基于组合 SVR 的非平稳时间序列的模糊建模方法

林树宽, 支力佳, 张少敏, 乔建忠, 王国仁, 于 戈
(东北大学信息科学与工程学院, 辽宁沈阳 110004)

摘 要: 本文介绍一种对非平稳时间序列建模的新方法. 参考 Janos Abonyi 提出的应用于时间序列的模糊分块算法, 将该算法与改进的支持向量回归模型结合起来. 首先, 提出一种改进的支持向量回归的表达形式; 然后, 通过启发式的加权方法将模糊分块的信息与 SVR 结合起来; 最后, 提出一种基于组合 SVR 的建模方法. 实验结果表明, 本文提出的方法对于非平稳时间序列的建模具有较高的实用价值.

关键词: 非平稳时间序列; 模糊分块; 启发式的 ε 加权方法; 组合支持向量回归
中图分类号: TP181 文献标识码: A 文章编号: 0372-2112 (2006) 10-1929-04

A Multi-SVR Based Fuzzy Modeling Method for Non-Stationary Time Series

LIN Shu-kuan, ZHI Li-jia, ZHANG Shao-min, QIAO Jian-zhong, WANG Guo-ren, YU Ge
(*College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China*)

Abstract: A new approach for modeling non-stationary time series was introduced in this paper. Combine the fuzzy segmentation which was proposed by Janos Abonyi with Support Vector Machines(SVMs). Firstly, a modified Support Vector Regression (SVR) was proposed; Secondly, fuzzy segment information was combined with SVR by heuristic weighting method; Thirdly, we discussed a model based on multi-SVR. Experimental results show that the method proposed in this paper has great practical values for non-stationary time series modeling.

Key words: non-stationary time series; fuzzy segmentation; heuristic weighting method; multi support vector regression

1 引言

近年来, 支持向量机(SVMs) 因其较好的性能被广泛应用于数据分类和回归. 同样, 它也可以应用于时间序列的建模[1].

现实世界中的时间序列数据通常来自于工业生产、商业活动等等, 通常这些时间序列数据都是非平稳的. 处理非平稳时间序列的一种有效的方法是先对它进行分块, 使得同一个分块中大部分的数据来自同一个数据源[2], 然后再对每一个分块进行建模.

在非平稳时间序列中, 两个相邻分块的变化通常是模糊渐变的, 而不是在某一个时间点上瞬间完成. 所以, 要准确定义这些分块的边界是不现实的. 通常, 对于交叉部分的每一个时间点可能属于一个分块多一些, 而属于另一个分块少一些. 自然地, 模糊集就被用来处理交叉这部分数据点. 模糊逻辑已被广泛用于处理重叠和含糊的数据[3], 很多研究表明可以将模糊逻辑和时间序列分析结合起来考虑[4].

本文根据模糊分块的信息, 用 SVR 来对非平稳时间序列进行建模. 特别考虑了两个相邻分块之间的交叉部分. 这部分数据点并不明显地属于相邻的任何一个分块, 同时它们又被相邻的分块影响, 因此, 对于这些交叉部分应采用一种组合的方式建模.

除了用改进的 SVR 表达形式对非平稳时间序列进行建模, 本文还将交叉部分的数据点的模糊信息和组合 SVR 结合起来, 建立一种基于组合 SVR 的模糊建模方法.

2 根据模糊信息对非平稳时间序列进行建模

2.1 时间序列的模糊分块

时间序列 $T=\{x_k \mid 1\leqslant k\leqslant l\}$ 是由 l 个样本点组成的有限集, 并被时间点 $t_1, t_2, \cdots, t_l$ 标记. 传统的 C-分块策略是将时间序列 T 分成 C 个不重叠的分块 $S_T^C=\{S_i(a_i, b_i) \mid 1\leqslant i\leqslant C\}$, T 中的任何一块都是连续时间点的集合, $S_i(a_i, b_i)=\{x_k \mid a_i\leqslant k\leqslant b_i\}$; 然后分别对每一块用一个函数建模.

