识管理”。用于“知识发现”的方法方式包括数学模型、数据挖掘、数据同化和数据驱动等; 用于“知识管理”的方法方式包括地理信息系统、决策支持系统和商业软件等。水信息学在近岸海域生态环境模拟中已有广泛的应用。美国 Chesapeake 湾^[7]、日本 Osaka 湾^[8]、澳大利亚 Port Phillip 湾^[9]、香港 Tolo 港^[10]等生态环境问题突出的海域也都建立了各自的近岸海域生态环境模拟模型, Chen 等^[11]运用模糊数学模拟藻类的生长取得了很好的效果; Hugh 等^[12]和 Barciela 等^[13]将人工神经网络应用于赤潮的预测计算之中, 穆迪等^[14]利用 ArcGIS 系列软件对渤海湾监测数据进行了空间分布特性和时间变化趋势的分析, 并对水体营养状况进行评价。

但总的来说, 水信息学在近岸海域生态模型中的应用还处于初级阶段, 以单一的技术为主, 各方法交叉结合的研究还比较少, 未形成系统的方法。笔者以渤海湾生态环境模拟为背景, 重点讨论遗传算法优化支持向量机和模糊模式优化遗传神经网络在近岸海域生态环境模拟中的应用。

1 遗传算法优化支持向量机

在统计学理论基础之上发展起来的支持向量机

基金项目: 国家自然科学基金(10872144); 教育部博士点基金(20090032110007)

作者简介: 向先全(1984—), 男, 安徽庐江人, 博士研究生, 从事水信息学及水生态环境模拟研究。E-mail: xqxuan@tju.edu.cn

通讯作者: 陶建华。E-mail: jhtao@tju.edu.cn

水利水电科技进展, 2011, 31(S1): Tel: 025-83786335 E-mail: jz@hhu.edu.cn http://kbb.hhu.edu.cn

· 99 ·

(support vector machine, SVM) 算法具有严格的理论基础, 采用结构风险最小化准则(而非传统的经验风险最小化原则) 把学习问题转化为一个二次规划问题来获得全局最优解, 在小样本情况下具有良好的外推能力。支持向量机的基本思想可概括为: 首先通过非线性变换将输入空间变换到一个高维空间, 然后在这个新空间中求取最优线性分类面, 这种非线性变换是通过定义适当的内积函数实现的。

目前支持向量机已经发展了很多种, 如 C-SVC, ν -SVC, ϵ -SVR, ν -SVR 等, 生态环境模拟模型主要是基于支持向量回归方法进行的, 采用的是 ϵ -SVR, 下面就具体介绍 ϵ -SVR 的算法原理。

设样本集 (x_i, y_i) , x_i 为样本输入, y_i 为样本输出, $i = 1, \dots, m$ 。函数逼近问题就是要找出一个函数 $f(x)$, 使之通过样本训练后尽量逼近 y_i , 并具有一定的推广能力。首先考虑线性回归函数 $f(x) = \omega \cdot x + b$, 根据结构风险最小化原理, 函数估计问题就是寻找 $f(x)$ 使得风险函数 R_{reg} 最小。

$$R_{\text{reg}} = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m L_{\epsilon}(y_i - f(x_i)) \quad (1)$$

式中: $\frac{1}{2} \|\omega\|^2$ 反映了模型的复杂度; $\sum_{i=1}^m L_{\epsilon}(y_i - f(x_i))$ 反映了训练误差, 即经验风险; C 为设定的误差惩罚参数, 用于控制对错分样本惩罚的程度, 实现错分样本数和模型复杂度之间的折中。

将风险最小化问题转化为求解约束优化问题, 并采用对偶理论和 Lagrange 乘子法进行变换, 该优化问题的对偶形式为

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i,j=1}^m (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) + \\ & \epsilon \sum_{i=1}^m (\alpha_i^* + \alpha_i) - \sum_{i=1}^m y_i (\alpha_i^* - \alpha_i) \\ \text{s.t.} \quad & \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0 \\ & 0 \leq \alpha_i, \alpha_i^* \leq C \end{aligned} \quad (2)$$

