一种基于SVR的综合预测方法及应用

张云龙1,2,　潘　泉1,　张洪才1

(1. 西北工业大学自动化学院，陕西西安 710072；2. 空军第一航空学院基础部，河南信阳 464000)

摘　要：针对一类因变量具有复杂自变量、且不具备相同采样周期的预测问题，综合运用支持向量回归估计(SVR)、多元回归和主成分分析等多种数据分析技术，提出了一种综合预测方法，建立起了飞机故障率与其错综复杂的影响因素间的一种数学关系，并且采用航空装备质量控制的统计数据对所提出的方法进行了实验，预测结果显示了方法的有效性。在影响因素量化过程中，还引入了Pearson相关系数方法。

关键词：SVR；多元回归；主成分分析；飞机故障率；综合预测

中图分类号：TB114　　文献标识码：A　　文章编号：1009-3516(2005)03-0019-03

预测方法可概括为3类:对比预测、惯性预测和相关分析预测[1-2]。虽然相关分析预测方法需要建立复杂的数学模型,但由于它能够找出预测对象变化的原因,所以受到越来越多的重视。这类方法大多采用单一的线性或非线性回归方程来建立预测模型[1],且一般都要求因变量与自变量有相同的采样周期和相同的数据长度,变量个数越少越好。针对飞机故障率、完好率、任务成功率等技术指标都有繁杂的影响因素,且采样不同步的实际特点,集成SVR、多元回归和主成分分析等多种数学方法,提出了一种基于SVR的综合相关分析预测方法,预测结果验证了这种方法的有效性。

1　预测方法与步骤

1.1　基本思想

把自变量(影响因素)分为两类,一类比因变量(预测变量)的采样周期密集,而另一类则与因变量的采样周期相同,对于比因变量采样周期稀疏的自变量,按与因变量相同的采样周期提取数据。预测过程分两步完成:先用SVR,对密集采样的自变量进行初步预测,将所得到预测结果按因变量采样周期内求和,得到包含这些所有密集采样因素信息的一个或多个与因变量具有相同采样周期新变量 $\hat{y}^{(i)}_{\mathrm{SVR}}(i=1,2,\cdots,m_1)$,使得其余的自变量连同新变量都与因变量具有相同的采样周期。再采用多元回归或其它非线性方法,将其余的自变量和新变量一起与因变量进行回归分析,得到因变量的一种综合预测。为了提高预测精度、减小变量间的耦合作用、并减少变量的个数,还将集成变量的归一化、标准化和主成分分析等多种数学方法。

1.2　基于SVR的初步预测

Vapnik于1995提出的支持向量机(SVM)学习方法[3],较好地解决了小样本二值分类问题,得到了广泛的应用。但对于连续回归问题,这种方法却需要产生数目庞大的支持向量(SV)。为此,提出了支持向量回归估计(SVR)方法[4],采用一种称为 ε -不敏感(ε -insensitive)的损失函数:

$$|y-f(\boldsymbol{X},\delta)|_\varepsilon=\begin{cases}0 & (|y-f(\boldsymbol{X},\delta)|\leqslant\varepsilon)\\ |y-f(\boldsymbol{X},\delta)|_\varepsilon & (\text{otherwise})\end{cases} \tag{1}$$

收稿日期:2004-05-19
基金项目:国家自然科学基金资助项目(60172037)
作者简介:张云龙(1966-),男,河南宁陵人,博士生,主要从事质量控制、信息处理等方面的研究;
潘　泉(1961-),男,上海人,教授,博士生导师,主要从事数据融合、小波分析等方面的研究;
张洪才(1938-),男,江苏江阴人,教授,博士生导师,主要从事估计理论、系统工程等方面的研究.

式中:$y=(w\cdot X)+b$ 为期望的输出(注:SVM 的 y 仅取右式的符号),w 为权阵,X 为自变量向量;$f(X,\delta)$ 为预测函数,$\delta\in\Lambda$ 为预测函数的参数;ε 为函数不敏感值。SVR 问题就是在训练样本集中,解$(w\cdot w^{T})\leqslant c_n$约束下的经验风险的最小化问题:$R_{emp}(w,b)=(\sum_{i=1}^{n}|y_i-(w\cdot X_i)-b|_{\varepsilon}^{2})/n$。

化成 Wolfe 对偶问题并解出最优参数 α^{*}。可得到 SVR 回归估计式为

$$y=\sum_{i=1}^{n}(\alpha_i^{*}-\alpha_i)k(X_i,X)+b \tag{2}$$

采用SVR 解决回归估计问题,首先确定三类自由参数:ε、C(正则化参数),以及核函数 $k(X_i,X)$的自身参数(如 RBF 函数的 gamma 值、多项式基的阶次 d 等)。SVR 参数选取方法以及回归精度评判见文献[5]。