在文献[5]中, 作者提出一种基于时间序列分块算法的模糊聚类方法, 它假定数据被当作一个混合多变量的高斯分布被建模(包括时间变量). 这里, 我们参考文献[5], 定义如下的测度函数并按之进行模糊分块:

$$J=\sum_{i=1}^{c}\sum_{k=1}^{l}(\mu_{i,k})^m D_{i,k}^2(\eta_i, z_k)$$
$$=\sum_{i=1}^{c}\sum_{k=1}^{l}(\mu_{i,k})^m D_{i,k}^2(v_i^t, t_k)\, D_{i,k}^2(v_i^x, x_k) \quad (1)$$

详细的说明和聚类过程参见文献[5~7], 这里不再赘述.

收稿日期: 2005-10-11; 修回日期: 2006-07-03
基金项目: 国家自然科学基金(No. 60473074); 辽宁省自然科学基金(No. 20042015); 沈阳市自然科学基金(No. 1041036-1-06.07)

2.2 改进的 **SVR** 表达式

SVR 体现了核方法的思想, 即将非线性的输入数据映射到高维空间, 使之在高维空间中能被线性处理, 用以近似估计给定数据之间的潜在关系.

标准的 SVR 对于所有的数据点使用同样的不敏感损失 ε. 但是, 事实上, 不同的样本数据应该对应不同的 ε. 为此, 对于每一个数据点的 ε 使用不同的权值 q^t, 以使得每个样本数据点对应大小不同的 ε 管道, 则可以得到如下的 SVR 变换形式:

$$\min_{w_i,\, b_i,\, \xi\, \xi^*} \frac{1}{2} w^T w + C \sum_{t=1}^{l} (\xi^t + \xi^{t*}) \tag{2}$$

$$\text{subject to} \quad \begin{aligned} &-\varepsilon q^t - \xi^{t*} \leqslant y_t - (w^T \phi(x_t) + b) \leqslant \varepsilon q^t + \xi^t, \\ &\xi^t \geqslant 0, \xi^{t*} \geqslant 0, \ t = 1, 2, \cdots, l, \end{aligned}$$

其中,

$$q^t = (1-d)^t, \ t = 1, 2, \cdots, l \tag{3}$$

这里, d 是变化因子(介于 0 与 1 之间), 权值 q^1 所对应的样本 x_1 是离预测点最远的数据, 权值 q^l 所对应的样本 x_l 是离预测点最近的数据. 通过求解式(2)得到模型.

2.3 模糊分块信息和改进的 **SVR** 的结合

给定一个平稳的时间序列 $T = \{y_t \mid 1 \leqslant t \leqslant l\}$, $y_t = f(x_t)$, 其中 $x_t = [y_{t-1}, y_{t-2}, \cdots, y_{t-d}]^T$, d 称为步长. 然而, 对于非平稳时间序列则不然, 因为不能用单一的 f 函数来拟合整个时间序列, 相反, 要用 m 个函数(也就是说, m 个模式)[8], 即用

$$y_t = \sum_{i=1}^{m} p_i^t f_i(x_t), \ \forall\, t = 1, \cdots, l \tag{4}$$

$$\text{subject to} \quad \begin{aligned} &\sum_{i=1}^{m} p_i^t = 1, \ \forall\, t = 1, \cdots, l \\ &0 \leqslant p_i^t \leqslant 1, \ \forall\, t = 1, \cdots, l, \ i = 1, \cdots, m \end{aligned}$$

来对给定的时间序列进行建模. 通常, 当处理现实数据时, 并不能准确得到数据源的个数. 所以, 通常要基于某种信息来估计 m 的值.

在这一部分, 我们的目标是定出式(4)中模式 m 的个数; 并且对于上述改进的 SVR 表达形式, 采用模糊分块的信息来对 ε 进行启发式加权, 估计出每一个 $f_i(x_t)$.

首先如 2.1 所述对给定的时间序列进行分块, 产生多个重叠的分块 $S_i (i = 1, \cdots, R)$. 然后定义一些条件来将所有的分块分为两类: 较大的分块 $LS_j (j = 1, \cdots, h, h < R)$ 和两个较大分块之间的交叉部分 $I_k (k = 1, \cdots, g, g < R)$, h 和 g 之和可能小于 R.