利用最优化理论中的二次型规划方法求解可得拉格朗日乘子 α 和 α^* , 它们中间只有一小部分不为零, 其对应的样本就是支持向量。再根据凸二次规划的 KKT 条件求得偏置 b 。由此可得支持向量回归机的决策函数

$$f(x) = \sum_{i=1}^m (\alpha_i^* - \alpha_i)(x_i \cdot x) + b \quad (3)$$

当样本集为非线性可分问题时, 引入核函数将输入样本通过非线性函数 ϕ 映射到一个高维的 Hilbert 空间, 在该空间中构造并求解最优分类面。所谓核函数就是在输入空间中存在一个函数

$K(x, x')$, 满足 $K(x, x') = (\phi(x) \cdot \phi(x'))$ 。由统计学习理论可知, 只要任意函数满足 Mercer 条件, 就可以作为核函数。应用核函数可以避免直接计算 $(\phi(x) \cdot \phi(x'))$, 从而解决了由于模式升维而引起的维数灾难问题, 使得在没有增加复杂度的情况下, 智能学习在高维特征空间成为可能。目前常见的核函数有以下几种: 多项式核 $K(x, x_i) = (\gamma x \cdot x_i + \gamma_0)^d$; 径向基核 $K(x, x_i) = \exp(-\gamma \|x - x_i\|^2)$; Sigmoid 核 $K(x, x_i) = \tanh(\gamma x \cdot x_i + \gamma_0)$ 。

由于有多种核函数, 设计 SVM 的一个重要步骤就是选择核函数和核参数。RBF 核函数是一个普适的核函数, 而且只有 1 个可控参数, 通过参数的选择, 它可以适用于任意分布的样本且输出权值由算法自动确定。同时考虑到富营养化的高度非线性和复杂性, 笔者选择了径向基核函数, 此时, SVM 回归方法中的参数主要有不灵敏参数 ϵ 、惩罚参数 C 和径向基核参数 γ 。遗传算法是一种全局搜索算法, 具有很强的鲁棒性, 且不依赖于特定的求解模型, 故笔者采用遗传算法对支持向量机的参数进行智能选取。利用遗传算法求解问题, 首先必须对每个可行解进行编码, 从而将解空间变换为染色体空间, 并定义染色体的适应度, 使得较优的个体适应度也较高。然后再对种群施加遗传算子如选择、交叉、变异等操作, 使群体不断向着最优解的方向进化。

渤海湾位于渤海西侧, 为陆地环抱的浅海盆, 海水交换能力和自净能力很弱, 近岸海域水体的富营养化严重, 赤潮频繁发生。选定渤海湾近岸的一个监控区作为研究对象。在监控区 $N38^{\circ}49'20'' \sim 38^{\circ}55'00''$, $E117^{\circ}37'00'' \sim 117^{\circ}50'00''$ 的范围内共设 6 个站位。2006 和 2007 年 6—10 月期间, 每 2 周对监控区的表层海水进行 1 次水质监测, 监测指标有: 水温 (T)、盐度 (S)、透明度 (SD)、pH、化学耗氧量质量浓度 ($\rho(\text{COD})$)、溶解氧质量浓度 ($\rho(\text{DO})$)、溶解氧饱和度 (ρ'_{DO})、活性硅酸盐质量浓度 ($\rho(\text{Si})$)、磷酸盐质量浓度 ($\rho(\text{P})$)、硝基氮质量浓度 ($\rho(\text{NO}_3\text{-N})$)、亚硝基氮质量浓度 ($\rho(\text{NO}_2\text{-N})$)、氨氮质量浓度 ($\rho(\text{NH}_4\text{-N})$) 以及叶绿素 a 的质量浓度 ($\rho(\text{Chl-a})$) 等, 共 13 个指标。

分析叶绿素 a 的动态含量, 可以快速简便地了解浮游植物生物量状况及其变化趋势, 在一定程度上反映水体富营养化的状况。利用 ϵ -SVR 算法建立以下模型:

$$\rho(\text{Chl-a}) = f(T, SD, \text{pH}, S, \rho(\text{COD}), \rho(\text{DO}), \rho'_{\text{DO}}, \rho(\text{Si}), \rho(\text{P}), \rho(\text{NO}_2\text{-N}), \rho(\text{NO}_3\text{-N}), \rho(\text{NH}_4\text{-N})) \quad (4)$$

