支持向量机回归的碳通量预测

陈 强,吴慕春,薛月菊,杨敬锋,刘国瑛
CHEN Qiang,WU Mu-chun,XUE Yue-ju,YANG Jing-feng,LIU Guo-ying

华南农业大学 工程学院,广州 510642
College of Engineering,South China Agricultural University,Guangzhou 510642,China
E-mail:mcwui66@scau.edu.cn

CHEN Qiang,WU Mu-chun,XUE Yue-ju,et al.Research of predicting methods for carbon flux based on support vector regression.Computer Engineering and Applications,2009,45(21):235-238.

Abstract: Precisely predicting the carbon flux through impact factors has attracted many ecologists' interest.However,there is still no perfect method to predict carbon flux effectively.In this paper,ε-support vector regression(ε-SVR) is used to predict carbon flux,and the results of ε-SVR and BP neural network(BPNN) for the prediction of carbon flux are compared.ε-SVR with different kernel functions and parameters and BPNN with different numbers of the neurons in hidden layer are analyzed.The experiment results show that the correlation between the carbon flux predicted by ε-SVR and BPNN and the observation values is high.However,ε-SVR converges global optimal more easily than BPNN.And the ε-SVR predicts more accurately than BPNN.
Key words: ε-support vector regression;Back Propagation(BP) neural network;carbon flux;predicting

摘 要:如何根据影响因素较好地预测碳通量是许多环境监测者非常关注的问题。但至今尚无一种非常有效的预测模型,为此研究 ε-支持向量回归机在碳通量预测中的具体应用,并与 BP 神经网络模型的预测结果做了比较,分析了两种方法在核函数及相关参数、网络结构、神经元数目选择方面各自不同的特点。实验结果表明,基于 ε-支持向量回归机和 BP 神经网络模型的碳通量预测结果与碳通量实测值之间存在显著相关性。但 ε-支持向量回归机方法的预测过程更易掌控,整体预测精度高于 BP 神经网络的精度。
关键词:ε-支持向量回归;反向传播神经网络;碳通量;预测精度
DOI:10.3778/j.issn.1002-8331.2009.21.068 **文章编号:**1002-8331(2009)21-0235-04 **文献标识码:**A **中图分类号:**TP391

近年来,大气中 CO_2 等温室气体浓度不断升高,引起了全球气候变暖等一系列环境问题,严重威胁到人类的可持续发展[1-2]。全球碳通量观测网络(http://www.fluxnet.ornl.gov/fluxnet/index.cfm)作为获取生态系统与大气之间的 CO_2 交换信息的有效手段,积累了大量的相关数据,为全球碳循环的研究提供了重要的数据基础[3]。但这种只通过硬件手段观测的碳通量数据,不但仪器设备造价昂贵,而且观测过程容易受环境、仪器故障等因素影响。利用现代科学的方法对以往观测数据进行研究,通过发现影响碳通量变化的主要因素, 寻找好的碳通量预测方法,是碳通量研究的重要手段之一。

由于碳通量与影响碳通量的各主要因素间更多的是存在一种非线性关系,传统的线性模型难以精确地预测碳通量。与回归分析相比,前馈神经网络不需要先验假设,理论上已证明在选择适当的隐含层及相应的神经元数目下,可以任意精度逼近任意非线性函数,通过学习可以获得输入与输出之间的最优关系,建立更为精确的碳通量预测模型。如 M.T.Van Wijk 等[4]和 Assefa M.Melesse 等[5]从观测数据本身出发,利用神经网络的方法选择了最小的输入因素集,并建立了该方法的碳通量预测模型,得到了理想的预测效果。Makoto Ooba 等[6]采用遗传算法改进了一般的神经网络模型,发现改进后的模型对净生态系统 CO_2 交换(Net CO_2 Ecosystem Exchange,NEE)的预测效果要明显好于一般神经网络模型和非线性回归模型。