一个分块 $S_i(a_i, b_i)$ 是一个较大的分块, 当且仅当存在 bl_i 和 $br_i (a_i < bl_i < br_i < b_i)$, 它们满足下列条件:

$$\begin{aligned} &p(z_k \mid \eta_i) > p_{\text{thre}} (\forall\, k \quad bl_i \leqslant k \leqslant br_i) \\ &\text{and} \quad |S_i(bl_i, br_i)| \geqslant L_{\text{low}} \end{aligned} \tag{5}$$

这里, $p(z_k \mid \eta_i)$ 表示向量 $z_k = [t_k, x_k]^T$ 属于第 i 类的概率, p_{thre} 是为概率 $p(z_k \mid \eta_i)$ 设定的一个阈值, L_{low} 是为每一个分块的长度设定的一个下限. 使用 bl_i 和 $br_i (i = 1, \cdots, h)$ 作为边界点, 我们重新将整个数据集分为几个不重叠的数据集 LS 和 I.

由于内在的同质性, 对每一个较大的分块 LS, 像对待平稳时间序列一样使用一个标准的 SVR 来建模. 即, 对于每一个 $LS_i (i = 1, \cdots, h)$, 式(4)中 m 的值为 1. 采用 2.2 中所述的方法对 LS 建模.

对于交叉部分 I_i, 需要更仔细的考虑. I_i 由不同模式交叉部分所包含的数据点组成. 除了受相邻两个较大的分块 LS_i 和 LS_{i+1} 的影响, I_i 可能还包含几个分块 S. 所以, 要考虑这些分块的影响. 使用 $z_k \in I_i$ 来说明数据点 z_k 在 I_i 中; $t \in I_i$ 用来说明时间点 t 在 I_i 中; $\eta_j \cap I_i \neq \varnothing$ 和 $S_j \cap I_i \neq \varnothing$ 用来说明同样的意思, 即分块 S_j 在 I_i 中. 对于在 I_i 中的每一个数据点 $z_k (z_k \in I_i)$, 检查 $p(z_k \mid \eta_j)(\eta_j \cap I_i \neq \varnothing)$, 发现它们大多数都明显小于 1. 这意味着在一个交叉部分 I_i 的数据点除了受到分块的影响, 还有它自己的属性.

关于 I_i 自己的属性, 假设它可以被一个高斯分布有效地建模. I_i 中模式 m 的个数可以从分块 S_j 在 $I_i (S_j \cap I_i \neq \varnothing)$ 中的个数 n 决定. 对于每一个分块 S_i, 用一个函数 $f_i(x_t)$ 就可以对它建模. 之后, 使用改进的 SVR 来近似估计 $f_i(x_t)$. 再加上一个假设的高斯函数 $f_m(x_t)$, 就可以得到 I_i 中的 m, 即 $m = n + 1$.

对于每一个交叉部分 I_i, 分别对 $f_1(x_t), \cdots, f_n(x_t)$ 的 ε 进行加权, 加权因子为 $q_j^t = 1 - p(z_t \mid \eta_j)(z_t \in I_i, \eta_j \cap I_i \neq \varnothing) (j = 1, \cdots, n)$, 得到近似函数 $f_1(x_t), \cdots, f_n(x_t)$;

对于高斯假设, 根据高斯分布函数对 $f_m(x_t)$ 的 ε 进行加权:

$$q_m^t = \mathrm{e}^{-[(t-\mu)^2/2\sigma^2]} \tag{6}$$

综上所述, 对于每一个较大的分块 $LS_j (j = 1, \cdots, h)$ 和交叉部分 $I_k (k = 1, \cdots, g)$, 我们得到 m 个函数 $f^*_1(x_t), \cdots, f^*_m(x_t)$. 接下来, 我们计算式(4)中每一个函数的权值, 完成整个模型.