具体操作步骤为: ①对数据进行平滑处理, 以剔除奇异值; ②为消除各属性数据的单位差异及数值量纲上的差异, 对原始数据进行归一化处理; ③将数

据集分为训练样本集和测试样本集,前者用于模型的建立,后者用于模型的验证。④采用训练样本集建模数据,利用遗传算法对 SVM 参数寻优,包括编码、生成初始种群、确定适应度函数以及相应的遗传操作(选择、交叉、变异)。⑤根据优化过的参数,建立 SVM 生态环境模型,并用测试样本对模型进行检测。

文献[15]通过对支持向量机各参数响应的研究,利用自适应的遗传算法对 C 和 γ 进行启发式寻优,GA 适应度函数除考虑 SVM 的预测精度,也考虑了 SVM 最优分类面的结构简单性。以渤海湾实测资料为依据,选 2 号站位实测数据作为测试样本集,其余 5 个站位实测数据作为训练样本集。SVM 对训练样本集进行模拟,叶绿素 a 质量浓度模拟值与实测值的均方误差为 $1.668 \mu\text{g/L}$;用 2 号站位实测数据对模型进行检测,叶绿素 a 质量浓度模拟的均方误差为 $1.831 \mu\text{g/L}$,表明所建立的 SVM 生态模型具有较高的认知、泛化能力。

2 模糊模式优化遗传神经网络

人工神经网络是模仿人的大脑神经元结构、特性和大脑认知功能构成的信息处理系统,具有学习、联想、容错、并行处理等能力,容易建立模型且能够快速反应,适用于实时监测和预报,尤其适用于机制尚不清楚的高维非线性系统的模拟。神经网络中应用最广泛的是 BP 网络,它是一种反向传递并且能够修正误差的多层映射函数。但由于 BP 网络是在梯度法的基础上推导出来,要求目标函数连续可导,在进化学习过程中收敛速度慢,容易陷入局部最优,并且由于 BP 网络的初始权值和阈值是随机选择的,每次的初始值都不一样,造成每次训练学习预测的结果都有所差别。将 BP 神经网络和遗传算法有机融合,充分发挥遗传算法的全局搜索能力和 BP 神经网络局部搜索能力,利用遗传算法来弥补权值和阈值选择上的随机性缺陷,可得到更好的预测结果。

为了更好地对测试集进行预测,利用测试样本对训练样本进行模式识别,从而遴选出和测试样本较匹配的训练样本输入到 BP 网络中进行训练。利用模糊模式识别进行样本的遴选,通过贴近期来计算匹配程度。采用高斯型隶属度函数将普通集模糊化,即通过隶属函数将普通集合映射到 $[0, 1]$,两样本越靠近,其隶属度值越接近于 1。

模糊贴近期有多种算式,如海明贴近期法、欧几里得贴近期法、测度贴近期法等。笔者采用海明贴近期法。可以得出样本 A 与测试样本中其他样本的贴近期。根据模糊模式识别的“择近原则”,取样本 A 与测试样本集中最大的贴近期记为样本 A 对

测试样本的贴近期。对于每一个训练样本都可以计算其对测试样本的贴近期,根据具体要求,剔除训练样本对测试样本靠近程度较小者,从而达到对训练样本遴选的目的。

以渤海湾实测数据为例,研究遗传神经网络对叶绿素 a 的预测模型。确定 t 时刻叶绿素 a 的质量浓度作为预测模型的输出,用 $t-1$ 时刻的各监测指标值来预测 t 时刻的叶绿素 a 的质量浓度,时间步长为 2 周。建立的模型如下:

$$\begin{aligned} \alpha(\text{Chl-a})_t = & g(T_{t-1}, SD_{t-1}, pH_{t-1}, S_{t-1}, \alpha(\text{COD})_{t-1}, \\ & \alpha(\text{DO})_{t-1}, \alpha'_{\text{DO}_{t-1}}, \alpha(\text{Si})_{t-1}, \alpha(\text{P}_{t-1}), \alpha(\text{NO}_2\text{-N})_{t-1}, \\ & \alpha(\text{NO}_3\text{-N})_{t-1}, \alpha(\text{NH}_4\text{-N})_{t-1}, \alpha(\text{Chl-a})_{t-1}) \end{aligned} \quad (5)$$

