支持向量机应用于大气污染物浓度预测

陈 俏[1],曹根牛[1],陈 柳[2]
(1.西安科技大学 理学院,陕西 西安 710054;2.西安科技大学 能源学院,陕西 西安 710054)

摘 要:支持向量机是基于统计学习理论的新一代机器学习技术,其非线性回归预测性能优越于传统统计方法。提出了一种大气污染物浓度预测模型,该方法将支持向量机应用于大气污染物浓度预测,首先对各类影响因子进行分析并进行建模预测;而后利用主成分分析的方法对输入因子降维,从而形成支持向量机的训练样本集;在此基础上建立了基于RBF核函数支持向量回归法的大气污染预模型。大气污染预测实例表明,该方法具有泛化能力强、预测精度高、训练速度快、稳定性好、便于建模等优点,有良好的应用前景。

关键词:支持向量机;大气污染预测;核函数
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2010)01-0250-04

Application of Support Vector Machine to Atmospheric Pollution Prediction

CHEN Qiao[1], CAO Gen-niu[1], CHEN Liu[2]
(1. College of Science, Xi'an University of Science and Technology, Xi'an 710054, China;
2. College of Energy, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: The support vector machine (SVM) as a new generation machinery learning technology based on statistical theory, has been reported to have better prediction performance of non-liner regression than traditional statistical methods. First, the input variables are analyzed, then dimensionality of input variables are reducted using principal component analysis(PCA) to form the training sample of the support vector machine. The appropriate forecasting methods are chosen and an SVM regression model for atmospheric pollution prediction is established. The testing results showed that the model based on support vector machine exhibited its properties of high forecast accuracy, fast training, high generalization capability and easy modeling.

Key words: support vector machine (SVM); atmospheric pollution prediction; kernel function

0 引 言

支持向量机(SVM)是Vapnik开发的基于统计学习理论的新一代机器学习技术[1],能较好地解决小样本、非线性、高维数和局部极小点等实际问题,已成为机器学习界的研究热点之一,并成功应用于分类、回归和时间序列预测等领域[2-4]。其遵循结构风险最小化原则,预测性能和推广能力优于神经网络,因而成为应用领域研究的热点。目前,大气污染物浓度统计预测方法多是传统统计模型,难以模拟复杂多变的大气污染变化[5]。神经网络较传统统计方法能更好地模拟大气污染因素的非线性关系,在大气污染预测应用中取得较好结果[6]。然而,神经网络具有推广能力差、过拟合、易于陷入局部最优、寻找结构参数复杂等缺点。文中通过实例论证,探讨支持向量回归方法应用于大气污染物浓度的可行性。

1 支持向量机原理

利用SVM进行回归与预测的基本思想[7,8]是通过非线性映射将数据映射到高维特征空间 Ω 中,并在该特征空间进行线性回归:

$$f(x) = (w \cdot \varphi(x)) + b \quad (1)$$

考虑 l 个独立分布的学习样本 $T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (X, Y)^l$,其中 $x_l \in X \in R^n$, $y_l \in Y \in R$, $i = 1, 2, \cdots, l$,在高维特征空间 Ω 中构造回归超平面。

用于回归分析的SVM主要有 ε-SVR和 v-SVR。在 ε-SVR中,需要事先确定 ε-不敏感损失函数中的参数 ε,然而在某些情况下选择合适的 ε 并非易

收稿日期:2009-04-27;**修回日期**:2009-07-09
基金项目:陕西省教育厅专项科研计划项目(07JK312)
作者简介:陈 俏(1980-),女,湖北武汉人,硕士研究生,研究方向为支持向量机的研究与应用;曹根牛,副教授,研究方向为最优化理论研究。

事。相比之下,v - SVR能够自动计算。因此文中以 v - SVR 为例予以说明。v - SVR 将回归分析问题转化为求解以下优化问题:

