【文章编号】 1004-1540(2008)04-0355-05

EMD-SVM 非线性组合模型对高炉铁水含硅量的预测

王义康
(中国计量学院 理学院,浙江 杭州 310018)

【摘 要】 提出了一种基于经验模式分解(EMD)和支持向量机(SVM)的非线性组合模型的预测方法.该方法运用 EMD 将原始铁水含硅量的时间序列分解成若干个频率不同的平稳分量,分解后的分量突出了原序列的局部特征.通过 Lempel-Ziv 复杂度分析选用不同的核函数,并利用 10-fold 交叉检验方法取定相应的参数,从而对各个分量构建不同的支持向量机模型,并对各分量进行预测.仿真结果表明,EMD-SVM 非线性组合模型预测命中率达到 90%.

【关键词】 经验模式分解;支持向量机;铁水含硅量;组合模型
【中图分类号】 TF531 【文献标识码】 A

Prediction of silicon content in hot metal based on EMD-SVM nonlinear combined model

WANG Yi-kang
(College of Sciences, China Jiliang University, Hangzhou 310018, China)

Abstract: In the blast furnace (BF) ironmaking process, the silicon content in hot metal which reflects the thermal state of BF is an important index. In order to predict the silicon content in hot metal effectively and level up the forecasting accuracy, a combined model based on Empirical Mode Decomposition (EMD) and Support Vector Machine (SVM) is proposed. Firstly, the time series data of silicon content in hot metal are decomposed into a series of stationary intrinsic mode functions (IMFs) in different scale space via the EMD sifting procedure. The local features of original time series data are prominent in the IMFs. Secondly, based on the analysis of Lemple-Ziv complexity and 10-fold cross validation, the right kernel functions and their parameters are chosen to build different SVMs respectively to predict each IMF. Finally, the predicted results of all IMFs are reconstructed to obtain the final predicted result. The result shows that the prediction is successful and the hit rate increased to 90%.

Key words: empirical mode decomposition; support vector machine; silicon content in hot metal; combined model

【收稿日期】 2008-08-18
【作者简介】 王义康(1976 -),男,安徽寿县人,讲师.主要研究方向为系统优化与控制、数学建模.

在高炉冶炼过程中,铁水含硅量是评定生铁质量的重要指标,也是表征高炉热状态及其变化的标志之一,还是高炉操作管理的重要参数,而且铁水的含硅量对于炼钢过程中的渣量生成以及钢液的脱硫、脱磷条件有很大的影响[1]. 因此,研究较为准确的预测方法,就成为炼铁生产中的重要课题,但迄今为止,这个问题仍未得到很好的解决. 现有的铁水含硅量预测方法主要有回归分析法[2]、人工神经网络法[3,4]、混沌预测法[5]、分形预测法[6]等. 这些模型各有优缺点和适用条件. 为了取长补短,高炉铁水含硅量的组合预测方法将是一个重要的发展方向.

高炉冶炼过程高度复杂,它的操作机制具有非线性、大时滞、高维数、大噪声、分布参数等特征. 针对高炉铁水含硅量时间序列非平稳、非线性的特征,我们在经验模式分解的基础上构建支持向量机预测模型,并对包钢6号高炉在线采集的[Si]时间序列数据进行离线预测. 结果表明基于经验模式分解和支持向量机的非线性组合模型可以达到较高的预报命中率.

1 经验模式分解(EMD)

经验模式分解(EMD)是在1998年Huang等人提出的一种基于经验模式分解的Hilbert-Huang变换数据处理方法[7]. 它通过对信号的"筛选",将信号分解成不同频率的本征模态函数(IMF). IMF具有如下特点:从全局特征上看,极值点数必须和过零点数一致或至多相差一个;任意时刻,其极大值包络线和极小值包络线的均值必须是零. 具体的分解步骤为:

1) 对于给定序列 $x(t)$,找出 $x(t)$ 所有的极大值点和极小值点,然后用三次样条曲线将所有的局部极大值点和局部极小值点连接起来形成上包络线和下包络线.

