第18卷 第5期
2008年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.18 No.5
May 2008

基于相空间重构的汇率预测研究

王亚惠,谢维波
(华侨大学 信息科学与工程学院,福建 泉州 362021)

摘 要:汇率在宏观经济政策、商业经营和个人决策制定上的作用越来越重要,使其成为了研究的热点。根据混沌动力系统的相空间延迟坐标重构理论,基于支持向量机的强大的非线性映射能力,提出了一种基于支持向量机回归的超短期汇率预测方法,并建立了模型,对美元港币的即时汇率进行了实证计算,且与BP神经网络模型进行了比较。结果表明,所建立的模型能很好地跟踪即时汇率的变化趋势,预测精度比较高且算法运行速度比BP神经网络模型快得多。

关键词:超短期汇率预测;相空间重构;支持向量机;神经网络
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)05-0015-04

Prediction Research of Exchange Rate Based on Phase Space Reconstruction

WANG Ya-hui, XIE Wei-bo
(Information Technology and Engineering College of Huaqiao University, Quanzhou 362021, China)

Abstract: Exchange rate is hotspots because it becomes more and more important in macro economic policy, business operation and private decision. Presents a new method - support vector machines(SVM) for the prediction of supper-short-term exchange rate. The model is established for predicting exchange rate of dollar to Hongkong dollar and compared with the neural network predicting model. The result shows that the model above can track the supper-short-term exchange rate's change well. And it is better than the neural network model at both the accurate rate and the run-time length.

Key words: prediction of the supper-short-term exchange rate; phase space reconstruction; SVM; neural network

0 引 言

汇率在宏观经济政策、商业经营和个人决策制定上的作用越来越重要,这种重要性使汇率预测成为了国内外学者研究的热点。目前,汇率交易是全球全天24小时通过网络进行的,不同时段的交易获利不同。因此,对企业经营和个人炒汇来说,汇率即时数据的跟踪是十分有意义的。神经网络[1]具有良好的逼近能力,是目前非线性系统研究的热门工具之一。但由于神经网络的结构过于复杂且难以选择,需要估计的参数相对于较少的数据样本显得太多,导致所得到的神经网络模型相对于数据容易产生过拟合,泛化能力不够,而使其预测精度不高,在短期汇率预测应用中受到了限制。Vapnik等人根据统计学习理论提出的支持向量机学习方法[2],近年来受到了国际学术界的广泛重视,并且已经广泛用于解决分类和回归问题。支持向量机的最大特点是它服从结构风险最小化原理而非经验风险最小化原理。研究发现支持向量机的各项性能尤其是泛化能力好于传统的人工神经网络。目前讨论较多的是支持向量机在模式识别方面的性能与应用,而在预测方面的应用研究还相对较少。

基于以上问题,将支持向量机回归原理应用于汇率预测,提出了一种适合超短期汇率预测的模型,模型所采用的输入向量通过相空间重构成技术得到。应用SVM回归算法对美元港币的即时汇率数据进行仿真计算,并将该算法与传统的BP神经网络预测算法进行比较,从而验证文中提出算法的有效性。

1 预测模型的建立

1.1 相空间重构技术

相空间重构的关键在于嵌入维数 m 和时滞 τ 的确定,对于给定的时间序列,应该存在一个最优的 m 和 τ[3]。如果 τ 太小,则不能覆盖捕捉信号的动力学需要的最小时间距,m 将变得相当大;相反,如果 τ 大于最

收稿日期:2007-08-29
基金项目:福建省自然科学基金(A0540005)
作者简介:王亚惠(1982-),女,河北人,硕士研究生,研究方向为图像处理与模式识别;谢维波,硕士生导师,研究方向为信号处理、模式识别、智能信息处理。

佳值,作为结果的模型的性质变得太离散,会导致捕捉不到信号的动力学性质。

Gautama[4] 等人提出一个基于样本时间序列及其替代数据的相空间的微熵率方法。同步确定 τ 和 m。该方法主要的优点是用一个简单的测度同时优化 m 和 τ,避免了分别求取 m 和 τ 引起的不一致性。该方法的物理意义明显,实际效果较好,故文中采用此方法估计这两个嵌入参数的最优值:m_{opt} 和 τ_{opt}。