回归超平面的对应优化问题是:

$$\min_{w,b,\zeta^{(*)}} \frac{1}{2}\|w\|^2 + C(v\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\zeta_i + \zeta_i^*)),$$

$$\text{s.t.}\begin{cases}(w\cdot\varphi(x)) + b - y_i \leqslant \varepsilon + \zeta_i, i = 1,2,\cdots,l \\ y_i - (w\cdot\varphi(x)) - b \leqslant \varepsilon + \zeta_i^*, i = 1,2,\cdots,l \\ \zeta_i^{(*)} \geqslant 0, \varepsilon \geqslant 0, i = 1,2,\cdots,l\end{cases} \tag{2}$$

式中:$\zeta^{(*)} = (\zeta_1, \zeta_1^*, \cdots, \zeta_l, \zeta_l^*)^T$,$C$ 是惩罚因子;v 为控制支持向量机的个数;ε 为不敏感损失函数。引入 Lagrange 乘子构造 Lagrange 泛函,得到原问题的对偶问题:

$$\max_{a^{(*)}} \sum_{i=1}^{l} y_i(a_i^* - a_i) - \frac{1}{2}\sum_{i,j=1}^{l}(a_i^* - a_i)(a_j^* - a_j) \cdot K(x_i, x_j)$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{l}(a_i^* - a_i) = 0 \\ 0 \leqslant a_i^* \leqslant \frac{C}{l} \quad i = 1,2,\cdots,l \\ \sum_{i=1}^{l}(a_i + a_i^*) \leqslant C \cdot v\end{cases} \tag{3}$$

其中 $v \geqslant 0, C > 0$ 是常数。

所求的最优回归超平面可表示为:

$$f(x) = \sum_{i=1}^{l}(-a_i + a_i^*)K(x_i, x) + b \tag{4}$$

2　大气污染物浓度预测模型

建立基于支持向量机的大气污染物浓度变化的预测模型,关键问题是输入模式的确定、训练样本的选取以及模型结构参数的选取。文中拟以 SO_2 为例,建立大气污染物浓度预测模型。

2.1　建立大气污染物浓度预测模型的步骤

(1)构建有效的预测因子。由于大气污染物浓度(y)主要受污染源的源强和气象因子的影响,故考虑将前一天的 SO_2 浓度(x_1)、平均风速(x_2)、日均温度(x_3)、日均湿度(x_4)、日均气压(x_5)、日照时数(x_6)、总辐射量(x_7)、净辐射量(x_8)、总云量(x_9)共 9 个因子作为预选预测因子。

(2)选择核函数及参数值。常用的核函数有线性核函数、多项式核函数、径向基函数(RBF)核函数和 sigmoid 核函数。

(3)用训练样本训练具有优化参数的支持向量机预测模型,获得支持向量,确定支持向量机的结构。

(4)用训练过的支持向量预测器对测试样本预测。

2.2　预测模型的具体应用

2.2.1　资料来源

文中采用的 SO_2 浓度资料由西安市环境监测站提供,监测 SO_2 浓度值为全市日平均浓度值。资料取 2001,2002 年 1～12 月。对应的气象资料由陕西省气象局提供。

2.2.2　试验软件

Libsvm 是台湾大学林智仁等开发设计的一个简单、易于使用、快速、有效的 SVM 模式识别与回归的软件程序。它不但提供了编译好的基于 Windows 操作系统的执行文件,还提供了有关的软件程序源代码,方便改进、修改以及在其他操作系统上应用。

2.2.3　数据的预处理

将前一天的 SO_2 浓度、平均风速、日均温度、日均湿度、日均气压、日照时数、总辐射量、净辐射量、总云量共 9 个因子作为 SVM 预测模型的预选输入因子,输出为当日 SO_2 浓度,为了防治数据溢出,同时加快运算速度,对训练前、后的数据均进行归一化处理,将输入输出数据变换为[-1,1]区间的值。

为了消除各输入因子间的相关性,对输入的 9 个因子进行主元分析,主元变换后的 7 个因子的矩阵的累计贡献率为 93%,因此,用主元变换后的 7 个因子的矩阵作为 SVM 预测模型的输入因子。

