第33卷 第10期 计 算 机 工 程 2007年5月
Vol.33 No.10 Computer Engineering May 2007
·博士论文· 文章编号：1000—3428(2007)10—0019—03 文献标识码：A 中图分类号：TP18

基于支持向量回归的批处理增量学习方法

王 玲，穆志纯，郭 辉
（北京科技大学信息工程学院，北京 100083）

摘 要：针对生产实际中数据批量增加的情况，为了提高所建立的模型准确性和模型更新问题，提出了一种基于支持向量回归的批处理增量学习方法。算法通过对钢材力学性能预报建模的工业实例进行研究，结果表明，与传统的支持向量机增量学习算法相比，提高了模型的精度，具有良好的应用潜力。

关键词：支持向量回归；批处理；增量学习

Batch Processing Incremental Learning Method Based on Support Vector Regression

WANG Ling, MU Zhichun, GUO Hui
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

【Abstract】 A new batch processing incremental learning method based on support vector machines is proposed to improve the model accuracy and update the model for the increasing batch data in the real work. The proposed method has been applied to a practical case of modeling prediction ability of mechanical property of steel materials. Compared with the traditional support vector machine incremental learning algorithm, the obtained model results demonstrate this promising method improves the model accuracy.

【Key words】 Support vector regression; Batch processing; Incremental learning

支持向量机(support vector machines, SVM)是20世纪90年代由Vapnik等人提出的一种新的学习机[1,2]，是统计学习理论中的结构风险最小化思想在实际中的一种体现。在解决小样本、非线性问题方面表现了良好的泛化能力，且不存在局部最优问题，已经在分类、时间预测、回归估计等领域得到了广泛应用。SVM用于非线性系统的回归估计，通常称为支持向量回归(SVR)[3,4]。它在非线性系统辨识、预测预报、建模与控制的潜在应用，使得对其研究显得非常重要。

对于一个预测模型而言，其性能很大程度上取决于所使用的训练样本。用于训练的样本越具有代表性，得到的预测性能越好。但在很多情况下，难以获得所有具有代表性的样本，常常需要采用增量学习技术， 即在利用已有训练样本完成学习后，对新获得的样本以增量的方式进行训练。但是，经典的SVM 学习算法并不直接支持增量式的学习。目前已有许多基于支持向量机的增量学习方法[5~7]用于分类问题中，而回归问题相对研究较少。本文讨论了一种基于支持向量回归的批处理增量学习算法，将其应用于钢材力学性能预报模型。由于其在新的训练中充分利用了历史的训练结果，从而显著地减少了后继训练的时间，同时提高了预测精度。

1 支持向量回归

支持向量回归[1]问题可以表述为：给定训练样本集为

$\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots, (\vec{x}_N, y_N)\}$

其中，$\vec{x}_i \in R^m$ 且 $y_i \in R$，N 为样本数。要求拟合的函数是

$$f(\vec{x}_k) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(\vec{x}_k, \vec{x}_i) + b \tag{1}$$

式中，α_i^*、α_i 是拉格朗日乘子，核函数 $K(\vec{x}_k, \vec{x}_i)$ 把输入向量映射到高维特征空间，b 为阈值。核函数采用径向基函数。

$$k(\vec{x}_k, \vec{x}_i) = \exp\left(-\frac{\|\vec{x}_k - \vec{x}_i\|^2}{2\gamma^2}\right)$$

通过引入两个松弛变量，不敏感损失函数 ε 和惩罚因子 C，再根据拉格朗日函数的极值满足条件，这样的回归问题可以转化成以下的二次优化问题：

$$\begin{aligned}\min_{\alpha,\alpha^*} w(\alpha,\alpha^*) = \min_{\alpha,\alpha^*} &\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\vec{x}_i, \vec{x}_j) \\ &-\sum_{i=1}^{N}(\alpha_i^* - \alpha_i)y_i + \varepsilon\sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\end{aligned} \tag{2}$$

约束条件：$\begin{cases}\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \\ \alpha_i, \alpha_i^* \in [0, C]\end{cases}$

解上述优化问题求得拉格朗日乘子，其中拉格朗日乘子为非零的训练样本点称为支持向量。

根据Karush-Kuhn-Tucker (KKT)条件[2]，可计算阈值

$$b = \frac{1}{2}\left[y_k + y_j - \left(\sum_{i=1}^{N}(\alpha_j - \alpha_j^*)k(\vec{x}_k, \vec{x}_j) - \sum_{k=1}^{N}(\alpha_k - \alpha_k^*)k(\vec{x}_k, \vec{x}_j)\right)\right] \tag{3}$$