2.4 求解组合模型中的权值 p_i^t

对于每一个较大的分块 LS_j, 可以使用单一的 SVR 来建模, 所以它和式(4)中函数的加权无关. 在这部分, 仅考虑如何对交叉分块 I_i 建模.

在已得到的函数 $f^*_i(x_t), (i = 1, \cdots, m)$ 基础上, 需要找到最优的 p_i^t. 假设对于每一个 i, 使得 $p_i^t, t = 1, \cdots, N$, 都变化的很缓慢, 可以把这个求解过程看成以下的优化问题,

$$\min_{p_i^t} \sum_{t=1}^{N} (y_t - \sum_{i=1}^{m} p_i^t f^*_i(x_t))^2 + \lambda \sum_{t=2}^{N} \sum_{i=1}^{m} (p_i^t - p_i^{t-1})^2 \tag{7}$$

$$\text{subject to} \quad \begin{aligned} &\sum_{i=1}^{m} p_i^t = 1, \ t = 1, 2, \cdots, N \\ &0 \leqslant p_i^t \leqslant 1, \ \forall\, t = 1, 2, \cdots, N, \ i = 1, 2, \cdots, m. \end{aligned}$$

其中, N 代表在交叉部分 I_i 数据点的个数, λ 作为惩罚参数以保证 p_i^t 和 p_i^{t-1} 之间足够小, 反映出 p_i^t, $t = 1, \cdots, N$ 变化很慢[8].

3 实验及其结果评价

为了验证本文提出的基于组合 SVR 的非平稳时间序列建模方法的有效性, 分别以炼铁高炉含硅量数据和太阳黑子活动数据作为数据集设计完成了两个实验. 因篇幅限制, 这里只给出第一个实验.

使用 350 个连续的高炉硅含量的数据(如图 1)作为数据

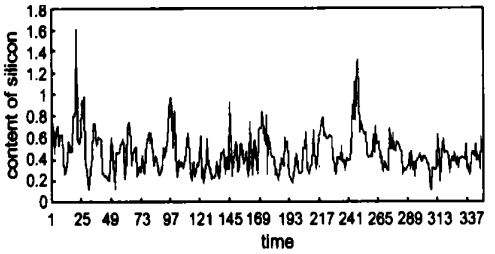

图 1　原始硅含量数据

集, 设计以下三组方案:

方案一: 对 350 个数据只用一个SVR 建模;

方案二: 通过模糊分块, 将 350 个数据分为两个较大的分块和两个交叉分块, 如图 2 所示. 对于 4 个部分分别用一个SVR 建模;

方案三: 通过模糊分块, 将 350 个数据分为两个较大的分块和两个交叉分块. 对于两个较大的分块分别用一个 SVR 建模, 与方案二的方法相同. 对于第一个交叉分块, 根据 2.3 所述, 得到 $m=3$, 即用 3 个函数建立组合的 SVR. 同理, 对于第二个交叉部分, $m=4$, 用 4 个函数建立组合的 SVR.

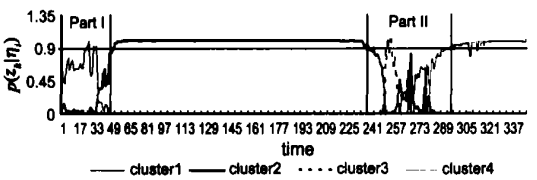

图 2　对原始硅含量数据进行模糊分块的结果

文中采用以下误差计算公式来评价结果的准确性:

$$NMSE=\frac{\sum_{j=1}^{N}(t_j-t_j^*)^2}{\sum_{j=1}^{N}(t_j-\bar{t})^2} \tag{8}$$

其中, N 为数据集中样本数据的数目, t_j 为某一点的真实值, t_j^* 为某一点的估计值, $\bar{t}$ 为所有点的估计值的平均.

方案一和方案二的实验结果如表 1 和图 3 所示, 从中可以看出方案一对四个分块估计的整体误差要小于方案二, 这是因为方案一用一个 SVR 对 350 个数据进行学习, 数据量较方案二充足, 学习得比较充分; 而方案二将 350 个数据分为 4 部分, 每一部分的数据量相对减少, 所以整体误差比方案一大.