2.2.4　SVM 的学习训练及预测

把 2001 年全年共 365 组数据作为训练样本,每组数据包括 7 个输入因子一个 SO_2 实际值。把 2002 年全年共 365 组数据作为测试样本,每组数据包括 7 个输入因子,对每日的 SO_2 进行预测。由于核函数和惩罚因子 C 是 SVM 模型的主要参数,它们对预测结果影响很大。如何合理选择 SVM 模型的参数,目前尚无有效的方法。文中通过交叉试验的方法选取核函数及惩罚因子。

(1) 核函数对分析结果的影响。

核函数反映了训练数据样本的特性,对于系统的泛化能力影响较大。通过交叉试验选择不同核函数建立的 SVM 预测模型,其预测的平均相对误差和均方根误差如表 1 所示。

表 1　不同核函数对分析结果的影响

核函数	平均相对误差	均方根误差
线性核函数	0.1814	0.0106
多项式核函数	0.1749	0.0107
sigmoid 核函数	0.1671	0.0099
RBF 核函数	0.1581	0.0098

对比表 1 中的数据可知,核函数的选择不同,分析结果也不同。因此,大气污染物支持向量机预测模型中核函数选用 RBF 核函数。

(2)惩罚因子 C 对分析结果的影响。

惩罚因子 C 为正常数,惩罚因子 C 决定了对超出误差 ε 的样本惩罚程度。从结构风险的角度考虑,C 值取得过大,问题倾向于经验最小,忽略对结构复杂程度的考虑;反之则更多地考虑了问题的复杂程度,忽略了经验数据的作用。因此也可以说,C 是支持向量机回归和泛化能力的平衡参数。

通过交叉试验选择不同惩罚因子 C 建立的 SVM 预测模型,表 2 为选择 RBF 核函数,惩罚因子 C 变化、其他参数不变时对分析结果的影响。

表 2 惩罚因子 C 对分析结果的影响

惩罚因子 C	平均相对误差	均方根误差
1	0.1748	0.0102
2	0.1755	0.0103
4	0.1790	0.0105
6	0.1808	0.0106
9	0.1815	0.0106
10	0.1810	0.0106
20	0.1793	0.0105
40	0.1754	0.0103
80	0.1734	0.0102
100	0.1722	0.0102
500	0.1661	0.0100
1000	0.1581	0.0098
1500	0.2419	0.0151
2000	0.2527	0.0159
10000	0.3116	0.0200

惩罚因子 C 取不同的常数值,对结果有不同的影响。当 C 大于一定值时,其变化对分析结果产生的影响变小。文中惩罚因子 C 的范围为 1 ~ 10000,表 2 中表明 C 取 1000 时,误差最小,故大气污染物支持向量机预测模型中 C 取 1000。

(3)算例分析。

针对以上对核函数及惩罚因子 C 的分析,通过 libsvm 建立模型如下:

用于建模的 SVM 为 v - SVM,SVM 采用的核函数为径向基内积函数(RBF 函数):

$$K(x,y) = e^{-\gamma \| x-y \|^2} \tag{5}$$

由交叉验证选取惩罚因子 $C = 1000$,核函数中的参数 $g = 0.001$,训练误差 $e = 0.0001$,对训练样本进行训练,最后对测试样本进行预测,结果如图 1 ~ 图 3 所示。

2.2.5 预测结果分析

图 1 可以看出,SO_2 的监测值和预测值符合得较好,特别是 1~3 月的 SO_2 拟合程度更高。因此模型对 SO_2 的浓度总体变化趋势反映较敏感;由图 2 可以看出,2002 年逐日 SO_2 监测值与预测值相关系数 $R = 0.841$,说明 SO_2 监测值与预测值相关性比较高;由图 3 可以看出,2002 年逐日 SO_2 监测值与预测值相对误差除个别突变点外,大部分在 25% 左右,而突变的原因可能和该时间段内的天气异常有关。