给定信号 $x(t)(t=1,2,\cdots,N)$ 的 N_s 个替代数据 $x_{s,i}(t), i=1,\cdots,m$,定义熵率(ER)为:

$$R_{ent}(m,\tau)=I(m,\tau)+\frac{m\ln n}{n} \tag{1}$$

其中:n 是延迟矢量数;$I(m,\tau)$ 为

$$I(m,\tau)=\frac{H(x,m,\tau)}{<H(x_{s,i},m,\tau)>_i} \tag{2}$$

而微熵 $H(x)=\sum_{j=1}^{N}\ln(N\rho_j)+\ln 2+C_E$,$N$ 是数据长度,ρ_j 是第 j 个延迟矢量与其最近邻点之间的欧氏距离,欧拉常数 $C_E\approx 0.5772$。熵率图上的最小值在 m 和 τ 轴上分别对应 m_{opt} 和 τ_{opt};然后,重构相空间

$$\boldsymbol{X}(k)=[x(k),x(k+\tau_{opt}),\cdots,x(k+(m_{opt}-1)\tau_{opt})]^{\mathrm{T}} \tag{3}$$

1.2 支持向量机回归

设有训练样本集 $\{(x_i,y_i)\}_{i=1}^{N}$,其中 $x_i\in R^n$,$y_i\in R$ 分别为输入值和对应的输出值,N 为样本的个数,R^n 和 R 分别为 n 维和1维实数空间。SVR采用如下的回归函数[5]:

$$y=f(x)=(\boldsymbol{w}\cdot\phi(x))+b \tag{4}$$

式中:$(\cdot)$ 为内积运算,b 为偏置项,$\phi(x)$ 是输入空间到高维特征空间的非线性映射,SVR就是将实际问题通过非线性映射转换到高维特征空间,并在这个空间中构造线性回归函数来实现原空间中的非线性回归函数,其特殊性质保证了SVR有较好的推广能力,同时也巧妙地解决了维数问题,使算法复杂程度与样本维数无关。系数 $\boldsymbol{w}$ 和 b 通过最小化下列泛函来估计:

$$R(\boldsymbol{w})=c\frac{1}{N}\sum_{i=1}^{N}e(f(x_i)-y_i)+\frac{1}{2}\|\boldsymbol{w}\|^2 \tag{5}$$

式中:$e(\cdot)$ 代表了经验风险,它通常用下式给出的 ε 不敏感损失函数来度量。

$$e(f(x)-y)=\begin{cases}0, & |f(x)-y|<\varepsilon\\ |f(x)-y|-\varepsilon, & |f(x)-y|\geqslant\varepsilon\end{cases}$$

$\frac{1}{2}\|\boldsymbol{w}\|^2$ 是正则化部分。通过控制 c 和 ε 两个参数,就可以控制SVM的泛化能力。c 为平衡系数(也称惩罚系数),是正常数,用来平衡经验风险和正则化部分,c 越大对数据的拟合程度越高,但是SVM的复杂度也会越大;ε 称为管子尺度,它决定了对训练样本拟合的精确程度,ε 值越大,支持向量数目就越少,因而解的表达就越稀疏。

为了寻找 $\boldsymbol{w}$ 和 b 需要引入松弛变量 ξ_i 和 ξ_i^*,使下式最小化:

$$\min\phi(\boldsymbol{w},\xi,\xi^*)=\frac{1}{2}\|\boldsymbol{w}\|^2+c\sum_{i=1}^{N}(\xi_i+\xi_i^*) \tag{6}$$

$$\text{s.t.}\begin{cases}y_i-(\boldsymbol{w}\cdot\phi(x))-b\leqslant\varepsilon+\xi_i^*\\ (\boldsymbol{w},\phi(x))+b-y_i\leqslant\varepsilon+\xi_i\\ \xi_i^*,\xi_i\geqslant 0\end{cases}$$

通过引入拉格朗日乘子 a_i 和 a_i^*,可以获得上式的对偶形式:

$$\max-\frac{1}{2}\sum_{ij=1}^{N}(a_i-a_i^*)(a_j-a_j^*)K(x_i,x_j)-\varepsilon\sum_{i=1}^{N}(a_i+a_i^*)+\sum_{i=1}^{N}y_i(a_i-a_i^*) \tag{7}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{N}(a_i-a_i^*)=0\\ a_i,a_i^*\in[0,c]\end{cases}\quad i=1,2,\cdots,N$$

