新型 ε-不敏感损失函数支持向量诱导回归算法及售后服务数据模型预测系统*)

罗泽举　朱思铭

(中山大学数学与计算机科学学院　广州 510275)

摘　要　对含有噪声的数据序列根据预测置信度进行去噪处理,将训练集和测试集及预测数据共同作为训练向量集,以此建立新型支持向量诱导回归算法。本文利用该算法对实时售后服务的"千车故障数"进行了时间序列分析,并建立了新型的 ε-不敏感损失函数小样本模型预测系统。预测显示误差小于 5.3%的值占了总体的 98.1%,其预测置信度达到 0.983,与二次和 Huber 损失函数相比其 MAPE 值只有 2.3%。用计算机模拟仿真单批次预测显,当时间参量 $t\to+\infty$,"千车故障数"将收敛于定值 74.0601,这和实际相当吻合,表明所建预测模型的有效性。文章最后还和传统神经网络模型作了比较,说明新型 SVM 机比神经网络处理小样本能力更强。

关键词　诱导回归算法,售后服务,预测系统

A New ε-insensitivity Function Support Vector Inductive Regression Algorithm and After-sales Service Data Model Forecast System

LUO Ze-Ju　ZHU Si-Ming

(School of Mathematics Computational Science, Sun Yat-Sen University, Guangzhou 510275)

Abstract　To filter noises according to prediction confidence level in the sequence of data that contains noises, set the training set, prediction set and testing set as the training set, we set up a new support vector inductive regression algorithm. To analyze the real-time after-sales service data time sequence of "the number of thousand cars malfunction" by using this algorithm and set up a new support vector machines models forecast system based on small sample and ε-insensitivity function. The predict value whose error is less than 5.3% is 98.3% of total. Further more, the confidence level come to 0.981. Compareed with quadratic loss function and Huber loss function, the MAPE is only 2.3%. From the computer analog simulation the single batch, we find that when time parameter $t\to+\infty$, the "the number of thousand cars malfunction"will converge to fixed value 74.0601, this is correspond with the reality and show the model is very availability. In the end of this paper, contrasted with neural network, the new SVM is superior to traditional neural network in the capacity for handling small sample.

Keywords　Inductive regression algorithm, After-sales service, Forecast system

1　引言

产品质量是企业的生命线,售后服务是产品质量的观测点,如何用好售后服务的数据是现代企业管理的重要问题之一。

整车或某个部件的"千车故障数"是一个从售后服务中了解轿车等机动车质量的很重要的指标。由于从售后服务中了解信息是时滞的,若干年后返回的信息对于目前的生产已经没有多大作用,因此如何更科学地利用少量现有的数据预测未来情况是售后服务中非常重要的问题。

支持向量机(Support Vector Machines, SV-MS)[1,2]是一种基于小样本学习的新型模式识别方法。它采用结构风险最小化原则和核函数方法,克服了模式分类器的复杂性和应用性之间的矛盾,显示出极强的推广能力,是一门正在蓬勃发展的理论。它计算的复杂性并不在于维数的高低,相反,它正是克服了维数灾难,利用少数支持向量,解决了向量在高维特征空间的分类问题,亦即解决了向量在普通二、三维空间中线性不可分问题;支持向量机采用最优分类超平面将一类成员和非该类成员分开,当用支持向量进行回归分析时便成了回归支持向量机。

本文利用某企业 2002 年 1 月至 2003 年 12 月(2004 年 1 月前统计的)的最新数据资料,分析了近两年共 24 个月的"千车故障数",建立了一整套置信度较高、基于 ε-不敏感损失函数支持向量诱导回归算法的 SVM 预测系统,并对数据作了年级以上的长期预测。仿真结果显示数据序列随时间参数收敛于 74.0601,这是"千车故障数"的极限,是企业售后服务的重要参考值。

2　支持向量机回归模型(support vector machine regression models, SVMRM)

2.1　统计学习的基本回归算法

统计学习的目标是:寻找属于某个函数集中的函数 $f(x,\alpha)$,使得它在函数集 $f(x,\alpha)$, $\alpha\in\Lambda$ 上最小化下面的风险泛函:

$$R[\alpha]=\int L(y,f(x,\alpha))P(x,y)\mathrm{d}x\mathrm{d}y,\alpha\in\Lambda \qquad (2.1)$$