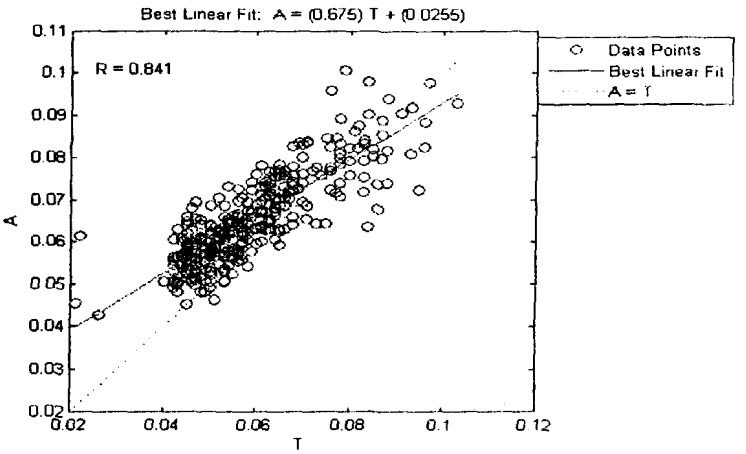

图 2 2002 年逐日 SO_2 监测值与预测值相关系数图

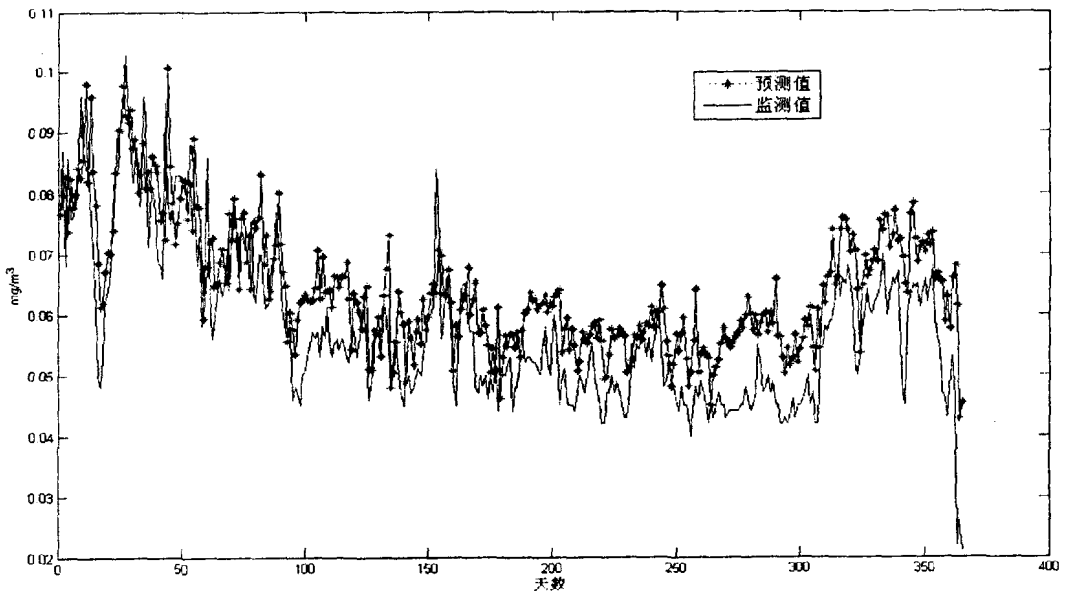

图 1 2002 年逐日 SO_2 监测值与预测值浓度对比图

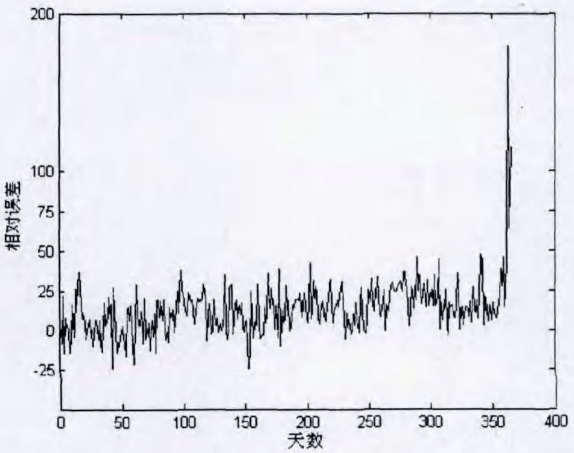

图 3　2002 年逐日 SO_2 监测值与预测值相对误差图

3　结束语

(1)由于支持向量机的优良特性,特别适合于那些模糊、随机、不确定性、样本数有限和非线性的复杂问题,支持向量回归模型能很好捕捉大气污染物浓度与其影响因子的非线性关系,通过实例证实支持向量机对大气污染的预测具有预测精度高和训练速度快的优点,并有广泛的应用前景。