其中 $K(x_i,x_j)=\phi_i\cdot\phi_j$ 为核函数,它是满足Mercer条件的任何对称的核函数对应于特征空间的点积。权值向量 $\boldsymbol{w}$ 为

$$\boldsymbol{w}=\sum_{i=1}^{N}(a_i-a_i^*)x_i$$

于是得回归函数 $f(x)$ 的表达式为:

$$f(x)=\sum_{i=1}^{N}(a_i-a_i^*)K(x_i,x)+b \tag{8}$$

从上面可以看出SVM回归算法在计算 $f(x)$ 时,无需计算权值向量 $\boldsymbol{w}$ 和非线性映射 $\phi(x)$ 的具体数值,而只需计算出拉格朗日乘子 a_i 和 a_i^* 以及核函数 $K(x_i,x_j)$ 即可,从而巧妙地解决了维数灾难问题,使得算法的复杂度与样本维数无关。同时通过选择合适的核函数,就可提高预测模型的精度,降低随机噪声对预测模型的影响。目前常用的函数主要有多项式函数、RBF函数、Sigmoid函数等,文中选择的是RBF核函数:

$$K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$$

其中 σ 为核函数的超参数,是预先选择的一个常数。

1.3 预测模型的建立过程

预测模型的建立过程如下[6]:

(1)给定时间序列 $x(t)(t=1,2,\cdots,N)$,利用1.1节的微熵率法求出 m_{opt} 和 τ_{opt},重构相空间作为SVM回归预测模型的输入矢量 $\boldsymbol{X}(N\times m$ 维):

$$\boldsymbol{X}(t)=[x(t),x(t+\tau_{opt}),\cdots,x(t+(m_{opt}-$$

$1)\tau_{opt})],t = 1,2,\cdots,N$

对应的输出矢量为 $\boldsymbol{Y} = x(t)'(t = 1,2,\cdots,N)(N \times 1$ 维)。

(2) 用 Cross - validation 法选择核函数的超参数 σ 和调节参数 c 与 ϵ 的值。

(3) 数据标准化处理:对输入向量 $\boldsymbol{X}$ 和输出向量 $\boldsymbol{Y}$ 进行标准化处理:

$$\boldsymbol{X}(k,i) = \frac{\boldsymbol{X}(k,i) - \text{mean\_}x(i)}{\text{std\_}x(i)}$$

$$\boldsymbol{Y}(k) = \frac{\boldsymbol{Y}(k) - \text{mean\_}y}{\text{std\_}y}$$

$k = 1,2,\cdots,N \quad i = 1,2,\cdots,m_{opt}$

其中 mean_ $x(i)$,std_ $x(i)$ 分别是输入向量 $\boldsymbol{X}$ 的第 i 列的算术平均值和标准方差;mean_ y 和 std_ y 分别是输出向量 $\boldsymbol{Y}$ 的算术平均值和标准方差。

(4)对输入输出数据进行训练建立模型,即利用式(8)求解模型的参数。

2　预测实例

文中的实验数据是从中国银行福建分行网页获取的实时数据[7],选取数据 2006 年 5 月 26 日 13:55 至 18:55 的 516 条美元港币即时汇率(如图 1 所示),前 500 条数据用于建立预测模型,后 16 条数据用于模型预测效果检验。采用微熵率方法求解时间序列的最优嵌入维数和最佳时滞。求得的熵率图如图 2 所示。求得的最优嵌入维数和最佳时滞为(2,6)。

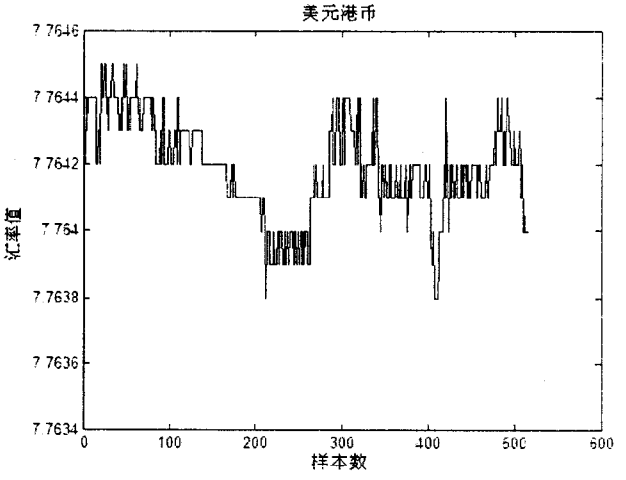

图 1　原始样本时间序列

在 SVM 预测模型中,取参数 $c = 10,\epsilon = 0.001$,核函数取为:

$$K(x_i,x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}),\sigma^2 = 0.75$$