表 1　方案一, 方案二和方案三的误差

	方案一 NMSE	方案二 NMSE	方案三 NMSE
第一个交叉部分	0.744613	0.785709	0.502015
第一个较大分块	0.710932	0.700227	0.700227
第二个交叉分块	0.553803	0.684760	0.414596
第二个较大分块	0.963554	0.928448	0.928448
四部分整体误差	0.623658	0.655586	0.535811

在表 1 中, 对方案一和方案二在四个分块中的的局部误差进行比较, 可以发现对于两个较大的分块, 方案二的误差 NMSE 比方案一的误差稍小一些, 这是因为两个较大分块中数据量相对较多, 样本学习得较充分, 同时数据的变化比较平缓, 其变化规律趋于一致; 而方案一中虽然样本数量更多, 样本学习得也很充分, 但对整个数据集(包括两个较大分块和两个交叉分块) 用一个 SVR 建模, 结果要受具有不同变化规律的数据的影响, 因此误差较方案二稍大一些. 而对于两个交叉部分, 由于数据变化剧烈, 规律不明显, 同时方案二中对应两个交叉分块的数据量较少, 所以方案一的误差明显小于方案二的误差. 图 3 中方案一和方案二的比较也说明了上述结论.

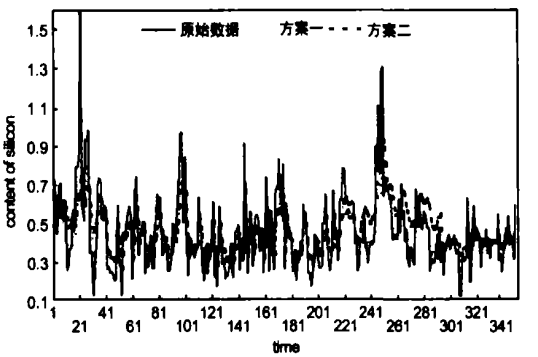

图 3　方案一和方案二对硅含量数据建模的整体比较

方案一与方案二实验结果的比较说明: 如果只是对数据进行分块处理, 从整体上和数据变化较剧烈的交叉分块来看, 结果反而不如单一建模, 而对于变化较平稳的较大分块, 结果也并无明显优势. 因此, 在对数据进行分块处理的基础上, 还应对交叉分块进一步加以处理, 交叉分块的数据量很小, 且变化规律不明显, 很难用单一的 SVR 学习到数据间潜在的关系. 从图 2 可以看出, 交叉部分的数据被相邻的多个模式影响, 同时还有它自己特有的属性, 因此本文提出用多个 SVR 组合的方法(即方案三中所用的方法) 对这些分块中的数据进行建模, 以提高整体和局部性能.

为了进一步说明组合 SVR 的有效性, 对于两个交叉分块, 将三种方案的实验结果进行比较, 分别如图 4 和图 5 所示, 图 4 表示出第一个交叉分块的 39 个数据点的对比结果, 图 5 表示第二个交叉分块的 62 个数据点的对比. 从这两个图和表 1 中可以看出, 对于两个交叉分块, 方案一的局部误差明显小于方案二的局部误差, 而方案三在这两块上的局部误差明显小于方案一的局部误差; 对于两个较大分块, 方案三与方案二均对应一个 SVR, 其模型及其局部误差是相同的, 均比方案一的局部误差小; 从整体上看, 方案三的整体误差也明显小于方案一、方案二的整体误差.

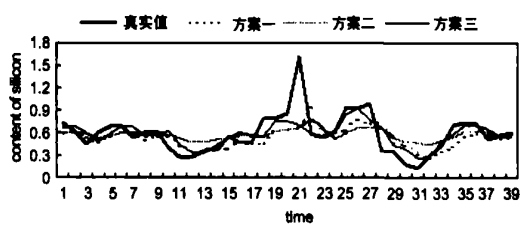

图 4　三种方案对第一个交叉分块硅含量数据建模的结果

4　结论

本文针对非平稳时间序列提出了一种新的建模方法. 这种方法将组合 SVR 和模糊分块信息结合起来. 首先, 对非平稳时间序列建模采取“分而治之”的办法, 采用模糊聚类算法