对密集采样的自变量之中的某一组具有相同采样周期的自变量建立 SVR,将式(2)的正输出按因变量的采样周期求和,我们可以得到这些密集采样的自变量对因变量的一个预测值,称之为 SVR 初步预测 $\hat{y}_{SVR}$,这个预测值由于考虑的自变量不够全面,因而会有较大的误差,但可以用作下一步回归修正的一个自变量。这样做既可以把密集采样的影响因素信息集成到预测结果中,又可以同步采样周期。

1.3 多元回归修正

设 P 为自变量的总个数,对其中 p 个密集采样自变量,通过 SVR 已经得到了 m_1 个与因变量采样同步的 $\hat{y}_{SVR}^{(i)}$,其余 $P-p$ 个自变量$(X_{p+1},X_{p+2},\cdots,X_p)$则与因变量有相同的采样周期。设 $y=f(\hat{y}_{SVR}^{(i)},X_j)+e,i=1,2,\cdots,m_1,j=p+1,p+2,\cdots,P$。式中:$e$ 为零均值随机变量,对于 f,可用多元回归或其它非线性方法(包括 SVR 方法)进行进一步的回归分析,得到因变量的一种最终综合预测 $\hat{y}$。由于这个预测结果较全面地考虑了所有可量化的自变量因素,因而可以获得较高的预测精度。

1.4 模型检验与评价

采用误差均方差和小误差概率两个指标来检验和评价两级预测模型的精度。定义某一预测点的误差为 $e_i=y_i-\hat{y}_i$,则可以得到误差序列$\{e\}=\{e_1,e_2,\cdots,e_n\}$,用误差的均方值$\sqrt{(\sum_{i=1}^{n}e_i^{2})/n}$来表示预测精度。如果几种预测模型的误差相差不大,可用小误差概率来评价模型可信度[6]。小误差概率定义为 $p=p\{|e_i-\bar{e}|<0.674\,5S\}$。其中 $S=\sqrt{(\sum_{i=1}^{n}(y_i-\bar{y})^{2})/n}$。$p$ 越大,模型越可信。

2 飞机故障率预测实例

以飞机故障率的采样周期为“月”。通过分析,确定 34 个影响因素变量作为飞机故障率预测的自变量。为减少分析结果受自变量间量纲差别的影响,先对所有 34 个自变量用 $\tilde{x}_i=(x_i-\bar{x})/\sqrt{S_{xx}}$ 进行了标准化处理,使各自变量的均值为 0,方差为 1。

2.1 整理数据

数据整理的任务就是将原质量控制数据库中数据转换成预测所需要的格式。如将原来表示时间的分到小时的 60 进制转换为 10 进制浮点数、将原来用整数存储的起落次数转为用浮点数表示,以提高预测精度。数据整理的另一项任务则是对非数值数据的量化。以“发现时机”为例。原数据库中故障发现时机有“机械日”等共 19 种。统计各自在故障表中的出现频率,由专家论证,确定其中前 11 种取值(见表 1)为参考向量,记为 $X=(x_1,x_2,\cdots,x_{11})^{T}=(28.337\,4,\cdots,0.182\,0)^{T}$,统计各月对应的 11 种发现时机的频率作为比较向量,记为 $K=(k_1,k_2,\cdots,k_{11})^{T}$。采用 Pearson 相关系数 $r_{xk}=s_{xk}/(\sqrt{S_{xx}}\sqrt{S_{kk}})$,作为当月“发现时机”的量化值。其中 S_{xx}、S_{kk}和 S_{xk}分别为 X 与 K 的方差和协方差。

表 1 “发现时机”基准向量

类别	值
机械日	28.337 4
特定检查	24.393 2
换季检查	12.742 7
飞行中	9.890 8
预先机务准备	5.036 4
再次出动	4.672 3
定时检查	2.609 2
更换发动机	0.788 8
直接机务准备	0.728 2
日历检查	0.303 4
大检查	0.182 0

2.2 基于 SVR 的初步预测

以每架飞机的规定寿命、规定日历年等共 24 个与寿命有关的影响因素变量为自变量$(X_1,X_2,\cdots,X_{24})^{T}$。

考虑到这些变量的采样周期比飞机故障率的采样周期密集,且表现出较强的非线性,不便直接建立回归模型,以本架飞机从某一飞行日期到下一飞行日期之间发生的故障数为因变量 y 建立 SVR,将预测值按月求和,得当月所有飞机的故障数,称为 SVR 预测数 $\hat{y}_{SVR}$。

这 24 个自变量基本上都是与飞机寿命有关的因素,数据采样到每架飞机的每个飞行日。为保证各自变量之间的独立性,采用主成分分析[2]方法对它们进行了正交化处理,提取前 10 个(累计贡献率 93.1524%)进行训练和测试。为使 SVR 产生理想的输出,对这 10 个变量用 $\hat{x}_i = (x_i - \min x)/(\max x - \min x)$ 进一步做了归一化处理,取值限制在 0 ~ 1 之间。按月分组,将所有非当月的样本存入训练样本集文件,而将所有当月的样本存入测试样本集文件,共有 36 对训练/测试样本集。