其中 $P(x,y)$ 为未知分布,$L(y,f(x,\alpha))$ 为损失函数;由于 $P(x,y)$ 未知,为了最小化(2.1)式,采用经验风险最小化原则(empirical risk minimization, ERM),将风险泛函 $R[\alpha]$ 替换为下面的经验风险泛函:

*)国家自然科学基金资助项目(No. 10371135)。罗泽举　博士研究生,研究方向为机器学习与模式识别,生物信息学。朱思铭　教授,博士生导师,主要研究方向为应用数学、常微分方程、计算机应用。

$$R_{emp}[\alpha]=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(x_i,\alpha)) \tag{2.2}$$

于是得到学习函数

$$f_l(x,\alpha_0)=\arg\min_{\alpha\in\Lambda}R_{emp}[\alpha] \tag{2.3}$$

对于损失函数，一般选取 ε 不敏感损失函数，二次损失函数，Huber 损失函数及 Laplace 损失函数。本文选用 ε 不敏感损失函数，因为这种函数由于有参数 ε 可调，因而可以提高对学习的泛化能力；ε 不敏感损失函数定义为：

$$L(y,f(x,\alpha))=|y-f(x,\alpha)|_\varepsilon$$
$$=\begin{cases}0 & ;|y-f(x,\alpha)|\leqslant\varepsilon;\\ |y-f(x,\alpha)| & ;\text{其它}\end{cases} \tag{2.4}$$

设数据集是 $(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)$，其中 $x_i\in R^n$，$y_i\in R,i=1,2,\cdots,l$，

回归函数是：

$$f(x,\alpha)=w\cdot x+b; \tag{2.5}$$

就是要寻找 w_0,b_0 使得在满足一定的条件下最小化下面的经验风险泛函：

$$R_{emp}(w,b)=\frac{1}{l}\sum_{i=1}^{l}|y_i-(w\cdot x)-b| \tag{2.6}$$

问题转化为最小化下式：

$$\Phi(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_i(\xi_i^*+\xi_i) \tag{2.7}$$

约束为

$$y_i-(w\cdot x_i)-b\leqslant\varepsilon+\xi_i^*,i=1,\cdots,l$$
$$(w\cdot x_i)+b-y_i\leqslant\varepsilon+\xi_i,i=1,\cdots,l$$
$$\xi_i^*\geqslant0,i=1,\cdots,l$$
$$\xi_i\geqslant0,i=1,\cdots,l \tag{2.8}$$

优化(2.7)式用 Lagrange 乘子法求下述函数 L 中自变量的鞍点：

$$L(w,\xi^*,\xi,\alpha,\alpha^*,C^*,\gamma,\gamma^*)=\sum_{i=1}^{l}(\xi_i^*+\zeta_i)-\sum_{i=1}^{l}\alpha_i[y_i-(w\cdot x_i)-b+\varepsilon+\xi_i]-\sum_{i=1}^{l}\alpha_i^*[(w\cdot x_i)+b-y_i+\varepsilon+\xi_i^*]-(C^*/2)(c_n-(w\cdot w))-\sum_{i=1}^{l}(r_i^*\xi_i^*+\gamma_i\xi_i) \tag{2.9}$$

(2.9)式转化为

$$\max_{\alpha,\alpha^*}[-\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)<x_i,x_j>+\sum_{i=1}^{l}\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)] \tag{2.10}$$

α,α^* 满足 $0\leqslant\alpha,\alpha_i^*\leqslant C,i=1,2,\cdots,l$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{2.11}$$

$\alpha_i\alpha_i^*=0$.(KKT 条件)

因而求得的回归参数为：

$$\tilde{\omega}=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i;\bar{b}=-<\bar{\omega},(x_r+x_s)>/2 \tag{2.12}$$

式中 x_r,x_s 为两类支持向量。

若采用内积回旋时，则相应的回归函数变形为：

$$f(x,v,\beta)=\sum_{i=1}^{l}\beta_iK(x,v_i)+b \tag{2.13}$$

其中 $\beta_i,i=1,\cdots,l$ 是标量，$v_i,i=1,\cdots,l$ 是向量，$K(u,v)$即为满足 Mercer 条件的核函数[3,4]。