2) 记上、下包络线的平均值为 $m_1(t)$,求出:$h_1(t) = x(t) - m_1(t)$,如果 $h_1(t)$ 是一个IMF分量,那么 $h_1(t)$ 就是 $x(t)$ 的第一个分量.

3) 如果 $h_1(t)$ 不满足IMF的条件,把 $h_1(t)$ 作为原始数据,重复步骤(1)和(2);得到上、下包络线的均值 $m_{11}(t)$,再判断 $h_{11}(t) = h_1(t) - m_{11}(t)$ 是否满足IMF的条件. 如不满足,则重复循环 k 次,得到 $h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t)$,使得 $h_{1k}(t)$ 满足IMF的条件. 记 $c_1(t) = h_{1k}(t)$,则 $c_1(t)$ 为信号 $x(t)$ 的第一个满足IMF条件的分量.

4) 将 $c_1(t)$ 从 $x(t)$ 中分离出来,得到 $r_1(t) = x(t) - c_1(t)$;$r_1(t)$ 应包含较长周期分量,把它作为新的原始信号,重复步骤(3)的筛选过程,得到 $x(t)$ 的第二个满足IMF条件的分量 $c_2(t)$. 以上过程重复循环 n 次,得到信号 $x(t)$ 的 n 个满足IMF条件的分量:$r_1(t) - c_2(t) = r_2(t), \cdots, r_{n-1}(t) - c_n(t) = r_n(t)$. 通过筛选过程可以消除叠加波,使不平整的振幅平滑. 直到 $c_n(t)$ 或 $r_n(t)$ 满足给定的终止条件时筛选结束. 最后原始序列 $x(t)$ 可表示为:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{1}$$

式(1)表明,可以将信号 $x(t)$ 分解成从高到低不同频率段的 n 个IMF分量和一个趋势项 $r_n(t)$ 之和,因每个IMF分量代表一个特征尺度的数据序列,故"筛选"的过程实际上是将原始数据序列分解成各种不同特征波动序列的叠加.

2 支持向量机(SVM)原理

支持向量机(SVM)是由Vapnik等人提出的一种建立在VC维学习理论和结构风险最小化原则基础上的一种学习方法[8]. 该方法具有结构简单、学习速度快、全局最优、泛化性能好等优点,能较好解决小样本、非线性、高维数和局部极小等问题. 其模型的基本思想[9]是:对给定的训练样本集合 $T = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$(其中 $\boldsymbol{x}_i$ 为输入向量,y_i 为输出数值,n 为样本总数),通过一个非线性映射 ϕ,把数据空间中的输入向量 $\boldsymbol{x}$,映射到高维特征空间,然后在特征空间中进行线性回归,构造出最优学习器:

$$f(x) = \omega^T \phi(x) + b \tag{2}$$

式(2)中 ω 和 b 是通过正则化和结构风险准则来估计的. 根据结构风险最小化准则得:

$$\min \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)^2$$

$$s.t. \begin{cases} y_i - \omega^T \phi(x_i) - b \leqslant \varepsilon + \xi_i \\ \omega^T \phi(x_i) + b - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0, \quad i = 1, 2, \cdots, n \end{cases} \tag{3}$$

式(3)中 ξ_i 和 ξ_i^* 为松弛变量,分别表示在不敏感

损失 ε 的约束下的训练误差的容许上限和容许下限,$C(\geqslant 0)$ 为可调参数(惩罚因子),控制对误差超出 ε 的样本的惩罚程度.

