小样本条件下航空装备费用预测

钟诗胜[1]，付旭云[1]，胡淑荣[2]

（哈尔滨工业大学 机电工程学院，150001 哈尔滨，fuxy_hit@163.com；2. 国家林业局管理干部学院，102600 北京）

摘 要： 为了提高小样本条件下的航空装备费用预测的精度，将信息扩散方法和支持向量机相结合，提出了信息扩散支持向量机预测模型，对模型的拓扑结构、建模步骤进行了描述. 为了采用粒子群优化算法为模型选择合适的参数，考虑到模型参数既有连续变量，又有离散变量，提出对粒子位置的各个分量采用不同的更新策略. 将信息扩散支持向量机应用于军用飞机机载电子设备的生产费用预测，预测结果的平均相对误差绝对值为3.3%，表明该方法可以满足工程的实际需要.

关键词： 小样本；费用预测；信息扩散；支持向量机；粒子群优化算法

中图分类号： V267 **文献标志码：** A **文章编号：** 0367－6234(2011)05－0052－04

Aviation equipment cost prediction under small sample size

ZHONG Shi-sheng[1], FU Xu-yun[1], HU Shu-rong[2]

(1. School of Mechatronics Engineering, Harbin Institute of Technology, 150001 Harbin, China, zhongss@hit.edu.cn;
2. State Academy of Forestry Administration, 102600 Beijing, China)

Abstract: To improve prediction accuracy of aviation equipment cost under small sample size, a prediction model named information diffusion support vector machine is proposed after combining the information diffusion method and support vector machine. The topology and modeling process of the model are also described. The model parameters include both continuous parameters and discrete parameters, so that a different update strategy is adopted to each component of the particle position when solving the problem of model parameter selection using the particle swarm optimization. Finally, the information diffusion support vector machine is applied to the production cost prediction of military aircraft avionics equipment. The average absolute relative error of the result is 3.3%, which can satisfy actual requirement.

Key words: small sample; cost prediction; information diffusion; support vector machine; particle swarm optimization

在航空领域，经常需要对航空装备的研制费用、单次送修费用或寿命周期费用等进行预测. 传统的费用预测方法主要有工程法、参数法、类比法和专家判断法，这些方法都存在着一些不足[1]. 一些比较新的理论包括偏最小二乘回归法、灰色理论、神经网络[2]、遗传算法[3]也已应用于费用预测，取得了不错的效果. 因为主观和客观的原因，很多费用预测问题的样本容量很小，经常不足30条，属于小样本问题[4]. 小样本问题的实质是信息不足，不能反映整个样本空间的分布，样本空间是不完备的[5]. 样本空间的非完备性带来了样本的模糊性. 支持向量机是植根于VC维理论的结构风险最小化原则基础上的机器学习模型，能够较好地处理小样本的分类和回归问题[6]，但传统的支持向量机不能处理样本的模糊性，不能学习样本包含的模糊信息，限制了支持向量机的性能.

信息扩散是一种对样本进行集值化的模糊数学处理方法，可以将单值样本变成集值样本. 本文将信息扩散引进到支持向量机中，建立了信息扩散支持向量机的预测模型. 针对模型建立过程中参数选择的难点，提出了一种能够适应于连续变量和离散变量并存的粒子群优化算法.

收稿日期： 2010－02－28.
基金项目： 国家自然科学基金重点资助项目(60939003).
作者简介： 钟诗胜(1964—)，男，教授，博士生导师.

1　信息扩散支持向量机预测模型

信息扩散原理[7]：当用一个不完备数据估计一个关系时，一定存在合理的扩散方式将观测值变为模糊集，以填充由不完备性造成的部分缺陷，从而改进非扩散估计.

对数据的扩散是通过扩散函数实现的. 最常用的扩散函数是正态扩散函数：

$$\mu(x,u) = \exp\left(-\frac{(x-u)^2}{2h^2}\right),\ x \in X,\ u \in U. \quad (1)$$

式中：$X = \{x_i \mid i = 1,2,\cdots,n\}$ 为样本集；$U = \{u_j \mid j = 1,2,\cdots,m\}$ 为等距的监测集；h 为扩散系数，可以按照式(2) 进行计算.

$$h = \begin{cases} 0.6841(b-1), & \text{当 } n = 5; \\ 0.5404(b-1), & \text{当 } n = 6; \\ 0.4482(b-a), & \text{当 } n = 7; \\ 0.3839(b-a), & \text{当 } n = 8; \\ 2.6851(b-a)/(n-1), & \text{当 } n \geqslant 9. \end{cases} \quad (2)$$

式中，$b = \max\limits_{1\leqslant i\leqslant n}\{x_i\}$，$a = \min\limits_{1\leqslant i\leqslant n}\{x_i\}$.

