图书文献经费的支持向量机组合预测

丁报林[1],倪天权[2]
DING Bao-lin[1],NI Tian-quan[2]

1.扬州环境资源职业技术学院,江苏 扬州 225009
2.中国船舶重工集团 723 研究所,江苏 扬州 225001
1.Yangzhou Vocational College of Environment and Resource,Yangzhou,Jiangsu 225009,China
2.No.723 Research Institute,Chinese Ship Industry Corporation,Yangzhou,Jiangsu 225001,China
E-mail:ntq72356@sina.com

DING Bao-lin,NI Tian-quan.Library budge forecast based on SVM combining forecasting.Computer Engineering and Applications,2009,45(6):221-223.

Abstract: The grey system forecasting model,neural network forecasting model and SVM(Support Vector Machine) forecasting model are proposed in this paper.Taking library budge of a library from year of 1996 to 2003 as a study case,the forecasting results are gotten by three methods.From the forecasting results,it is concluded that the accuracy of the SVM forecasting method is higher.Analyzing the characteristic of combining forecasting method,based on grey system forecasting model,neural network forecasting model and SVM,the linear combining forecasting model and SVM combining forecasting model are set up.Compared with single prediction methods and linear combining forecasting method,the accuracy of the SVM combining forecasting method is higher.
Key words: grey system;neural network;Support Vector Machine(SVM);combining forecasting;library budge

摘 要:对灰色、神经网络和 SVM(支持向量机)的 3 个预测模型进行了研究,以某图书馆 1996 年~2003 年图书文献总经费为例,对图书文献总经费进行了预测,经过比较,SVM 的预测方法精度较高。在分析组合预测特性的基础上,提出了对灰色系统、神经网络和 SVM 三种预测方法结果进行了线性组合预测方法和 SVM 的组合预测方法。与单一预测方法结果和线性组合预测进行对比,SVM 组合预测方法比较精确。
关键词:灰色系统;神经网络;支持向量机;组合预测;图书文献总经费
DOI:10.3778/j.issn.1002-8331.2009.06.064 **文章编号:**1002-8331(2009)06-0221-03 **文献标识码:**A **中图分类号:**U491

1 引言

图书馆担负收集情报、传播知识、为社会提供信息服务的职责。作为高等教育三大支柱之一,图书文献经费支出已占学校经费相当的比例。如何预测未来几年的经费需求,提出较合理的图书馆经费预算,是每个图书馆负责人较为难的工作。图书文献经费支出的影响因素众多,尤其是在学校大规模扩大招生和图书成本不断上涨的情况下,许多影响的因素不可预知,用常规方法很难估算来年的图书文献经费。文[1]采用灰色理论对图书文献经费进行了预测,文[2]利用 BP 神经网络对图书文献经费进行了预测。组合预测方法是由 Bates.J.M.和 Granger.C.W.J.于 1969 年首次提出的实用预测方法[3]。它是对同一个预测对象采用不同的单项预测模型,以适当的加权平均形式在某个准则下求得最优加权系数,从而可以充分利用各种单项预测方法所提供的信息,达到提高预测精度的目的。能增强预测的稳定性,具有较高的适应未来预测环境变化的能力。组合预测已经取得了很大发展,因而引起了众多学者浓厚的研究兴趣。本文分别采用灰色预测、神经网络预测和支持向量机预测,最后对 3 种方法进行了线性组合预测和支持向量机组合预测,并对它们精度进行了比较。

2 三种预测方法

2.1 灰色预测

将序列建成具有微分、差分、近似指数律兼容的模型,称为灰色建模[4]。1 阶 1 个变量的 GM 模型记为 $GM(1,1)$,若有序列 $x^{(0)}$:

$$x^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$$

对 $x^{(0)}$ 进行累计和:$x^{(1)}=AGOx^{(0)}$,则方程

$$\frac{\mathrm{d}x^{(1)}}{\mathrm{d}t}+ax^{(1)}=b \tag{1}$$

如果令 $\hat{\boldsymbol{a}}$ 为参数向量,即 $\hat{\boldsymbol{a}}=(a,b)^{\mathrm{T}}$,则在最小二乘准则下,有 $\hat{\boldsymbol{a}}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}y_N$,$GM(1,1)$ 的下述形式