根据对偶理论及鞍点条件,原始问题(3) 式的对偶形式为:

$$\min \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\phi(x_i),\phi(x_j)) + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)$$

$$s.t.\begin{cases}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0,\\ 0 \leqslant \alpha_i \leqslant C, \quad 0 \leqslant \alpha_i^* \leqslant C,\\ i = 1,2,\cdots,l\end{cases} \tag{4}$$

且有

$$\omega = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi(x_i) \tag{5}$$

其中 α_i 和 α_i^* 是非负的拉格朗日乘子,(4) 式为凸二次规划问题,α_i 和 α_i^* 可以通过(4) 式得到. 根据 Karush-Kuhn-Tucker 条件[10],可求得 α_i,α_i^* 和 b,其中只有少数 α_i,α_i^* 不为零,这些参数对应的样本,即在不灵敏区边界上或外面的样本,称为支持向量(*SV*). 将(5) 式带入(2) 式,从而得到:

$$f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)k(x_i,x) + b \tag{6}$$

式(6) 中 $k(x_i,x_j) = (\phi(x_i),\phi(x_j))$ 称为核函数,任何函数只要满足 Mercer 条件[11],均可作为核函数. 常用的核函数有:

1) 线性核函数:

$$k(x,y) = (x,y) \tag{7}$$

2) 多项式核函数:

$$k(x,y) = ((x,y) + t)^d \tag{8}$$

3) 径向基核函数:

$$k(x,y) = \exp\left(-\frac{\| x - y \|^2}{\sigma^2}\right) \tag{9}$$

4)Sigmoid 核函数:

$$k(x,y) = \tanh(\alpha(x,y) + \beta) \tag{10}$$

其中 t,d,σ^2,α 和 β 为核函数的参数.

3　基于 EMD 分解的 SVM 模型对铁水含硅量的预测

3.1　预测方法

高炉铁水含硅量时间序列具有一定的非平稳、非线性特征,EMD 对硅值序列具有平稳化的作用,能将铁水硅序列按其内在特性自适应分解为若干个不同频率的平稳 IMF. 利用 EMD 对某一时段内的样本数据进行分解,分解后的 IMF 突出了原序列的局部特征,能更明显地看出原序列的数据特征,对其进行分析,可以更清楚地把握硅序列的特性. 在此基础上,根据 IMF 的变化特点分别选用不同的 SVM 模型建立预测模型,最后将预测结果进行重构得到最终预测结果. 其预测结构图如图 1.

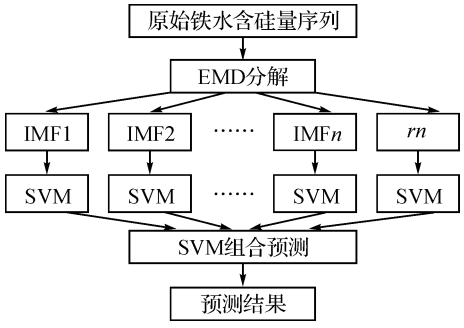

图 1　基于 EMD 分解的 SVM 预测模型结构图

Figure 1　The architecture of the combined model of EMD and SVM

3.2　铁水含硅量序列的 EMD 分解

以包钢 6 号高炉 2007 年 5 月 3 日到 2007 年 6 月 10 日采集的高炉铁水含硅量数据(炉号:No 1827—2362)为样本,其时间序列图如图 2.

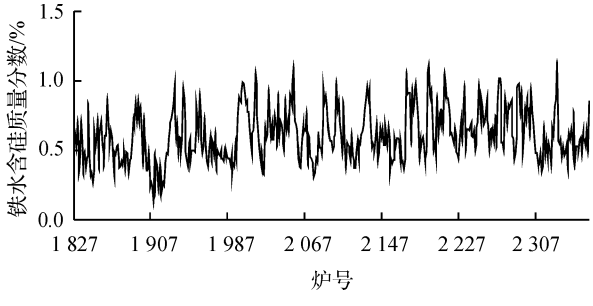

图 2　包钢 6 号高炉铁水含[Si]量时间序列图

Figure 2　Time series data of silicon content from blast furnace No. 6 of Bao Steel

对图 2 的[Si]时间序列做 EMD 分解,可得到其本征模态分量和剩余分量,如图 3 和图 4.