2 基于支持向量回归的批处理增量学习算法

2.1 基于支持向量回归的批处理增量学习算法的设计思想

尽管 SVR 在很多方面都具有其它学习方法难于比拟的优越性，但随着样本数量的增多，SVR 缺乏对增量式学习的支持，当新增样本与训练好的数据集相差甚远时，模型的预测准确度将非常低。要想使模型具备增量学习能力，也就是

基金项目：国家“863”计划基金资助项目(2002 AA412010)；国家科技部资助攻关项目(2003EG113016)；北京市教委重点学科共建基金资助项目
作者简介：王 玲(1974－)，女，博士生，主研方向：人工智能，机器学习；穆志纯，教授、博导；郭 辉，博士生
收稿日期：2006-06-21 **E-mail：**linda_gh@sina.com

要求训练好了的模型不仅能准确地预测出新的数据，而且也不能忘记原来的知识。目前所采用的批样本处理算法，随着时间的推移，需要将一些样本数据摒弃掉。但可能导致离线学习时一些有用的样本被丢弃，而最新的但不太可能被再次经过该状态空间的数据却被加入到训练样本集中，这是不合理的。另一种可能是当对象的输出从一个区域到另一个区域时，训练集将丢弃一批刚才学习的数据，而加入另一批数据，这就可能导致这一次的数据破坏上次已学完的数据。

为了克服以上两种方法的缺点，本文对学习样本的选取作了一些改进。如果保存所有的训练样本，需要庞大的存储空间，从某种意义上而言也不是真正的增量学习算法，因此希望仅保存一部分重要的样本，使得在增量学习时不但只要较少的存储空间，而且也可以保持大部分的知识不被改变。而通过支持向量回归建模得到的支持向量集可以完全描述整个样本集的回归特性，支持向量集和训练样本集之间的等价关系可以得到证明[4]。通常支持向量集只是样本集的一部分，通过对支持向量集来研究增量学习是可行且有效的。首先建立一个后备样本数据集，其内容包括以前不同批次的支持向量集数据和当前运行时的最新数据，后备数据库一般比较大，当容量达到其上限时，将数据库中最前一个批次的支持向量集数据丢掉；然后用当前新批次的样本建立的模型去预测后备数据库中前几批次的数据，根据预测误差来确定前几批次的支持向量集的权重，重新建立新的增量模型。这样选出的训练数据能兼顾上述两种方法的优点。仿真实验表明该方法是相当有效的。

2.2 算法实现步骤

在实际应用中，数据通常是批量增加的，假设已经存在历史数据集按时间批分为$A_1,A_2,\ldots,A_w$，其中A_w为第w时刻的批样本集合。支持向量集按时间批次可表示为 $A_{sv}^1, A_{SV}^2,\ldots, A_{SV}^W$，其中 A_{SV}^W 为第w时刻的支持向量集。假设 $A_1 \cup A_2 \cup \ldots \cup A_w$，且$A_i \cap A_j = \phi$，i≠j。新增样本集$A_{w+1}$为第w+1 时刻的批样本集合。如果在所有样本集合(包括新增样本集)$A_1 \cup A_2 \cup \ldots \cup A_w \cup A_{w+1}$的基础上重新训练建立预测模型，需要花费大量时间和占用很大的内存。文献[9]充分利用了支持向量集与训练样本集的等价关系，采用在已经学习过的数据集所形成的支持向量集的基础上与新增数据一起构成新的训练集合进行增量学习的方法，即

$$SV_1 \cup A_2 \rightarrow SV_2 \cup A_3 \rightarrow \ldots \rightarrow SV_{w-1} \cup A_w \rightarrow SV_w \cup A_{w+1}$$

本文在此基础上进一步考虑到数据批对建模贡献的大小，采用了一种局部加权建模的方法。它的训练样本的选择是基于这样一种思想：只选择那些性质与预测目标相似的数据作为训练样本。在第 1 步，首先根据最新的一批数据A_{w+1}训练学习得到预测模型SVR_{w+1}，尽管不能认为它的预测效果很好，但是可以肯定随着时间的变化，目前的模型是最新的。现在，可以用SVR_{w+1}模型对历史数据集各批次的支持向量集进行预估测试，通过比较找出与最新的批样本集最相近的批次。换句话说，也就是哪个批次的误差越大，哪个批次就不可能成为训练样本。第 2 步，根据不同批次的误差信息选择训练样本建立最终的预测模型。为了有效地利用误差，这里采用了一种局部加权的策略，也就是说，对于预测误差大的批次，则赋予该批次中的支持向量较小的权重；而对于预测误差小的批数据集，则赋予该批次中的支持向量较大的权重。

具体步骤如下：

Step 1 初始化集合 F 为空集。

Step 2 用数据集$A_1,A_2,\ldots,A_w$作为训练集进行学习，分别得到模型$SVR_1,SVR_2,\ldots,SVR_w$以及对应的支持向量集 $A_{SV}^1, A_{SV}^2,\ldots, A_{SV}^W$。