2.2 新型诱导回归支持向量算法

我们可将传统数据训练的方法总结为：

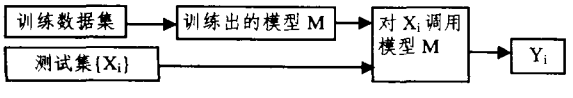


图 1　传统机器学习模型

用这种方法将训练集、确定集、测试集严格地分开，没有循环过程，而且对于含有大量噪声的数据不进行处理，再加上预测点和训练集离得较远，这样预测的效果往往不理想。如果在学习算法中根据置信度高低将训练集和测试集经过置信估计和去噪处理(就是去掉含有高噪声的样本)，再作训练集，这样的训练集由于降低了噪声，置信度高，从而预测出来的数据将更加可靠。经过本文的建模试验，数据预测准确率确实大大提高了。这是对传统训练算法的改进。新的算法模型是：

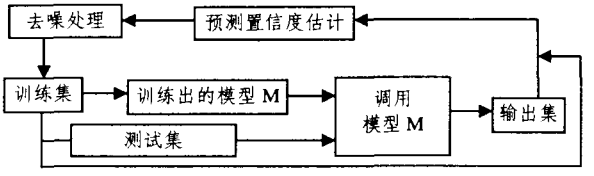


图 2　新型诱导回归算法机器学习模型

这种新型的诱导回归算法如下：

(1)用训练集训练出模型 M；

(1)对训练模型 M，用测试集进行测试，得到输出集；

(3)用训练集和测试集再加上输出集进行置信度估计；

(4)根据置信度大小进行去噪处理；

(5)重新划分去噪后的数据集，再用去噪后的数据集作训练集训练；

(6)转到(1)，如果预测误差小于某精度，算法终止；否则继续。

从上面的算法可以看出，经过若干步去噪循环处理后，其数据序列是高置信度的序列，也即是说序列中各点的预测误差是相当小的(预先给定误差范围)。

下面我们用上述方法进一步讨论关于售后服务中的“千车故障数”时间序列分析建模的过程。

3　售后服务预测模型的建立

3.1　模型中几个重要参数的过滤

模型中的几个重要参数为：

表 1　模型中的重要参数及其含义

参数类别	意义
$K(u,v)$	满足 Mercer 条件的实对称实函数
L	损失函数的类型
C	控制 α_i 的取值，影响支持向量和函数的 VC 维
ε	允许逼近函数的敏感程度

• 核函数。由于已知的数据点呈非线性相关性，因此可选取多种函数作为核函数，通过实验，发现选取线性核函数效果较好，体现了核函数和样本数据本身的固有规律。

• 损失函数 L 和敏感度 ε。选取二次损失函数的计算时间要比 ε 不敏感函数的计算时间少。这是由 ε 不敏感函数计算中的约束条件比二次损失函数的计算中的约束条件多引起的。但是这并不表示用二次损失函数计算精确度高，相反，选取 ε 不敏感函数和合适的 ε 值可以达到最理想的值，而二次损失函数是不可调整精度的。从这点来讲，ε 不敏感函数比传统的二次损失函数优越。

• 控制上界 C。影响支持向量数目和计算时间。C 值越小支持向量的数目和计算时间越少。C 值不能太小，否则由于不够支持向量使分类回归效果差，误差会增大。

对于本次预测，我们所选的几个参数如下：

表 2　参数的选取，其中 u,v 为核向量参数

$K(u,v)$	L	C	ε
$u*v$	ε 不敏感	300	0.23

上述经过调整选取的，可以使得 MAE、MAPE 、RMSE（意义见下述）几个值达到最小；对 SVM 的解来讲，由于讨论的是凸区域，局部最优解总是全局最优解。

3.2　数据预处理

数据是 2002 年 1 月至 2003 年 12 月（2004 年 1 月前统计）的“千车故障数”最新资料。

（1）*计算各批次相关系数*　用 Matlab 公式 corrcoef（$[M_1,M_2,\cdots,M_{10}]$）计算各批次相关数 r，其中 $M_1,M_2,\cdots,M_{12}$ 为各批次向量，经过计算得 $r=1$，说明各批次向量是显著相关的，预测是有规律可寻的。