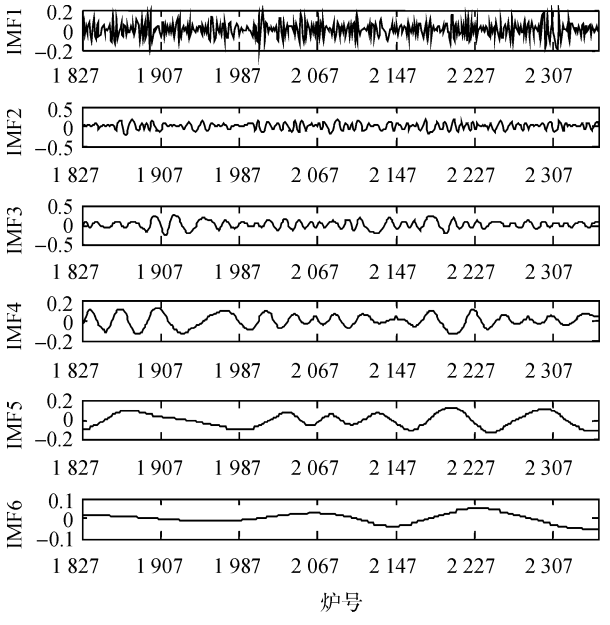

图 3 [Si]时间序列 EMD 分解的本征模态分量

Figure 3 The IMFs of EMD of the time series data of silicon content

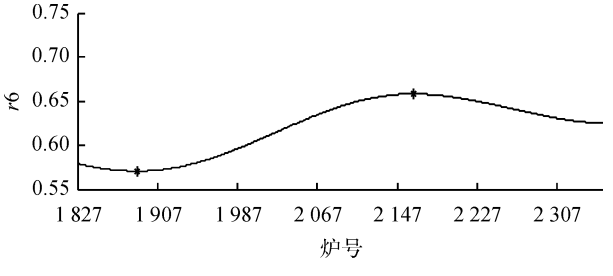

图 4 [Si]时间序列 EMD 分解的剩余分量

Figure 4 The residue of EMD of the time series data of silicon content

EMD 分解方法的合理性是由其分解的完备性和正交性来决定的[12]. 根据式(1),经过计算得到由所有本征模态分量与剩余分量重构的铁水含硅量序列与原始硅序列的差异绝对值之和为3.15×10^{-14},这个可以认为来自于计算机的舍入误差. 分解的正交指数经计算为 0.002 96. 因此,从数值上保证了分解的完备性和正交性.

3.3 SVM 模型核函数及其参数的选择

核函数的选取对 SVM 性能的优劣有比较大的影响. 由于人们对核函数所对应的特征空间的性质缺乏了解,核函数的选择还没有统一的标准. 笔者利用 Lanckriet 等人的研究结论[12]:核函数的选择与数据的复杂度相适应. 通过 Lempel-Ziv 复杂度分析[13]得到表 1:

表 1 Lempel-Ziv 复杂度计算结果

Table 1 The result of Lempel-Ziv complexity

序列	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	r_6
L-Z 复杂度	0.824 8	0.690 5	0.360 9	0.154 7	0.103 1	0.051 6	0.016 7

因此,对于剩余分量 r_6 以线性核构造向量机模型进行拟合,本征模态分量 IMF4、IMF5、IMF6 采用多项式核,复杂度较大的 IMF1、IMF2、IMF3 采用径向基核. 对于核函数的参数的选择采用 10-fold 交叉检验方法[14],主要结果如表 2. 模型中不敏感损失 ε 取为 0.001.

表 2 10-fold 交叉检验参数选择结果

Table 2 The main result of choice of 10−fold cross validation

参数	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	r_6
C	2.219	6.33	75.49	272.73	167.16	3.85	220.27
σ^2	0.343	0.261	0.272	—	—	—	—
d	—	—	—	2	2	2	—
t	—	—	—	15.76	12.25	0.99	—

3.4 预测结果

根据选定的核函数和相应的参数,取前 536 炉数据作为训练数据,对后 100 炉铁水含硅量数据作为测试数据,根据 EMD 分解的完备性以及公式(1),包钢 6 号高炉 2007 年 6 月上旬采集的连续 100 炉铁水含硅量的预测结果如图 5.