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60310213)。
作者简介:丁报林(1972-),男。
收稿日期:2008-01-15 **修回日期:**2008-06-10

$$\begin{cases}\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\dfrac{b}{a}\right)e^{-ak}+\dfrac{b}{a}\\ \hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)\end{cases} \tag{2}$$

称为 $GM(1,1)$时间响应式,上标 $\hat{}$ 表示计算值。某图书馆 1996 年~2003 年图书文献总经费和灰色预测数据见表 1。

表 1 某图书馆 1996 年~2003 年图书文献总经费和 3 种方法预测数据 万元

时间	实际	灰色预测 GM		神经网络预测 NN		支持向量机预测 SVM	
		预测值	相对误差/(%)	预测值	相对误差/(%)	预测值	相对误差/(%)
1996	98.4	98.40	0.00	101.20	2.85	98.40	0.00
1997	97.7	95.02	2.75	102.62	5.03	97.68	0.02
1998	97.6	98.01	0.42	104.06	6.61	98.04	0.45
1999	99.0	101.09	2.11	105.52	6.59	99.48	0.48
2000	103.9	104.27	0.36	107.02	3.00	102.00	1.83
2001	105.6	107.55	1.85	108.53	2.78	105.60	0.00
2002	111.6	110.94	0.59	110.07	1.37	110.27	1.19
2003	116.0	114.43	1.36	111.64	3.76	116.00	0.00
2004		118.03		112.95		122.78	
2005		121.74		114.56		130.60	
2006		125.57		116.19		139.43	
2007		129.52		117.84		149.26	
2008		133.60		119.51		160.07	

2.2 神经网络预测

人工神经网络(NN)[5]是一类模拟生物神经系统结构,由大量处理单元组成的非线性自适应动态系统。神经网络是通过把问题表达成单元间的权来解决问题的。这里采用三层前向神经网(BP),神经网络输入神经元有 1 个,隐层有 3 个神经元,1 个输出神经元。时间作为神经网络的输入向量,需要预测的量为输出(变换到[0,1]区间内),然后用前 10 年的数据训练这个网络,使不同的输入得到不同的输出。学习过程由正向传播和反向传播组成,正向传播过程中,输入信号从输入经隐层单元逐层处理,并传向输出层,每一层神经元的状态只影响下一层神经元的状态。如果在输出层不能得到期望的输出,则转入反向传播,将输出信号的误差沿原来的连接通路返回,通过修改各层神经元的权值,使得误差最小。B-P 及其改进算法较多,这里不再详述。由 B-P 算法就可得到训练好的权值和阈值,调用这些权值和阈值,通过正向传播,输出值就是预测值。经过训练好的网络所持有的权系数和阈值, 便是预测值与时间的内部表示。转移函数选的是 S 型函数,动量因子为 0.2,学习率为 0.5,允许误差为 0.01。1996 年~2003 年图书文献总经费的预测数据见表 1。

2.3 支持向量机(SVM)预测模型

支持向量机(SVM)具有完备的统计学习理论基础和出色的学习性能,是一类新型机器学习方法,已成为机器学习界的研究热点,并在如人脸检测、手写体数字识别、文本自动分类、多维函数预测等领域都取得了成功应用。SVM 解决回归问题的基本原理如下[6]:设训练样本集

$D=\{(x_i,y_i)|i=1,2,\cdots,l\},x_i\in R^n,y_i\in R$

首先介绍线性回归问题,线性回归方程为:

$$f(x)=\langle \boldsymbol{w},\boldsymbol{x}\rangle+b \tag{3}$$

常用的损失函数有 ε-insensitive 损失函数,Quadratic 损失函数,Huber 损失函数和 Laplace 损失函数等。这里采用 ε-insensitive 损失函数,其形式如下:

$$L_\varepsilon(y)=\begin{cases}0 & \text{当}|f(x)-y|<\varepsilon\\ |f(x)-y|<\varepsilon & \text{否则}\end{cases} \tag{4}$$