经过反复测试,确定径向基(RBF)函数 $\exp(-0.01 \times \| x - x_i \|^2)$ 为 SVR 核函数,并取 $\varepsilon = 0.001$, $C = 100$。所得 $\hat{y}_{SVR}$ 与飞机故障率 y 的预测误差均方差为 1.024 3,小误差概率 $p_1 = 0.916\ 7$。

2.3　基于多元线性回归的预测修正

以 $\hat{y}_{SVR}$ 和故障的发现时机(的量化值)、月飞行强度等 10 个因素为自变量 $(\hat{y}_{SVR}, X_{25}, X_{26}, \cdots, X_{34})^T$,以当月所有飞机的故障数为因变量 y。考虑到它们都具有相同的采样周期,且与飞机故障率的非线性不强,采用多元线性回归法对预测结果进行修正,得到最后的预测结果 $\hat{y}$。使用最小二乘法,得到下面的多元回归修正方程:$\hat{y} = 0.5324 + 1.689\hat{y}_{SVR} + 0.0521\tilde{X}_{25} + 2.4593\tilde{X}_{26} + 0.0011\tilde{X}_{27} + 0.3452\tilde{X}_{28} + 1.0021\tilde{X}_{29} + 0.2301\tilde{X}_{30}$。其中 $\tilde{X}_{25} \sim \tilde{X}_{30}$ 为 $X_{25} \sim X_{34}$ 的前 6 个主成分(累计贡献率 91.2536%)。此步误差均方差为 0.524 2, $p_2 = 1.0$。

2.4　结果分析

某团某机型 2000 ~ 2002 3 年的实有故障数、SVR 预测值及回归修正预测值的对照结果见图 1。可以看出,提出的基于 SVR 的预测方法是可行的,再用多元最小二乘线性回归方法对预测结果做进一步的修正是有效的。分析预测结果与实际故障数误差产生的原因主要有:各影响因素中的异常值和观测噪声未进行滤波处理;部分飞机和发动机的规定寿命、规定日历年、大修次数等数据缺失,采用其它飞机或发动机的对应平均数据代替,导致误差;一些因素,如机务人员素质、宏观政策等因难以量化而被舍弃。

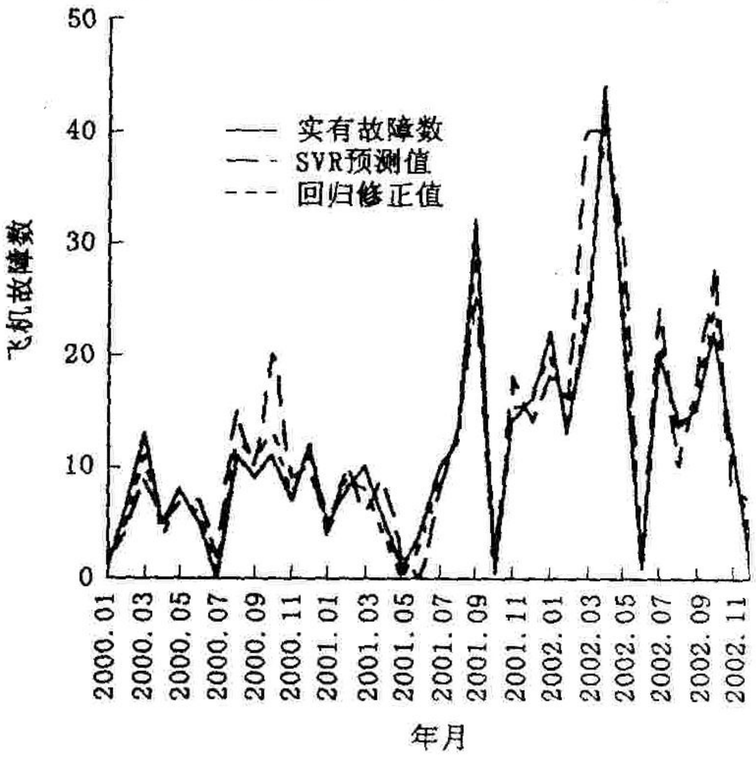

图 1　预测结果比较

3　结论

本文方法避免了一般预测方法面对复杂自变量和采样不同步可能带来的两个问题:一是参与建模的自变量数目多;二是对密集采样的自变量信息损失大。贡献还在于:集成多种数学方法,充分考虑不同采样周期自变量在其采样点上对因变量的贡献,因而保证了预测结果的客观性;首次为空军建立了飞机故障率与其错综复杂的影响因素的数学关系,为进一步提出减小故障率的方法和措施提供了理论依据。

参考文献:

[1] 徐国祥. 统计预测和决策[M]. 上海:上海财经大学出版社,1998.
[2] 张恒喜,郭基联. 小样本多元数据分析方法及应用[M]. 西安:西北工业大学出版社,2002.
[3] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer,1995.
[4] Vapnik V N. Statistical Learning Theory[M]. New York:Wiley, 1998.
[5] 张绍武,潘　泉. 基于支持向量机和贝叶斯方法的蛋白质四级结构分类研究[J]. 生物物理学报,2003,19(2):171 - 175.

(下转第 46 页)