(2)支持向量机预测大气污染模型的核函数以及惩罚参数 C 的选择将直接影响到支持向量机的学习效率和推广能力。文中核函数取 RBF 函数,通过对核函数核惩罚参数 C 的交叉测试, C 取 1000。

(3)SVM 方法属于参数预测方法,其预测精度在很大程度上依赖于预测模型的输入和输出参数的代表性。大气污染物浓度预测的可靠性和准确性,依赖对其各种影响因素的准确分析。

参考文献:

[1] 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,2006:224-235.

[2] Nello C,John S T.支持向量机导论[M].李国正,王　猛,曾华军译.北京:电子工业出版社,2005:98-105.

[3] 白　鹏,张喜斌,张　斌,等.支持向量机理论及工程应用实例[M].西安:西安电子科技大学出版社,2008:41-55.

[4] 许建华,张学工,李衍达.支持向量机的新发展[J].控制与决策,2004,19(5):482-493.

[5] 谈建国,邵德民,黄家鑫.上海城市空气质量预报(日报)业务系统探讨[J].气象,2001,27(6):33-35.

[6] 金　龙.人工神经网络技术发展及在大气科学领域的应用[J].气象科技,2004,32(6):12-13.

[7] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000(1):32-42.

[8] Vapnik V N.统计学习理论的本质[M].张学工译.北京:清华大学出版社,2000:96-98.

(上接第 249 页)

$e_t \sim \mathrm{IID}(0,1)$

3.4　模型比较

利用上面得到的两个模型分别对上证综指的 10 月 13 日,14 日,15 日三天的收盘价作个短期预测,得到结果如表 4 和表 5 所示。

表 4　AR 模型预测上证综指

日期	03.10.13	03.10.14	03.10.15
真实值	1399.66	1388.17	1383.10
AR 预测值	1405.40	1391.50	1421.69
绝对误差(%)	0.41	0.24	2.79

表 5　GARCH 模型预测上证综指

日期	03.10.13	03.10.14	03.10.15
真实值	1399.66	1388.17	1383.10
GARCH 预测值	1405.19	1390.47	1421.13
绝对误差(%)	0.395	0.166	2.75

4　结束语

文中简要介绍了时间序列预测方法,分析了其在证券市场分析中的应用的可能和效果。并通过对上证综指的日收益率进行的实证研究,发现市场波动存在集群性,总体而言,GARCH 模型能比 ARCH 模型更好地反映这类特征。

参考文献:

[1] 谷　赫.时间序列的数据挖掘在证券预测分析中的应用研究[D].长春:吉林大学,2005.

[2] Mills T C.金融时间序列的经济计量学模型[M].北京:经济科学出版社,2002:217-355.

[3] 叶　峰.汇率的可预测性实证分析[J].管理工程学报,2001(3):41-45.

[4] 孔淑慧.流数据时序模式依赖挖掘在股市行情分析中的应用[D].北京:北京交通大学,2008.

[5] Weigend A S. Time Series Analysis and Prediction[D]. Colorado: Univerisity of Colorado,1994.

[6] Faloutsos C, Rangsnathan M, Manoloppoulos Y. Fast stubsequence matching in time-series databases[C]//In: SIGMOD Proceedings of Annual Conference. Minneapolis: [s. n.], 1994:419-429.

[7] Xia B B. Similarity search in time series data set[D]. Canada: Simon Fraser University,1997.

[8] 佚　名. GARCH 模型对沪市行业指数的实证研究[EB/OL]. 2008-10-12. http://www.govyi.com/lunwen/2008/200810/262942.shtml.

[9] 王　珏,秦伟良,钱海荣.上海股市的时间序列模型研究[J].理论新探,2004(11):11-23.