支持向量回归[7]作为支持向量机的一种,是基于统计学理论[8]发展而来的一种核方法(Kernel Method,KM[9])。建立在结构风险最小化优化上的 SVR 使得神经网络应用中结构选择的问题,在 SVR 的应用中成为相对容易的核函数选择问题。核函数实现了数据空间与特征空间之间的非线性映射,有效地将数据空间中的非线性操作转变为特征空间中的线性操作,大大提高

基金项目:国家科技攻关计划项目(the Key Technologies R&D Program of China under Grant No.2002BA516A08);国家星火计划项目(No.2006EA780057);广东省自然科学基金(the Natural Science Foundation of Guangdong Province of China under Grant No.04300504,No.05006623);广东省科技攻关计划(the Key Technologies R&D Program of Guangdong Province,China under Grant No.2005B20701008,No.2005B10101028,No.2004B20701006)。
作者简介:陈强(1983-),男,硕士,主要研究方向为数据挖掘技术的应用;通讯作者:吴慕春(1965-),女,讲师,主要研究方向为机电一体化及数据挖掘;薛月菊(1969-),女,博士,副教授,主要研究方向为智能控制、数据挖掘及地理信息系统应用。
收稿日期:2008-10-08 **修回日期:**2008-12-24

了非线性处理能力，在非线性回归中具备非常优秀的性能，且已有不少实际应用，如进行土壤水利学参数预测[10]、岩爆预测[11]和非线性时间序列预测[12]等。但目前在国内外，运用支持向量回归的方法来提高碳通量预测精度的研究成果还极为少见，因此该文研究了其在这方面的具体应用，并与神经网络的方法进行了详细比较。

1 ε-SVR 算法

支持向量机的原问题是凸二次优化问题，其转换后的有更简单变量约束的对偶问题同样是凸二次优化问题，保证找到的解为全局最优解，能够很好地解决小样本、高维、非线性等实际问题。由于函数的求解只涉及到样本之间的内积运算$(x_i \cdot x_j)$，高维空间中的内积运算可以通过原空间核函数来实现，所以不需要知道非线性映射的显式表达式，也几乎不增加计算的复杂性。ε-SVR 算法如下[13]：

(1)设已知训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(\chi\times\gamma)^l$，其中 $x_i\in\chi=R^n$，$y_i\in\gamma=R$，$i=1,\cdots l$；

(2)选择适当的核函数 $K(x_i,x_j)$，选择适当的 ε 和 C；

(3)构造并求解原最优化问题的对偶问题

$$\min_{\alpha^{(*)}\in R^{2l}}\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x')+\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i)$$

$$\text{s.t.}\ \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \qquad (1)$$

$$0\leqslant\alpha_i,\alpha_i^*\leqslant\frac{C}{l},i=1,2,\cdots,l$$

得到最优解$\bar{\alpha}=(\bar{\alpha}_1,\bar{\alpha}_1^*,\cdots,\bar{\alpha}_l,\bar{\alpha}_l^*)^{\mathrm{T}}$。

(4)构造决策函数

$$f(x)=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)K(x_i,x)+\bar{b} \qquad (2)$$

其中$\bar{b}$按下列方式计算：选择位于开区间$(0,C/l)$中的$\bar{\alpha}_j$或$\bar{\alpha}_k^*$，若选到的是$\bar{\alpha}_j$，则

$$\bar{b}=y_i-\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(x_i\cdot x_j)+\varepsilon \qquad (3)$$

若选到的是$\bar{\alpha}_k^*$，则

$$\bar{b}=y_k-\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(x_i\cdot x_k)+\varepsilon \qquad (4)$$

SVM 性能的好坏取决于核函数及其参数的选取，不同的核函数会导致 SVM 的推广性能有所不同，如何根据具体的数据选择恰当的核函数是 SVM 应用领域遇到的一个重大难题，对核函数的研究还未能深入到足以指导我们如何选取核函数，目前应用较多的核函数有如下几种：多项式(Polynomial)核函数为 $K(x,y)=[\kappa(x\cdot y)+c]^d$ $(d=1,2,\cdots)$；径向基(Radial Basic Function，RBF)核函数为 $K(x,y)=\exp(-\kappa\|x-y\|^2)$；Sigmoid 核函数为 $K(x,y)=\tanh(\kappa(x\cdot y)+c)$，式中的 d,κ,c 为待定参数。