输入矢量为:

$$\boldsymbol{X}(t) = [x(t),x(t+6)] \quad t = 1,2,\cdots,N$$

预测结果如图 3 所示,可见所建立的模型能很好地跟踪即时汇率的变化趋势。为了说明 SVM 的预测效果,分别采用 BP 神经网络对美元港币即时汇率数据进行训练与预测。与支持向量机一样,把前 500 个数据样本用于训练,后 16 个数据样本作为预测效果检验样本,汇率真实值与 BP 神经网络预测值的对比如图 4 所示。

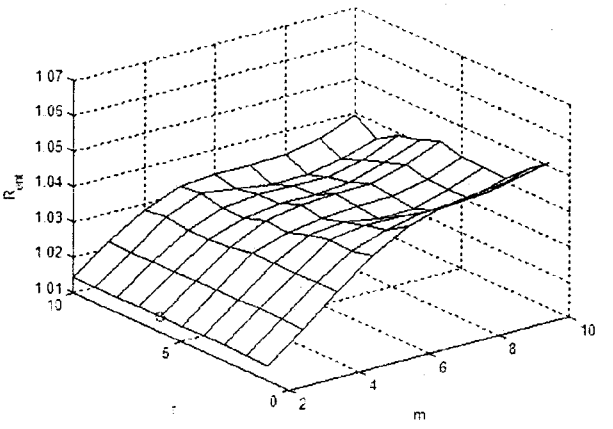

图 2　即时汇率熵率图

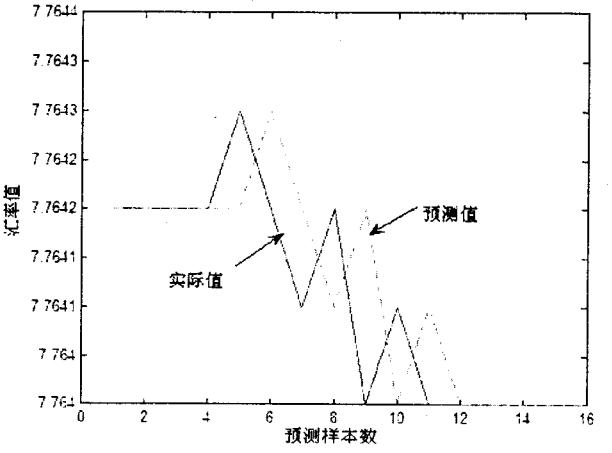

图 3　汇率实际值与 SVM 回归拟合曲线的比较

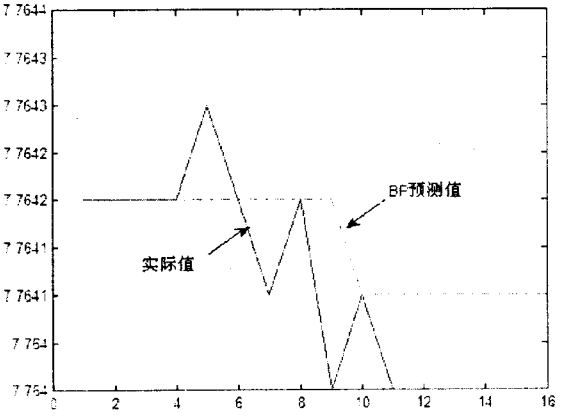

图 4　汇率实际值与 BP 神经网络拟合曲线的比较

为了更好地反映预测模型拟合效果和预测效果,对构建的模型采用以下统计量(其比较见表 1):