SVM 解决回归问题,可转化为求解下列数学规划问题:

$$\begin{aligned}&\min_{w,b,\xi_i,\xi_i^*}\ \Phi=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\ &\text{s.t.}\ ((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)-y_i\leqslant\varepsilon+\xi_i\quad i=1,2,\cdots,l\\ &\qquad y_i-((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)\leqslant\varepsilon+\xi_i^*\quad i=1,2,\cdots,l\\ &\qquad \xi_i,\xi_i^*\geqslant 0,i=1,2,\cdots,l\end{aligned} \tag{5}$$

其对偶问题为:

$$\begin{aligned}&\max_{\alpha,\alpha^*}\ W=-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle+\\ &\qquad\sum_{i=1}^{l}[\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)]\\ &\text{s.t.}\ \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ &\qquad 0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,2,\cdots,l\end{aligned} \tag{6}$$

解规划式(6),得到拉格朗日乘子 α_i,α_i^*,则回归方程式(3)中的系数为:

$$\boldsymbol{w}=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\boldsymbol{x}_i \tag{7}$$

对于非线性回归,回归方程为:

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

求解下列规划问题:

$$\begin{aligned}&\max_{\alpha,\alpha^*}\ W=-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)+\\ &\qquad\sum_{i=1}^{l}[\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)]\\ &\text{s.t.}\ \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ &\qquad 0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,2,\cdots,l\end{aligned} \tag{9}$$

式中 $K(x_i,x)$为核函数,核函数可以有不同的形式,如多项式核、高斯径向基核、指数径向基核、多层感知核、样条核等。这里把时间序列作为输入,实际值作为输出。核函数 $K(x_i,x)$采用高斯径向基核函数:

$$K(x_i,x)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}\|}{2\sigma^2}\right) \tag{10}$$

参数设置如下:C=1 000,ε=0.1,σ=18, 采用支持向量机 Matlab 工具箱(SVM 工具箱网址:http://www.isis.ecs.soton.ac.uk/isystems/kernel/)。1996 年~2003 年图书文献总经费的预测数据见表 1。表 2 显示了各种方法的误差的平方和、平均相对误差和

表 2 3 种预测方法的误差的平方和

方法	灰色预测	神经网络预测	SVM 预测
误差平方和	18.598	155.879	5.797
平均相对误差/(%)	1.180	4.000	0.500
最大相对误差/(%)	2.750	6.610	1.830

最大相对误差，从表 2 可以看出,SVM 预测方法精度较高,效果较差的是神经网络方法。灰色预测实质上是指数预测,原数据具有指数递增特性时,采用灰色预测效果比较好,反之,效果并不如意;神经网络预测方法具有收敛速度慢和容易陷入局部极值等缺点;而 SVM 预测方法具有速度快,精度高等优点,效果比较好。

3 组合预测模型

3.1 组合预测优点

预测模型有很多,任何一个模型都是对实际系统的简化和抽象,其所包含的变量和参数必定有所选择并十分有限。不同模型从不同的角度对系统进行模拟,往往各有特点。能否将各种不同的模型组合在一起,博采众长,达到更好的预测效果,是组合预测思想的出发点。组合预测方法[7]是通过求个体预测值的加权算术平均而得到它们的组合预测值。目前的研究已论证了组合预测的许多优点。例如,组合预测集结所有单一模型所包含的信息;用最少方差准则得到的组合预报,其误差方差不大于任一分量的误差方差等等。

3.2 线性组合预测

线性组合预测的基本原理如下:设原数据序列:$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$,第 i 种($i=1,2,\cdots,m$)预测方法的预测结果为:$\boldsymbol{X}_1=(x_{1i},x_{2i},\cdots,x_{ni})^{\mathrm{T}}$,$m$ 种预测的线性组合预测结果为:

$$\boldsymbol{Y}=w_1X_1+w_2X_2+\cdots+w_mX_m$$

组合预测误差平方和为:$Q=(\boldsymbol{X}-\boldsymbol{Y})^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{Y})$,为使 Q 达到最小,满足$\frac{\partial Q}{\partial w_i}=0$,$i=1,2,\cdots,m$,得到方程组:

$$\begin{cases} w_1\sum_{i=1}^{n}x_{i1}^2+w_2\sum_{i=1}^{n}x_{i1}x_{i2}+\cdots w_m\sum_{i=1}^{n}x_{i1}x_{im}=\sum_{i=1}^{n}x_ix_{i1} \\ w_1\sum_{i=1}^{n}x_{i2}x_{i1}+w_2\sum_{i=1}^{n}x_{i2}^2+\cdots w_m\sum_{i=1}^{n}x_{i2}x_{im}=\sum_{i=1}^{n}x_ix_{i2} \\ \cdots \\ w_1\sum_{i=1}^{n}x_{im}x_{i1}+w_2\sum_{i=1}^{n}x_{im}x_{i2}+\cdots w_m\sum_{i=1}^{n}x_{im}^2=\sum_{i=1}^{n}x_ix_{im} \end{cases} \tag{11}$$

这里把 3 种方法($m=3$)组合起来,预测结果为 $\boldsymbol{Y}=0.218\,4X_{GM}-0.059\,1X_{NN}+0.844\,6X_{SVM}$,图书文献总经费的预测数据见表 3。

3.3 基于支持向量机的组合预测

这里采用 SVM 的来进行非线性组合预测,把 m 种预测结果作为输入,需要预测的量作为输出。核函数 $K(x_i,x)$采用高斯径向基核函数,采用 Quadratic 损失函数,参数设置如下:$m=3$,$C=1\,000$,$\varepsilon=0.1$,$\sigma=1\,000$,采用支持向量机 Matlab 工具箱。图书文献总经费的预测数据见表 3。表 4 显示了各种组合预测方法的误差的平方和、平均相对误差和最大相对误差,从表 4 可以看出 SVM 组合预测的误差的平方和最小，平均相对误差和最大相对误差最小,精度较高。SVM 组合预测实质上是非线性组合预测,很显然其效果要比线性组合预测要好。

表 3 某图书馆 1996 年~2003 年图书文献总经费和 2 种组合预测的数据 万元

时间	实际	线性组合预测		SVM 组合预测	
		预测值	相对误差/(%)	预测值	相对误差/(%)
1996	98.4	98.62	0.22	98.40	0.00
1997	97.7	97.19	0.52	97.64	0.06
1998	97.6	98.06	0.47	98.01	0.42
1999	99.0	99.86	0.87	99.46	0.47
2000	103.9	102.60	1.25	102.99	0.87
2001	105.6	106.26	0.63	105.60	0.00
2002	111.6	110.86	0.67	110.27	1.19
2003	116.0	116.37	0.32	116.00	0.00
2004		122.80		122.77	
2005		130.12		130.57	
2006		138.32		139.35	
2007		147.39		149.11	
2008		1573.1		1598.2	

表 4 二种组合预测方法的误差的平方和

方法	线性组合	SVM 组合
误差平方和	4.090	2.966
平均相对误差/(%)	0.620	0.380
最大相对误差/(%)	1.250	1.190

4 结束语

依照过去 8 年的主要统计数据,图书文献总经费呈增长趋势。对图书文献总经费做出较为准确的预测,希望能对相关部门和人员把握运输市场或进行决策有所裨益。SVM 的训练问题本质上是一个经典的二次规划问题，它可避免局部最优解,并且有唯一的全局最优解。在核函数的选择上,本文采用高斯径向基核函数,其收敛速度快,具有全局收敛等特性。用 SVM 方法进行非线性组合预测,比传统的预测方法及神经网络方法有更高的计算精度,是一种很有价值的新方法。

参考文献:

[1] 王廷满,沈思.利用灰色系统理论预测图书文献经费[J].西安科技学院学报,2002,22(9):367-370.

[2] 张雪莲,李丽燕.BP 神经网络图书文献经费预测模型[J].情报杂志,2006(6):115-116.

[3] Bates J M.Granger C W J.Combination of forecasts[J].Operations Research,1969,20(4):451-468.

[4] 邓聚龙.多维灰色规划[M].武汉:华中理工大学出版社,1990:30-35.

[5] 张立明.人工神经网络的模型及其应用[M].上海:复旦大学出版,1994:32-47.

[6] Gunn S R.Support vector machines for classification and regression[D].Image Speech and Intelligent Systems Research Group,University of Southampton,1997.

[7] 唐小我,曹长修.组合预测方法研究[J].控制与决策,1993,8(1):35-38.