（2）*数据的标准化处理*　对于个别缺损数据和异常数据（比如远远超出某范围的数倍，则一定是异常数据），在程序中我们用样条插值的办法来进行补偿。对于大小不一的数据，我们用以下计算公式将数据限定在 $-1.0\sim+1.0$ 范围内，以统一尺度，达到数据规范化处理：

$$x_{norm}=2*\left(\frac{x-x_{\min}}{x_{\max}-x_{\min}}\right)-1.0 \tag{3.1}$$

然后用以下公式将数据还原：

$$x=\frac{(x_{norm}+1.0)*(x_{\max}-x_{\min})}{2}+x_{\min} \tag{3.2}$$

经过标准化处理后的数据由于计算单位减少，从而节省了计算时间和内存消耗。

（3）*几个重要的预测评价指标*　为了评价预测效果，我们使用以下四个统计指标值：

- 相对百分误差的绝对值（absolute relative percentage error，ARPE）：

$$ARPE=\left|\frac{y_i^p-y_i^a}{y_i^a}\right| \tag{3.3}$$

- 平均绝对百分比误差（相对误差）（mean absolute percentage error，MAPE）：

$$MAPE=\left(\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^p-y_i^a}{y_i^a}\right|\right)\times 100\% \tag{3.4}$$

- 平均绝对误差（mean absolute error，MAE）

$$MAE=\frac{1}{N}\sum_{i=1}^{n}|y_i^p-y_i^a| \tag{3.5}$$

- 根方差（root mean square error，RMSE）：

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^p-y_i^a)^2} \tag{3.6}$$

通过这些值的比较可以观察出其预测的准确度，其值越小，准确度越高。

4　模拟预测

4.1　输入模式的确定

（1）*样本集的划分*　新算法第一次迭代前，我们以 2002 年 1 月～2002 年 10 月共 10 个月的数据为训练集，以 2002 年 11 月～12 月的样本为确定集，以 2003 年 1 月～2 月共二个月的样本为测试集，然后再迭代计算。

（2）*输入模式的决定*　千车故障数是一个时间序列 $x(t)$，$t=1,2,\cdots$，其后某一时刻的数据值可以看为由前面 n 个时刻 $x(t-1),x(t-2),\cdots,x(t-n)$ 的值共同作用的结果（n 称为时延或时滞），t 时刻的值可以表示为函数：

$$x(t)=F(x(t-1),x(t-2),\cdots,x(t-n)) \tag{4.1}$$

这里关键是要确定时滞参数 n 的值，以 10 个月的训练数据为基础，我们假设当前值和前面 3～7 个时延相关（虽然理论上时延还可以有更宽的选择），限定窗口范围在 3～7，于是得到平均绝对误差如下。

表 3　模型窗口的确定

参数 n	MAE（平均绝对误差）
3	7.1966
4	7.4161
5	7.3116
6	7.2964
7	7.2938

其中最小的平均绝对误差是 7.1966，因此我们选取时延窗口大小 $n=3$，即用过去的三个值预测下一个值，确定输入模型为：

$$x(t)=F(x(t-1),x(t-2),x(t-3))$$

4.2　两组重要的预测指数

我们以时延为 3 进行预测，得到 2003 年 3 月至 2003 年 9 月计算，其平均绝对误差（MAE）、平均绝对百分比误差（MAPE）、根方差（RMSE）及相关系数（r）如表 4。

表 4　两组重要的预测指数

损失函数	MAE	MAPE	RMSE	r
ε-不敏感	1.8163	3.61%	1.23	0.985
Huber	1.880	4.2%	2.6	0.963
二次损失	2.3	5.83%	3.68	0.97

（A）

核函数	MAE	MAPE	RMSE	r
linear	1.8163	3.61%	1.23	0.985
ploy	1.99	4.67%	2.45	0.98
rbf	5.89	7.83%	5.98	0.92
BSPLINE	4.53	5.69%	3.65	0.93

（B）

表 4（A）是在不同损失函数下的预测指数，表 4（B）是在损失函数为 ε-不敏感的条件下，不同的核参数的预测指数，可见，只有基于 ε-不敏感损失函数和线性核才具有最小的 MAE、MAPE、RMSE 值，其平均相对误差只有 3.61%，根方差也只有 1.23，其相关系数也达到了最大，因此这组参数值最好。

从表 4 可以看出，并不是核函数越复杂越好，按理说，径向基核函数具有更好的拟合性，可见事实并非如此。

4.3　预测结果分析

（1）*按月的预测曲线*　用 SVM 方法进行的按月的连续分布曲线（如图 3），曲线是每个月的连接图，并已经进行了原数据表的空白位置的预测；按月的趋势看，随着时间的推移，千车故障数据是越来越减少的，这和实际是相符的，因为生产质量的改进，意味着其千车故障数会逐渐减少。