Step3 将 $A_{SV}^1, A_{SV}^2,\ldots, A_{SV}^W$ 和增量数据集A_{w+1}添加到F中。

Step4 对增量数据集A_{w+1}进行训练建立预估模型SVR_{w+1}。

Step5 运用SVR_{w+1}模型对历史各批次的支持向量集进行预估测试，计算该模型对各批次的支持向量集的预测误差，根据误差的大小来确定各批次支持向量的权重，即

$$p_i(x) = 1 - 2.5 \cdot (error_i - 0.1)$$

其中，i 表示时间批次。根据上式可知，某批次的误差大于0.5，权重设置为 0，误差小于 0.1，权重设置为 1。

Step6 由以前各批次中不同权重的支持向量集和新加入的增量支持向量集共同建立的模型，增量学习结束。

3 工业实例研究

在实际生产过程中，热连轧过程是一个复杂过程，随着原料化学成分、炼钢参数、轧制参数等众多变量的变化，产品的力学性能质量指标(如断裂延伸率)很容易改变而发生偏移，而轧钢生产厂家和用户都希望通过提前预报产品的力学性能[10]来满足需求，及时调整生产工艺，提高产品质量。随着系统的不断运行，成批的新数据的不断增加，必须实时修正钢材性能预报模型，以达到所要求的质量标准。针对某大型钢铁厂生产的热轧产品，本文利用基于支持向量回归的批处理增量学习算法建立预报系统。该系统采用Q235B带钢不同批次的生产数据，考虑到原始化学成分和热轧生产工艺参数是影响成品最终力学性能的主要因素，输入值主要包括原始化学成分和生产工艺参数，其原始化学成分除了常规的 5 项化学成分是C、Mn、P、S、Si，还包括了Cr、Ni、Cu、Mo等多种合金元素，同时还包括了 3 种气体成分H、N、O。生产工艺参数主要包括精轧开轧温度、终轧温度、卷取温度和压下率。需要预报的性能对于热轧带钢来说主要是屈服强度、抗拉强度和延伸率。

3.1 预处理

首先对属性进行标准化变化：

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_i^j$$

$$\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_i^j - \overline{x}_i)^2$$

$$\overline{x}_i^j = \frac{x_i^j - \overline{x}_i}{\sigma_i}$$

其中，$\overline{x_i}$ 和 σ_i 分别是属性的平均值和标准差，x_i^j 和 $\overline{x}_i^j$ 分别是属性的原始数据和标准化之后的数据，j=1,…,n 是维数。

3.2 结果与讨论

以预报延伸率为例，建立“化学成分－延伸率”的SVR模型。假设历史数据集中的训练数据按照时间批次划分，共有 5 批样本子集A_1~A_5，增量样本按时间批次分别为A_6、A_8、A_8，将它们依次加入原有数据集。样本的分布见表 1。本文中支持向量机选用的是libSVM，核函数为径向基函数

$$K(x,y) = \exp\left(-(x-y)^2/(2\sigma^2)\right)$$

核参数 σ =0.01，正则化参数 C=100。本文选用渐进标准误差 (Asymptote Standard Error, ASE)评价模型的拟合程度和预测效果。

$$ASE = \frac{\sum_{i=1}^{l}(y_i - \hat{y}_i)^2}{l}$$

其中，l 表示测试集的样本数，$\hat{y}_i$ 表示预测值，y_i 表示实际值。

表 1 钢厂实际数据集两种不同算法的学习结果

基于支持向量回归的批处理增量学习算法			支持向量增量学习算法		
训练样本集	ASE	SV 数目	训练样本集	ASE	SV 数目
A_1[217]	0.011 216	183	A_1[217]	0.011 216	SV_1[183]
A_2[184]	0.011 659	170	SV_1+184	0.011 327	SV_2[302]
A_3[113]	0.011 048	100	SV_2+113	0.011 489	SV_3[365]
A_4[251]	0.010 15	230	SV_3+251	0.011 612	SV_4[578]
A_5[162]	0.010 129	142	SV_4+162	0.011 975	SV_5[660]
A_6[309]	0.057 963	275	SV_5+309	0.013 129	SV_6[915]
A_7[340]	0.012 355	306	SV_6+340	0.015 307	SV_7[1 023]
A_8[136]	0.011 051	120	SV_7+136	0.015 523	SV_8[1 096]

表 1 列出了本文提出的算法和文献[9]中的算法的泛化结果和支持向量个数的比较。在基于支持向量回归的批处理增量学习算法的训练样本集这一列中，以A_8[136]为例，表示新增的第 8 批数据有 136 个样本。在支持向量增量学习方法[9]中的训练样本集这一列中，如SV_5+309 表示利用支持向量回归训练前 5 批样本得到的支持向量SV_5与新增的 309 个样本构成训练集。