2 实验数据

该研究数据来自于美国北卡罗来纳州布莱克伍德区 Duke 森林的 CO_2 流量测量塔(35.58′N，79.8′W)。该地区在 1983 年将原有森林砍伐后，重新种植火炬松林，火炬松的间距为 2.4×2.4 m。碳流量的测量始于 1996 年。采用以每 30 分钟采集一次碳的净生态系统交换量的数据。每一个净生态系统交换量的数据记录包含风速(WS)和空气温度(AT)等 90 个因素，去除时间和地点等因素，共有 67 个测量的因素[14]。参考相关文献并依据本研究的试验，选取其中对碳通量输出影响比较重要的 10 个因素作为输入变量，它们分别是空气温度(TA)、射入辐射(RN)、摩擦速度(UST)、绝对湿度(H)、潜热通量(LE)、光合有效辐射(PAR)、相对湿度(RH)、水气压差(VPD)、土壤水分含量(SWC)和土壤温度(TS)。测量中，由于设备故障引起了严重的数据缺失。由于重点研究各因素之间的关系，以及各因素与 CO_2 流量的关系，所以仅用这些数据中完整的 10 800 条数据记录来验证所提方法的有效性。表 1 为输入及输出数据值的变化范围。

表 1 碳通量以及各主要因素值的变化范围

变量(英文简写)	最大值	最小值	平均值	标准差
空气温度(TA)	38.211	−11.991	18.87	8.47
射入辐射(RN)	789.744	−135.835	118.16	217.20
摩擦速度(UST)	1.293	0	0.27	0.20
绝对湿度(H)	414.15	−143.412	28.48	71.30
潜热通量(LE)	587.302	−73.953	54.68	95.12
光合有效辐射(PAR)	2 760	−10	395.63	560.54
相对湿度(RH)	100	14	72.38	22.52
水气压差(VPD)	4.564	0	0.74	0.83
土壤水分含量(SWC)	70.3	13.9	39.44	14.34
土壤温度(TS)	23.61	3.6	16.16	4.46
碳通量(FC)	36.432	−42.455	−3.07	7.44

由表 1 可以看出，影响碳通量输出的输入因素值存在如下特点：因素内和因素间的数值变化范围均较大，为了消除由于量纲和单位不同的影响，有必要在建立预测模型之前，对数据进行归一化预处理，归一化公式如式(5)：

$$z_i=\frac{2(x_i-x_{\min})}{(x_{\max}-x_{\min})}-1 \qquad (5)$$

式中 x_i 与 z_i 分别为归一化前后的变量，$x_{\max}$ 与 $x_{\min}$ 分别为 x_i 的最大值和最小值。

3 实验结果与分析

实验中随机地抽取 10 800 条样本总数的 80%做为网络的训练样本，剩余的 20%用来做网络的测试样本。将数据按式(5)进行归一后输入网络。ε-SVR 和 BP 神经网络模型采用上述 10 个变量作为网络的输入，碳通量为输出。

3.1 ε-SVR 的预测结果分析

ε-SVR 性能的好坏关键取决于核函数及相关参数的选择。选择目前使用较为广泛的多项式核函数和 RBF 核函数，对于不同类型的核函数，产生的支持向量的个数变化不大，但核函数的相关参数却对模型的预测性能有重要影响。其中惩罚因子 c 用于控制模型复杂度和函数逼近误差的折中。c 越小，对错分样本的惩罚越小，那么样本的训练误差就越大，使得结构风险也变大；而 c 越大，惩罚就越大，对错分样本的约束程度就越大，从而使得第二项置信范围的权重变大，那么分类间隔的权重就相对变小，导致模型的泛化能力就变差。对多项式核函数来说，多项式次数 d 的取值对模型预结果有较大影响；对