平均绝对误差(MAE):$MAE = \frac{\sum | e_i |}{N}$

均方差误差(MSE):$MSE = \frac{\sum (e_i)^2}{N}$

均方根误差(RMSE):$RMSE = \sqrt{MSE}$

平均相对误差(MAPE):$MAPE = \frac{1}{N} \sum | \frac{e_i}{s_i} |$

算法运行时间 RUNT:为预测算法开始到结束的时间间隔。

其中 e_i 为汇率实际值与预测值的绝对误差,s_i 为汇率实际值。$i = 1:16, N = 16$。

表 1 各项性能指标比较

各项性能指标	BP 神经网络	SVM 回归
平均绝对误差(MAE)	6.8750e-005	5.0000e-005
均方差误差(MSE)	8.1250e-009	6.2500e-009
均方根误差(RMSE)	9.0139e-005	7.9057e-005
平均相对误差(MAPE)	4.5280e-007	1.0062e-007
算法运行时间 RUNT	400s	≈1s

仿真结果表明,SVM 回归算法的预测效果良好,能很好地跟踪即时汇率的变化趋势,预测精度高。与 BP 神经网络相比 SVM 回归算法可以提供更好的泛化能力,且运行速度要快的多。

3 结束语

汇率预测的研究一直是国内外学者研究的热点。超短期汇率预测的研究对汇率日常交易具有十分重要的意义。针对目前流行的神经网络预测模型不能适应超短期预测的要求,文中将支持向量机回归应用于即时汇率的预测中,获得了良好的拟合与预测精度。但是如何更加合理地选择各模型中的参数仍需进一步研究。

参考文献:

[1] Maguire L P, Roche B, Mcginnity T M. Predicting a chaotic time series using a fuzzy neural network[J]. Information Sciences,1998,112:125-136.

[2] Vapnik V N. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.

[3] Grassberger P, Procaccia I. Characterization of Strange Attractors[J]. Phys Rev Let,1983,50(5):346-349.

[4] Gautama T, Mandic D P, Vanhulle M M. A differential entropy based method for determining the optimal embedding parameters of a signal[C]// Proc of the Int Conf on Acoustics, Speech and Signal Processing. Hong Kong: [s. n.], 2003:29-32.

[5] Schreiber T, Schmitz A. Surrogate time series[J]. Physica D, 2000,142(3-4):346-382.

[6] 崔万照, 朱长纯. 混沌时间序列的支持向量机预测[J]. 物理学报,2004(10):3303-3309.

[7] 黄巧玲,谢维波. 超短期汇率的预测研究[J]. 计算机应用, 2007(4):1009-1012.

(上接第 14 页)

统计结果为 2.1%,可以暂时不予处理。

* 地名后缀问题。

如:珠海市尼科希电子有限公司→珠海//市尼科希//电子//有限公司。解决问题的办法是建立地名后缀库,如"市"、"区"、"省"、"镇"等加入库中。

4 结 语

由于文中涉及切分需求的特殊性,目前的分词系统针对该领域分词效果不理想,所以文中定制了一个基于合并策略的未登录词识别方法,实验表明,在针对机构名切分这个特定领域,文中分词系统有更好的切分性能。另外,在一些短语类汉语切分中,该方法同样有效,前提是增加相关的关键词库[9]。

参考文献:

[1] 吴 栋,滕育平. 中文信息检索引擎中的分词与检索技术[J]. 计算机应用,2004,24(7):128-131.

[2] 朱巧明,李培峰,吴 娴,等. 中文信息处理技术教程[M]. 北京:清华大学出版社,2005:193-196.

[3] Lai B Y, Sun M S. Chinese word segmentation and part-of-speech tagging in one step[C]// Proceedings of International Conference: Research on Computational Linguistics. Taipei: [s. n.], 1997:229-236.

[4] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000:4-10.

[5] Chen Keh-jiann, Chen Chao-jan. Knowledge extraction for identification of Chinese organization names[C]// In proceeding of the 19th International Conference on Computational Linguistics. Taipei:[s. n.],2002.

[6] Knuth D E, Morris J H, Pratt V R. Fast pattern matching in strings [J]. SIAM Journal on Computing, 1977,6(2):323-350.

[7] 张华平,刘 群. 基于 N-最短路径方法的中文词语粗分模型[J]. 中文信息学报,2002,16(5):1-7.

[8] 林碧英,赵 锐,陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展,2007,17(5):184-186.

[9] 钟良伍. 基于中文机构名简称的检索方法研究 [D]. 北京:清华大学信息科学技术学院,2005.