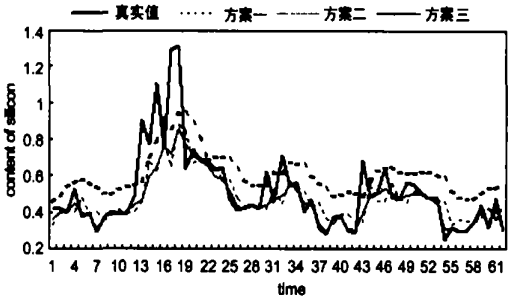

图 5 三种方案对第二个交叉分块硅含量数据建模的结果

进行模糊分块,并将各分块按照一定原则分为“较大分块”和“交叉分块”两种类型,针对每一种分块分别建立各自的模型.其次,提出一种改进的 SVR 模型,并根据模糊分块的特征,采用一种启发式的加权方法,用概率 $p(z_k|\eta_i)$ 作为对不敏感损失 ε 的启发权值来训练 SVR,得到估计函数 $f_i(x_t)(i=1,\cdots,n)$;对于交叉分块中的高斯假设,使 ε 的加权值符合高斯分布,得到函数 $f_m(x_t)(m=n+1)$.最后,使用这些函数在交叉分块中构建组合的 SVR 模型,并提出一种求解组合权值的方法,即通过定义一个目标函数,求解凸二次规划问题来找到每一个函数最优的权值 $p_i^t(t=1,\cdots,l,\ i=1,\cdots,m)$.实验表明本文提出的方法显著提高了交叉部分数据的估计值的精确度,从而改善了模型的整体性能.在以后的工作中,应进一步考虑在提高模型精度的同时,不显著增加模型复杂度.

参考文献:

[1] K Muller, A J Smola, G Ratsch, B Scholkopf, J Kohlmorgen, V Vapnik. Using support vector machines for time series prediction[A]. In Advances in Kernel Methods: Support Vector Learning[C]. Massachusetts: MIT Press, 1999. 243- 253.

[2] K Vasko, H Toivonen. Estimating the number of segments in time series data using permutation tests[A]. In Proc IEEE ICDM' 2002[C]. Maebashi: IEEE Computer Science Press, 2002. 466- 473.

[3] J Abonyi. Fuzzy Model Identification for Control[M]. Birkhauser: Boston, 2003. 1- 10.

[4] R J Hathaway, J C Bezdek. Switching regression models and fuzzy clustering[J]. IEEE Transactions on Fuzzy Systems, 1993, 1(3): 195- 204.

[5] J Abonyi, B Feil, S Nemeth, P Arva. Fuzzy clustering based segmentation of time-series[A]. In IDA 2003, LNCS 2810[C]. Berlin Heidelberg: Springer-Verlag, 2003. 275- 285.

[6] J Abonyi, R Babuska, F Szeifert. Modified gath-geva fuzzy clustering for identification of takagi-sugeno fuzzy models[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2002, 32(5): 312- 321.

[7] J Abonyi, B Feil, S Nemeth, P Arva. Modified Gath-Geva Clustering for Fuzzy Segmentation of Multivariate Time-series[OL]. http://www.fmt.vein.hu/softcomp/segment/Abonyi_segment.pdf.

[8] M W Chang, C J Lin, R C Weng. Analysis of nonstationary time series using support vector machines[A]. In Pattern Recognition with Support Vector Machines—First International Workshop, SVM 2002, LNCS 2388[C]. Berlin Heidelberg: Springer, 2002. 160- 170.

作者简介:

林树宽 女,副教授,1966 年 5 月出生于吉林省长春市,1988 年、1991 年分别在吉林大学、中科院沈阳计算技术研究所获理学学士和工学硕士学位,主要从事人工智能、机器学习等方面的研究. E-mail: linshukuan@ise.neu.edu.cn

支力佳 男,1977 年 11 月生于宁夏银川,1999 年毕业于湖南大学,现为硕士研究生,主要从事机器学习方面的研究.