RBF 核函数来说,核参数 κ 对模型的预测结果有较大影响。实验证明,对于用于控制支持向量个数和泛化能力的损失函数参数 ε,其取值在 0.000 1~0.1 范围内时对模型的预测结果无明显影响[15]。表 2 和表 5 为相同训练样本和测试样本下不同模型参数对预测效果的比较。

表 2　惩罚因子 c 对预测效果的影响

惩罚因子 c	多项式核函数次数 d	训练样本与实测结果间的相关性	测试样本与实测结果间的相关性
1	2	0.878	0.873
5	2	0.894	0.887
10	2	0.897	0.888
15	2	0.898	0.887

表 3　多项式核函数次数对预测效果的影响

惩罚因子 c	多项式核函数次数 d	训练样本与实测结果间的相关性	测试样本与实测结果间的相关性
10	1	0.866	0.861
10	2	0.897	0.888
10	3	0.884	0.870
10	4	0.848	0.828

表 4　惩罚因子 c 对预测效果的影响

惩罚因子 c	RBF 核函数参数 κ	训练样本与实测结果间的相关性	测试样本与实测结果间的相关性
1	0.1	0.898	0.891
5	0.1	0.902	0.893
10	0.1	0.903	0.893
20	0.1	0.904	0.892

表 5　RBF 核函数参数 κ 对预测效果的影响

惩罚因子 c	RBF 核函数参数 κ	训练样本与实测结果间的相关性	测试样本与实测结果间的相关性
10	0.01	0.888	0.883
10	0.05	0.900	0.892
10	0.10	0.903	0.893
10	0.20	0.906	0.891

由表 3 的实验结果可知在多项式核函数次数较低时,模型过于简单,不能很好拟合较复杂的非线性训练样本;但次数过高,则会使得函数集的 VC 维升高,从而提高学习机器的复杂性,模型的推广性能将下降,易出现“过拟合”现象。实验进一步验证了对于一般非线性拟合问题 d 的取值不宜超过 3[16]。由表 5 实验结果可知随着 κ 的不断增大,训练样本的预测性能不断提高,但上升到一定值后同样出现“过拟合”现象,样本的推广性能下降,其取值范围一般为(0.1~3.8)[15]。所以,对选取的核函数为多项式函数,惩罚因子 c 取值为 10,多项式次数为 2 时,预测结果最为理想,与实测结果之间的相关系数为 0.888;对选取的核函数为 RBF,惩罚因子取值为 10,核函数参数 κ 为 0.1 时,预测结果最理想,与实测结果之间的相关系数为 0.893。实验结果还表明,选择 RBF 作为 ε-SVR 核函数的碳通量预测效果要略好于多项式核函数。就相关性这一指标的预测结果来看,基于 RBF 核的 ε-SVR 碳通量预测和 BP 神经网络的碳通量预测有同样好的预测效果。

3.2　BP 神经网络预测结果分析

BP 神经网络模型采用三层的网络结构:一个输入层、一个隐含层和一个输出层。其中隐含层神经元数目为 20,输入层到隐含层间的传递函数为 tansig 函数;隐含层到输出层间的传递函数为线性函数。训练中,在大约迭代到 55 步时,网络收敛,如图 1 所示。通过对输出结果的分析,发现网络的预测结果与实测值之间存在明显的相关性,训练样本输出值和实测值间的相关系数为 0.891 4,而测试样本输出值与实测值间的相关系数也达到了 0.893 7,如图 2 所示。