表 1 显示的结果表明本文提出的学习算法优于文献[9]中的增量学习算法。以新增的批数据集A_8为例，图 1 和图 2 分别给出了本文提出的学习算法和文献[9]中提出的支持向量增量学习算法的钢材延伸率的模型估计结果。相比之下，对于钢材的断裂延伸率，本文算法建立的支持向量回归模型的估计值与实际值吻合得非常好，估计结果的精度很高，较好地反映了断裂延伸率的变化趋势，可见基于支持向量回归的批处理增量学习算法显示了很强的泛化能力。

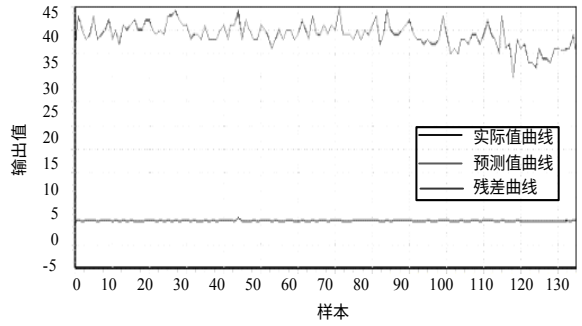

图 1 本文算法的断裂延伸率的模型预报与实际值的比较曲线

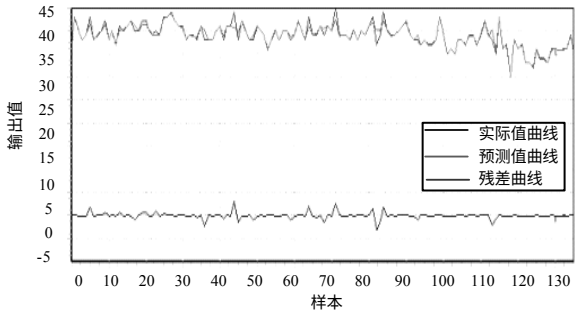

图 2 文献[9]算法的断裂延伸率的模型预报与实际值的比较曲线

4 结论

支持向量回归具有学习能力强、泛化能力好、对样本依赖程度低、模型参数确定方便等优点，但 SVR 缺乏对增量式学习的支持，本文研究了一种基于支持向量回归的批处理增量学习算法，实验结果表明，这种学习方法在精度要优于文献[9]中的支持向量机的增量学习方法，利用本文提出的算法进行钢材力学性能预报建模，取得了十分有效的应用结果。

参考文献

1 Vapnik N. The Nature of Statistical Learning Theory[M]. New York : Springer Press, 2000.

2 Vapnik V. Statistical Learning Theory[M]. New York: Willey, 1998: 21-22.

3 Trafalis T B, Inco H. Support Vector Machine for Regression and Applications to Financial Forecasting[C]//Proceedings of the IEEE INNS-ENNS International Joint Conference on Neural Networks. 2000-06: 348-353.

4 Tay F E H, Cao L. Application of Support Vector Machines in Financial Time Series Forecasting[J]. Omega, 2001, 29(4): 309-317.

5 Fung G, Mangasarian O L. Incremental Support Vector Machine Classification[R]. Wisconsin. Madison, 2001.

6 萧 嵘, 王继成, 孙正兴, 等. 一种 SVM 增量学习算法——ISVM[J]. 软件学报, 2001, 12(12): 1818-1824.

7 Ruping S. Incremental Learning with Support Vector Machines[C]//Proc. of ICDM. 2001: 641-642.

8 Drucker H. Neural Information Processing Systems[M]. Cambridge, MA: MIT Press, 1997.

9 Syed N A, Liu Huan, Sung Kah Kay. Incremental Learning with Support Vector Machines[C]//Proceedings of IEEE International Conference on Data Mining. 2001: 641-642.

10 Wang Yingjie, Tian Qingping. Metal Materials and Heat Treament [M]. China Railway Publish Company, 1999.

(上接第 12 页)

2 Estan C, Varghese G. New Directions in Traffic Measurement and Accounting[C]//Proc. of ACM SIGCOMM. 2002.

3 Abhishek K, Jun Xu. Space-code Bloom Filter for Efficient Per-flow Traffic Measurement[C]//Proc. of IEEE INFOCOM. 2004.

4 Fang Hao, Murali K. Fast, Memory-efficient Traffic Estimation by Coincidence Counting[C]//Proc. of IEEE INFOCOM. 2005.

5 Fang Wenjia, Larry P. Inter-as Traffic Patterns and Their Implications[C]//Proc. of IEEE GLOBECOM. 1999.

6 Vivek S. Zipf's L Z. Distribution: An Introduction[Z]. http://www.cs.unc.edu/~vivek/home/stenopedia/zipf/.