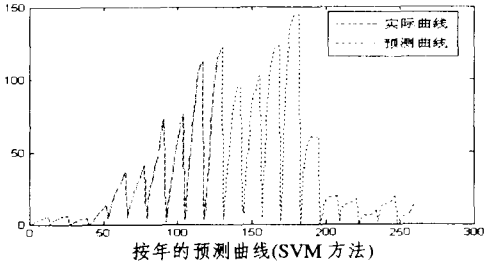


图 3　按批次的连续分布预测曲线

而图 4 是按批次的平行分布图，多数曲线分布在图的底部，最高的一条曲线是 2003 年 2 月的曲线图，是千车故障数最多的一年，可见这一年的生产质量出现了严重问题。

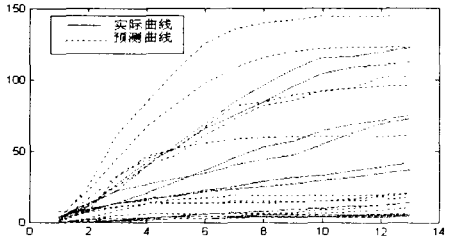


图 4　按批次的平行分布预测曲线

根据预测曲线，预测 0306 批次使用月数为 9 个月的千车故障数为：6.3757，预测 0310 批次使用月数为 12 个月的千车故障为：11.876。从已知数据中我们得到，预测误差小于 1.3%占总体的 96%，误差小于 2.3%占总体的 97.3%，误差小于 5.3%占总体的 98.1%，说明预测是相当准确的。

(3)新算法的 SVM 对小样本的进一步实验　为了检验诱导回归算法对更少数据量的样本的预测能力，我们单独取出该年该批次的只有 13 个数据的同一批次小小样本(千车故障数)，来预测当时间因子 $t\to+\infty$ 的千车故障数。因为这只有用这极少量的 13 个样本，完全可以测量 SVM 的对小样本的处理能力。通过预测，我们得到：

$$\lim_{t\to\infty}x(t)=74.0601 \tag{4.2}$$

其预测曲线如图 5，数据最终趋于稳定值 74.0601。实际上已经有：

$$x(t)=74.0601,t\geqslant 280 \tag{4.3}$$

这和实际是相当吻合的，因为前期数据点大约以步长为 2 的速度增长，例如 18 个月后理论上应为 46 左右，图中预测为 46.1932；另一方面，由实际情况知，数据不可能永远递增下去，随着企业生产质量的改进，其售后服务的“千车故障数”当 $t\to+\infty$ 应趋于固定值，因而 74.0601 是较合理的极限值。

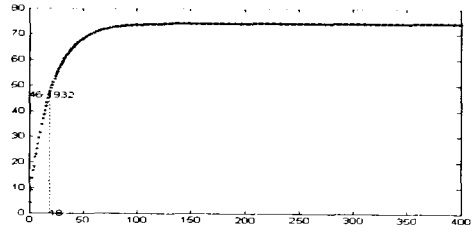


图 5　用 SVM 方法预测 0205 批次千车故障数

5　SVM 预测置信度估计

按文[5]提出的关于 SVM 预测的概率置信度估计公式进行置信度估计：

$$PB_\varepsilon(y_{M+1})=\frac{\sum_{i=1}^{l}I\{|y_i-f(x_i)|\leqslant\varepsilon\}}{l} \tag{5.1}$$

其中 $I(A)=\begin{cases}1,\text{若 } A \text{ 为真}\\0,\text{否则}\end{cases}$，$(x_{M+1},y_{M+1})$的数值对中，$x_{M+1}$为预测输入，而 y_{M+1}为预测输出，l 为邻近集的长度；我们进行随机抽样，任取预测序列中的 10 个值，得到迭代 80 次后的置信度如下表。

表 5　SVM 预测置信度估计

实际值	2.6	7.9	13.31	15.97	16.86
预测值	2.58	7.6	13.2	14.9	16.31
置信度	1	1	1	1	0.95

实际值	18.6	18.63	18.63	18.63	18.63
预测值	18.2	17.5	19.1	18.6	18.59
置信度	0.96	1	0.93	0.99	1

上述 10 个值其平均置信度为 0.983(98.3%)，因此，利用 ε-*insensitivity* 损失函数和线性核，新算法所建立的 SVM 预测模型是相当可信的。