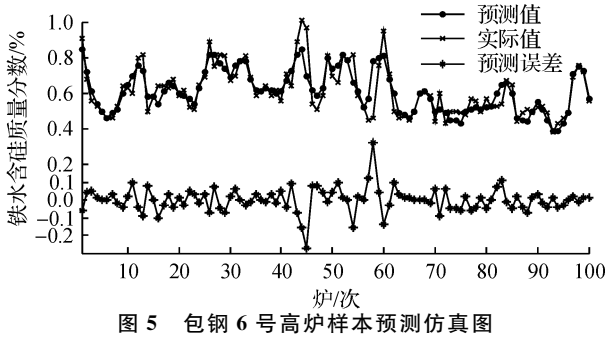

图 5 包钢 6 号高炉样本预测仿真图

Figure 5 Comparison between the real values and predicted values from blast furnace No. 6 of Bao Steel

经过统计计算,得到在行业规定的±0.1 误

差范围内预测命中率为 90%(有些 3 000 m^3 以上的大型高炉由于[Si]值波动较小,要求误差在±0.05),这是一个较理想的结果,比工程中常用的 n 阶自回归模型效果要好很多,表明基于 EMD 分解的 SVM 模型可以较好的预测铁水含硅量.

4 结　语

应用 EMD 方法对高炉铁水含硅量序列进行经验模式分解,把复杂的数据序列分解成为一系列简单的本征模态函数. EMD 分解不同于小波方法,其分解是自适应的、局部的,分解的基直接来源于实际数据,因此它特别适用于非平稳信号的处理. 结合 SVM 进行预测,预测效果良好,其命中率达 90%. 由于分解以后得到的各个本征模态函数及剩余分量的 Lempel－Ziv 复杂度显著不同,而支持向量机方法在核函数的选取上具有灵活性,因此,采用基于 EMD 分解的支持向量机模型,能够达到较高的命中率. 然而,该方法需要较多人为因素的参与,例如,EMD 分解效果的验证、各分解变量的复杂度分析、支持向量机核函数及其参数的确定等,这些使得模型有一定的不确定性.

【参　考　文　献】

[1] 刘祥官,刘　芳. 高炉炼铁过程优化与智能控制系统[M]. 北京:冶金工业出版社,2003:55-59.

[2] SAXEN H. Short-term prediction of silicon content in pig iron[J]. Canadian Metallurgical Quarterly, 1994,33(4):319-326.

[3] CHEN J. A predictive system for blast furnaces by integrating a neural network with qualitative analysis[J]. Engineering Applications of Artificial Intelligence, 2001,14(1):77-85.

[4] JIMENEZ J, MOCHON J, AYALA J S, et al. Blast furnace hot metal temperature prediction through neural networks-based models[J]. ISIJ Int, 2004,44(3):573-580.

[5] 郜传厚,周志敏,邵之江. 高炉冶炼过程的混沌性解析[J]. 物理学报,2005,54(4):433-436.

[6] LUO S H, LIU X G, ZHAO M. Using generalized iterated function systems to model BF hot metal silicon content sequences[C]// IEEE 6th WCICA. Dalian City: IEEE, 2006:7766-7770.

[7] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Pro R Soc Lond, 1998,454:903-995.

[8] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge Uni Press, 2000:93-122.

[9] BURGES C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998,2(2):121-127.

[10] 徐成贤,陈志平,李乃成. 近代优化方法[M]. 北京:科学出版社,2002:105-122.

[11] VAPNIK V N. Statistical learning theory[M]. New York: Wiley, 1998:1926-1940.

[12] LANCKRIET G, CRISTIANINI N, BARTLETT P, et al. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine Learning Research, 2004(5):27-72.

[13] LEMPEL A, ZIV J. On the complexity of finite sequences [J]. IEEE Transaction on Information Theory, 1976,22(1):75-81.

[14] CHIH W H, CHIH C C, CHIH J L. A practical guide to support vector classification [EB/OL]. (2003-08-10) [2008-08-10]. http://www.csie.ntu.edu.tw/～cjlin/papers/guide/guide.pdf.