文献[16]利用模糊模式识别对训练样本进行遴选处理。以渤海湾实测数据为依据,将 6 个站位按时间序列组合为样本集,取前 18 个时间序列的实测值为训练样本($18 \times 6 = 108$ 个),第 19 个时间序列的 6 个站位实测值为测试样本。以测试样本作为既定的模式,计算训练样本相对于测试样本的贴近期,选择贴近期较大的样本作为模型的输入。选取贴近期大于 0.8 的训练样本作为匹配样本,输入到 BP 网络中进行模型的训练。通过模糊模式识别,从 108 个训练样本中选取 86 个训练样本作为模型的输入,约占总样本的 80%。测试样本的预测结果和实测值的均方根误差为 $1.45 \mu\text{g/L}$,与未经模式识别遴选处理时模型预测的均方根误差为 $3.81 \mu\text{g/L}$ 相比,改进的效果较明显。

3 结 语

水信息学是一门新兴学科,它的许多概念体系和内容范畴至今还处于模糊阶段。由于生态环境的高度复杂性和非线性,水信息学在生态环境中的研究还存在一些不足,可从以下方面进一步完善:①基于数据的各方法手段的交叉。基于数据的方法包括数据挖掘、数据同化、数据驱动等,这些方法在外延上有一定的重合。对于水生态系统,当获得大量监测数据时,往往可以先采用基于数据的模型进行分析,以便对生态环境系统得到一些初步的认识。②基于数据和基于过程的结合。单纯用基于数据的模型很难反映出对象本身的特性,探索系统的内部机理和演变机理在水环境研究中不可或缺。可根据实际需要两者结合,发挥各自优势。③生态环境知识发现和知识管理的结合。知识发现将信息变为知识,同时知识管理在解决实际问题时必不可少。如将数学模型、数据挖掘等技术与 GIS 耦合,对生态环境的数据进行存储、分析、模拟和结果显示,并结合决策支持系统为决策者提供参考。

参考文献：

[1] ABBOTT M B. Introducing hydroinformatics[J]. Journal of Hydroinformatics ,1999 ,1(1) 3-19.

[2] 向先全 ,陶建华.水信息学及其在水环境中的应用研究综述[J].生态环境学报 ,2009 ,18(4) :1587-1593.

[3] 陶建华.计算水力学和水力信息学[C]//1994 年全国水动力学研讨会文集.北京 :海洋出版社 ,1994 :803-807.

[4] ABBOTT M B. Hydroinformatics :information technology and the aquatic environment[M]. Aldershot :Avebury Technical , 1991.

[5] ABBOTT M B. Towards the hydraulics of hydroinformatics era [J]. Journal of Hydraulic Research ,2001(4) 339-349.

[6] 顾正华 ,唐洪武 ,李云 ,等.水信息学与智能水力学[J].河海大学学报 :自然科学版 ,2003 ,33(5) 518-521.

[7] XU J , HOOD R. Modeling biogeochemical cycles in Chesapeake Bay with a coupled physical-biological model[J]. Estuarine ,Coastal and Shelf Science ,2006 ,69(1/2) :19-46.

[8] YAMANE N , NAKATSUJI K. Field data collection and modeling for verification of an ecosystem model in Osaka Bay [J]. Estuarine and Coastal Modeling ,1997 ,4 :196-210.

[9] MURRAY A G ,PARSLOW J S. Modelling of nutrient impacts in Port Phillip Bay :a semi-enclosed marine Australian ecosystem[J]. Marine and Freshwater Research ,1999 ,50(6) : 597-611.

[10] CHAUK W. A three-dimensional eutrophication modeling in Tolo Harbour[J]. Applied Mathematical Modelling ,2004 ,28 (9) 849-861.

[11] CHEN Q ,MYNETT A E. Integration of data mining techniques with heuristic knowledge in a fuzzy logic modelling of eutrophication in Taihu Lake[J]. Ecological Modelling ,2003 , 162 :55-67.

[12] HUGH W ,FRIEDRICH R. Towards a genetic artificial neural network model for dynamic prediction of algal abundance in freshwater lakes[J]. Ecological Modelling ,2001 ,146 :69-84.

[13] BARCIELA R M ,GARCIA E ,FERNANDEZ E. Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models an artificial neural networks[J]. Ecological Modelling ,1999 ,120 :199-211.