6　SVM 和传统神经网络的比较

图 6 和图 7 是 SVM 和神经网络的比较图，选取线性神经网络，经计算机模拟，在时间参数 $t\to+\infty$ 时有：

$$\lim_{t\to\infty}x(t)=157.6325 \tag{6.1}$$

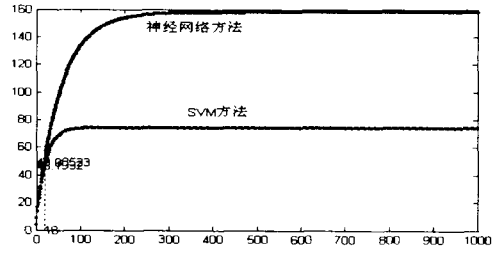


图 6　SVM 和神经网络预测比较图

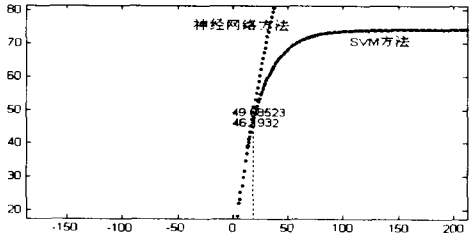
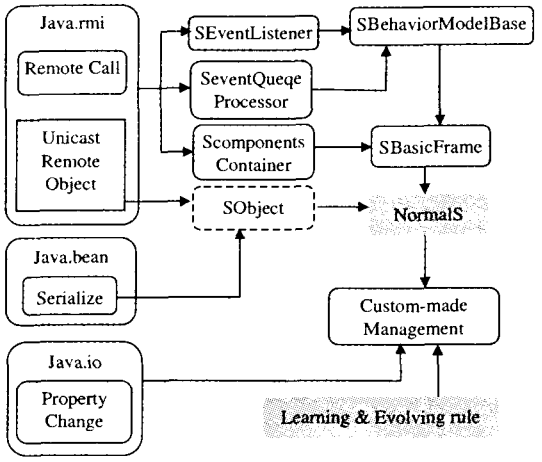


图 7　图 6 的局部放大

开始的时候，两种数据预测相差并不大(图 7，例如预测 18 个月时 SVM 的预测值是 46.4932，而神经网络的预测值是 49.0053)，但随着时间的增加，其差距越来越大，神经网络的预测值虽然最终也趋于固定值，但数据大大超过正常值，达

(下转第 154 页)



灰色圆角矩形部分是类的示意图，虚线圆角矩形部分为实现接口示意图，图中细实线是类继承关系或接口的实现

图 3 “软件人”系统平台基类和接口

5 未来的研究趋势

5.1 “软件人”将成为一种新兴共享资源

网络追求信息共享，网格追求的目标是计算资源化与协同工作，“软件人”技术的理想是“智能资源化”。在有效利用当地资源前提下，最大化地发挥“软件人”的潜力，借以达到满足个性化需求的目的，为“智能经济”[11]的到来打下坚实的基础。

5.2 网络环境下智能与移动的融合

Internet 的发展趋势要求智体同时具备移动能力和高智能。如在分布式信息查询中，被发送出去的智体不仅需要自主导航的移动能力，还需要信息的理解能力（如自然语言理解），这样才能找到用户真正需要的信息。“软件人”的思想对移动和智能的理解更上了一个新台阶，基于广义人工生命的思想，为网络环境中复杂问题的求解提供了更加可行的方案。

5.3 从平台无关到个性化平台资源充分利用

无法有效地利用当地资源来高效地工作成为目前所有“智体”系统共有的一个缺点：即无法解决“平台无关性和充分利用平台的个性”这一矛盾[12]。解决这一矛盾的方法之一是：结合广义人工生命[7]的思想，建立一种可以体现平台个性的机制，“软件人”的模型在充分利用本地环境和远程资源方面引入了新的思路，而这正是推动“软件人”技术进一步发展的动力之一。