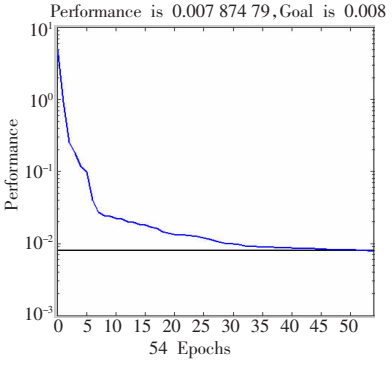

图 1　BP 神经网络的训练步长

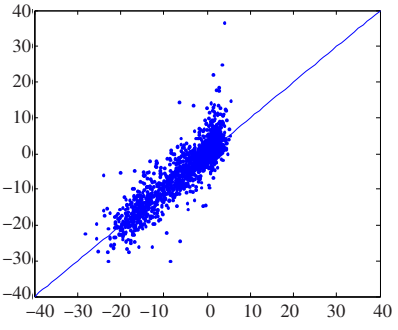

图 2　预测值与实测值之间的相关性

3.3　ε-SVR 与 BP 神经网络预测性能的比较

该节进一步比较研究 ε-SVR 和 BP 神经网络的方法在碳通量预测方面的效果。为了进一步完成两种方法的比较,根据碳通量观测数据的特点,将实测碳通量绝对值大于等于 10 的数据提取出来,利用提取的数据来从预测结果的准确率、相关性和平均相对误差三方面对两种方法进行细致比较。提取的数据样本总数为 2 043 组,依然随机地抽取其中的 80%作为训练样本,剩余的 20%为测试样本。BP 神经网络结构仍为三层网络,其中隐含层神经元数为 20。ε-SVR 的核函数选择 RBF,其中 c、κ 和 ε 的取值分别为 50、0.1 和 0.001。实验结果见表 6。

实验结果表明,虽然样本的预测值与实测值间的相关性仍很接近,但在适当的参数选择下,从准确率和平均相对误差方

表 6　ε-SVR 与 BP 神经网络预测性能的比较

实验方法	预测值与实测值间的相关性(训练样本)	预测值与实测值间的相关性(测试样本)	输出相对误差<20%的测试样本数/(%)	输出相对误差<30%的测试样本数/(%)	输出相对误差<40%的测试样本数/(%)	测试样本平均绝对误差/(%)	测试样本平均相对误差/(%)
BPNN	0.867	0.843	0.657	0.809	0.892	3.05	19.474
SVR	0.876	0.844	0.704	0.841	0.931	3.04	17.242

面来看，基于 RBF 核的 ε-SVR 的预测效果要明显优于 BP 神经网络。

4 结论

（1）在适当选择 ε-SVR 的相关参数和核函数以及 BP 网络结构、隐含层神经元数目情况下，基于 ε-SVR 的方法和基于 BP 神经网络的方法在预测碳通量整体输出中都能获得不错的预测效果，预测值与实测值之间的相关系数相近，都接近 0.9。但从单个样本的准确率和所有样本的平均相对误差方面来看，基于 ε-SVR 的方法要好于 BP 神经网络的方法。

（2）基于 ε-SVR 的方法的最大特点是能够每次得到全局最优解，这样可以通过对核函数及相关参数的调整确保模型朝好的方向发展。且目前其可供选择的核函数有限，核函数的确定相对比较容易。而基于 BP 神经网络的方法，其网络结构和隐含层神经元的数目不易确定，且网络不能保证每次都收敛到最优。

参考文献：

[1] Fang J Y,Chen A P,Peng C H,et al.Changes in forest biomass carbon storage in China between 1949 and 1998 [J].Science,2001,292:2320-2322.

[2] Houghton J T,Ding Y,Griggs D J.IPCC:Climate change 2001:'the scientific basis' contribution of working group 1 to the third assessment report of the intergovernmental panel on climate change[M].New York,NY,USA:Cambridge University Press,2001.

[3] Wylie B K,Fosnight E A,Gilmanov T G,et al.Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains[J].Remote Sensing of Environment,2007,106(4):399-413.

[4] Wijk M T van,Bouten W.Water and carbon fluxes above European coniferous forests modelled with artificial neural networks[J].Ecological Modelling,1999,120(2-3):181-197.

[5] Melesse A M,Hanley R S.Artificial neural network application for multi-ecosystem carbon flux simulation[J].Ecological Modelling,2005,189(3-4):305-314.

[6] Ooba M,Hirano T,Mogami J I,et al.Comparisons of gap-filling methods for carbon flux dataset:A combination of a genetic algorithm and an artificial neural network [J].Ecological Modelling,2006,198(3-4):473-486.