[14] 穆迪 ,袁德奎 ,陶建华.地理信息系统在渤海湾水体富营养化研究中的应用[J].海洋技术 ,2006 ,25(1) :73-77.

[15] 向先全 ,陶建华.基于 GA-SVM 的渤海湾富营养化模型 [J].天津大学学报 ,2011 ,44(3) 215-220.

[16] 向先全 ,陶建华.基于模糊识别和遗传神经网络的渤海湾叶绿素 a 预测研究[J].海洋环境科学 ,2011 ,30(2) : 239-242.

[17] 穆迪 ,袁德奎 ,陶建华.地理信息系统在渤海湾水体富营养化研究中的应用[J].海洋技术 ,2006 ,25(1) :73-77.

[18] 向先全 ,陶建华.基于 GA-SVM 的渤海湾富营养化模型 [J].天津大学学报 ,2011 ,44(3) 215-220.

[19] CERNETIC J. The use of noise and vibration signals for detecting cavitation in kinetic pumps [J]. Mechanical Engineering Science ,2009 ,223 :1645-1655.

[20] WANG G ,OSTOJA-STARZEWSKI M. Large eddy simulation of a sheet/cloud cavitation on a NACA0015 hydrofoil[J]. Applied Mathematical Modelling ,2007 ,31(3) 417-447.

[21] OLIVIER-DELGOSHA ,FRNCOIS D ,JACQUES A A ,et al. Numerical prediction of cavitating flow on a Two-Dimensional symmetrical hydrofoil and comparison to experiments[J]. Journal of Fluids Engineering ,2007(129) 279-292.

[22] NECKER J ,ASCHEBRENNER T. Model test and CFD calculation of a cavitating bulb turbine[C]//25th IAHR Symposium on Hydraulic Machinery and Systems ,Timisoara , Romania :Earth Environmental science ,2010.

[23] LIU Shu-hong ,LI Sheng-cai ,ZHANG Liang ,et al. A mixture model with modified mass transfer expression for cavitating turbulent flow simulation[J]. Engineering Computations : International Journal for Computer-Aided Engineering and Software ,2008 ,25(4) 290-304.

[24] SRINIVASAN V ,SALAZAR A J ,SAITO K. Numerical simulation of cavitation dynamics using a cavitation-induced-momentum-defect (CIMD) correction approach[J]. Applied Mathematical Modeling ,2009 ,33 :1529-1559.

[25] INANC S ,WEI S. Interfacial dynamics-based modelling of turbulent cavitating flows ,part2 :time-dependent computations [J]. International Journal for Numerical Methods in Fluids ,

2004 ,44 :997-1016.

[26] INANC S ,WEI S. Interfacial dynamics-based modelling of turbulent cavitating flows ,part1 :Model development and steady-state computations [J]. International Journal for Numerical Methods in Fluids ,2004 ,44 :975-995.

[27] WU Lei ,LU Chuan-jing ,LI Jie ,et al. Numerical simulations of 2D periodic unsteady cavitating flows [J]. Journal of Hydrodynamics ,2006 ,18(3) 341-344.

[28] TAMURA Y. Improvement of bubble model for cavitating flow simulations[J]. Journal of Hydrodynamic ,2009 ,21(1) :41-46.

[29] QIN Z ,BREMHORST K ,ALEHOSSEIN H ,et al. Simulation of cavitation bubbles in a convergent-divergent nozzle water jet [J]. Journal of Fluid Mechanics ,2007 ,573 :1-25.

[30] NOURI N M ,MOGHIMI M ,MIRSAEEDI S M H. Unsteady modelling of cavitating flow with artificial viscosity[J]. Mechanical Engineering Science ,2009 ,223 :1-10.

[31] ZHU B ,WANG H. Numerical prediction of impact force in cavitating flows[C]//25th IAHR Symposium on Hydraulic Machinery and Systems ,Timisoara ,Romania :Journal of Fluids Engineering Transactions of the ASME ,2010.

[32] 徐宇 ,吴玉林 ,唐学林.用动理学方法推导空化流的控制方程[J].清华大学学报 ,2003 ,43(10) :1416-1419.

[33] 张新明 ,周超英 ,SLAM SHAMS ,等.用格子 Boltzmann 方法数值模拟三维空化现象[J].物理学报 ,2009 (12) : 8406-8414.

(收稿日期 2010-09-25 编辑 :方宇彤)