参 考 文 献

1 Wooldridge M, Jennings N R. Formalizing the cooperation problem solving process. In: Readings in Agents. 430～440
2 杜军平，庄力可，涂序彦. 移动智能体的研究与应用[J]. 计算机应用研究，2000，12：37～39
3 涂序彦，尹怡欣. 人工生命及应用[C]. 见：中国人工智能学会第一届“人工生命及应用”专题学术会议论文集. 北京科技大学，2002
4 涂序彦. 人工智能及其应用[M]. 北京：电子工业出版社，1988. 58～59
5 涂序彦，等. 智能管理[M]. 北京：清华大学出版社，广西科学技术出版社，1995. 29～30
6 Tu Xuyan. AI, AL, and Robotics. In: proc. of FIRA World congress (plenary speech), Korea, 2002
7 Tu Xuyan. Generalized Artificial Life Race and Model. In: Proc. of the 8-th AROB, Japan, 2003
8 应力可，杜军平，涂序彦. 基于 Mobile Agent 技术的主动性网络管理策略的研究[J]. 计算机应用研究，2001，4：126～128
9 曾广平，涂序彦. 软件人[C]. 中国人工智能进展 2003 年. 广州：北京邮电大学出版社，2003，12：677～682
10 Asperti A, Busi N. Mobile Petri Nets. Technical Report UBLCS-96-10, Laboratory for Computer Science, Bologna, Italy, 1996. 71～85
11 涂序彦. 从知识经济到智能经济[C]. 见：中国人工智能学会第七届学术联合会议论文集. 西安，1999. 1～4
12 陆汝钤. 知识科学与计算科学[M]. 北京：清华大学出版社，2003. 126～155
13 王克宏. JXTA 技术与应用[M]. 北京：清华大学出版社，2003. 29～31
14 Bigus. A toolkit for building multi-agent autonomic systems. IBM system journal, 2002：41(3)：350～371

（上接第 141 页）

到 157.6325，因而不合理。而且收敛速度远不如 SVM，有：

$x(t)=157.6325, t\geqslant 825$；

SVM 只有 280 步就收敛（见式 4.3），预测的准确性方面远不如支持向量机（见表 6），说明 SVM 在处理小样本方面确实比传统神经网络更胜一筹。表 6 还显示了旧式的回归算法不如改进后的诱导回归算法，诱导回归算法要比原来的 SVM 回归算法提高了 6 个百分点，说明我们的算法是非常有效的。

表 6 SVM 和神经网络预测准确率比

预测误差	神经网络准确率	旧 SVM 准确率	新 SVM 准确率
1.3%	80.6%	90%	96%
2.3%	85.6%	91.3%	97.3%
5.3%	90.5%	93.8%	98.1%

结束语 本文建立了一整套基于 ε 不敏感损失函数的小样本支持向量诱导回归算法模型预测系统，并对实时售后服务数据进行了时间序列分析。由于通过置信度估计和反复迭代计算，去掉了大量含有噪声的样本，因而数据序列的规律性更强，预测可信度更高。通过筛选模型参数，显示出支持向量机诱导回归算法对小样本比神经网络具有更强的处理能力。关于支持向量机，关键在于参数（核函数，损失函数，控制上界，敏感度）的选取，参数的选取依赖于所分析的数据结构以及样本间的相关关系。窗口的确定是非常重要的，为了更好地控制预测误差，我们分析了平均绝对误差以确定输入模型窗口（时延）大小。另外，ε 不敏感函数虽然精度高，但也有一个不足之处就是处理回归方程较多，消耗的计算时间和内存较大，特别是数据间有非线性高阶相关性时十分明显。这样，在讨论回归模型时，必须在计算时间和精度方面作个折衷。因此，如何分析数据及优化算法、选择模型参数仍是将来进一步要做的工作。

参 考 文 献

1 Burgers C. A Tutorial on Support Vector Machines for Pattern Recognition. Data mining and Discovery, 1998, 2(2)
2 Vapnik V N. Statistical Learning Theory. John Wiley & Son, Inc. 1998
3 Platt J. Sequential Minimal Optimizat-ion: A Fast Algorithm for Training Support Vector Machines. In: Scholkopf B, Burges C. j. C, Smola A J. eds. Advances in KerneL Methods -SuppOrt Vector Learning, MIT Press, Cambridge, MA, 1999. 185～208
4 Joachims T. Making Large-Scale SVM Learning Practical. In: Schölkopf B, Burges C j C, Sm-ola A J, eds. Adva-nces in KernelL Methods -Support Vector Learning, MIT Press, Cambridge, MA, 1999. 169～184
5 孙德山，吴今培. 支持向量回归中的预测信任度. 计算机科学，2003，30(8)：126～127