[7] Smola A J,Scholkopf B A.Tutorial on spport vector regression,NeuroCOLT TR NC-TR-98-030[R].Royal Holloway College University of London,UK,1998.

[8] Vapnik V.The nature of statistical learning Theory[M].NewYork:SpringerVerlag,1995.

[9] Muller K R,Mika S,Ratsch G,et al.An introduction to kernel-based learning algorithms[J].IEEE Trans on Neural Networks,2001,12(2):181-201.

[10] 杨绍锷，黄元仿.基于支持向量机的土壤水利学参数预测[J].农业工程学报，2007，23(7)：42-47.

[11] 祝云华，刘新荣，周军平.基于 v-SVR 算法的岩爆预测分析[J].煤炭学报，2008，33(3)：278-281.

[12] Lau K W,Wu Q H.Local prediction of non-linear time series using support vector regression[M].Pattern Recognition,2008,41(5):1539-1547.

[13] 邓乃扬，田英杰.数据挖掘中的新方法—支持向量机[M].北京：科学出版社，2004：235-244.

[14] Zhang Y,Sha L,Yu G,et al.Annual variation of carbon flux and impact factors in the tropical seasonal rain forest of Xishuangbanna,SW China[J].Science in China Ser D,2006,49(Supp II):150-162.

[15] 杨虞微，左洪福，陈果.支持向量机时间序列预测模型的参数影响分析与自适应优化[J].航空动力学报，2006，21(4)：768-770.

[16] 王炜，郭小明，王淑艳，等.关于核函数选取的方法[J].辽宁师范大学学报：自然科学版，2008，31(1)：2-4.

（上接 214 页）

分类进行研究。因为 Hilbert-Huang 变换所进行的 EMD 总是从信号本身出发，基于原信号的局部性特征进行层层“筛选”，因此具有良好的自适应性；另外 Hilbert 变换中，通过与 $1/t$ 的卷积使得结果相当的局部化，无论在时间还是频率上都能取得较高的分辨率。实验表明通过 HHT 提取的目标注视任务下的脑电特征，能够成为模式分类的可靠依据。可以设想通过 BCI 的一些辅助设备，可将注视目标的自动识别方法应用到更多的领域，比如意念控制多个电器开关、思维拨号或者键盘控制中，将有可能为肢体残疾的病人与外界进行交流提供新的手段。

参考文献：

[1] Middendorf M,McMillan G,Calhoun G,et al.Brain-computer interfaces based on the steady-state visual-evoked response[J].IEEE Trans Rehab Eng,2000,8(2):211-214.

[2] 何庆华，吴宝明，彭承琳.基于小波和神经网络的视觉诱发电位识别方法[J].仪器仪表学报，2007，28(6)：1003-1006.

[3] Wang Y J,Wang R P,Gao X R.A practical VEP-based brain-computer interface[J].IEEE Trans on Neural System and Rehabilitation Engineering,2006,14(2):234-239.

[4] Huang N E,Zheng S,Long S R.The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[C]//Proceedings of the Royal Society of London,1998,454(A):903-995.

[5] Yang J N,Lei Y,Pan S,et al.System identification of linear structures based on Hilbert-Huang spectral analysis[J].Earthquake Engineering and Structural Dynamics,2003,32(9):1443-1467.

[6] Huang N E.The age of large amplitude coastal seiches on the Caribbean coast of Puerto Rice[J].Phy Oceanography,2000,30(8):405-409.

[7] Huang N,Wu M,Long S.A confidence limit for the empirical mode decomposition and Hilbert spectral analysis[C]//Proc Roy Soc London,2003,459(A):2317-2345.

[8] Zhao H W,Huang N.A study of the characteristics of white noise using the empirical mode decomposition method[C]//Proc Roy Soc London,2004,460(A):1594-1611.

[9] Veltcheva D.Wave and group transformation by a hilbert spectrum[J].Coastal Engineering Journal,2002,44(4):283-300.

[10] Huang N,Zheng S,Long S.A new view of nonlinear water waves:The hilbert spectrum[J].Annual Review of Fluid Mechanics,1999,31(1):417-457.