通过扩散函数，可以将 X 扩散为 U 上的模糊集.

小样本预测从某种意义上说是使用小样本数据估计总体模型. 因为样本容量小，可以看成是不完备数据. 根据信息扩散原理，一定存在某种扩散方式将小样本数据变为模糊集，并改进总体模型. 基于此，在传统支持向量机的输入层和隐藏层之间再增加一个隐藏层，建立信息扩散支持向量机. 信息扩散支持向量机由 4 层组成：输入层、第 1 隐层、第 2 隐层和输出层，其拓扑结构如图 1 所示. 其中，第 1 隐层实现输入数据向模糊集的转换，第 2 隐层实现模糊集从低维空间向高维空间的映射.

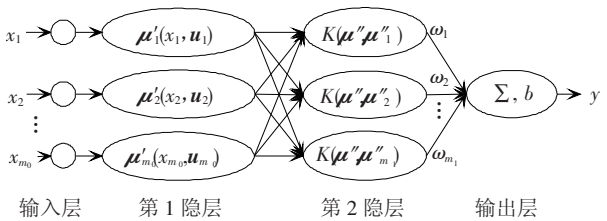

图 1　信息扩散支持向量机拓扑结构

图 1 中，$\boldsymbol{x} = (x_1,x_2,\cdots,x_{m_0})^{\mathrm{T}}$ 为输入向量，$\boldsymbol{\mu}'_i(x_i,\boldsymbol{u}_i) = (\mu'_{i,1},\mu'_{i,2},\cdots,\mu'_{i,n_i})^{\mathrm{T}}(i = 1,2,\cdots,m_0)$ 为 x_i 在 $\boldsymbol{u}_i = (u_{i,1},u_{i,2},\cdots,u_{i,n_i})^{\mathrm{T}}$ 上扩散的模糊向量，$\boldsymbol{\mu}'_{i,j} = \mu_{i,j}/\sum\limits_{j=1}^{n_i}\mu_{i,j}$，$(j = 1,2,\cdots,n_i)$，$\mu_{i,j} = \mu_i(x_i,u_{i,j})$，$\mu_i$ 为扩散函数，$\boldsymbol{\mu}'' = (\boldsymbol{\mu}'^{\mathrm{T}}_1,\boldsymbol{\mu}'^{\mathrm{T}}_2,\cdots,\boldsymbol{\mu}'^{\mathrm{T}}_{m_0})^{\mathrm{T}}$，$K$ 为核函数，$\boldsymbol{\omega} = (\omega_1,\omega_2,\cdots,\omega_{m_1})^{\mathrm{T}}$ 为权向量，m_1 为支持向量的数量，b 为偏差.

信息扩散支持向量机的输出可表示为

$$y = \sum_{k=1}^{m_1}\omega_k K(\boldsymbol{\mu}'',\boldsymbol{\mu}''_k) + b.$$

采用信息扩散支持向量机处理小样本预测问题的基本步骤如下：

1) 预处理样本数据.

包括对离散因素的数值化和数据的标准化 2 个方面.

离散因素的数值化处理按照是否引进虚变量可以分为直接法和间接法. 直接法是将离散因素的不同状态使用一个不同的数值表示. 间接法是通过引入虚变量，将离散因素的各个状态转换为布尔型，各个虚变量的取值为 0 或者 1. 因为引入虚变量会增加模型的输入数量，所以对于小样本问题，采用直接法进行离散因素的数值化处理更加合理.

数据标准化的目的是将数据的取值转换到某一指定的范围，一般为[-1,1] 或[0,1]，消除量纲影响. 本文采用 Z - Score 法. Z - Score 法不依赖于原始数据的最大值或者最小值，可以将原始数据标准化为均值为 0、方差为 1 的数据. Z - Score 法的标准化公式为

$$x' = (x - \bar{A})/\sigma_A.$$

式中：x' 为标准化后的值，x 为原始值，$\bar{A}$ 为均值，σ_A 为均方差.

2) 选择模型的输入输出.

模型的输入选择方法主要有相关系数分析法、逐步回归法、变量投影重要性分析法等. 变量投影法是基于偏最小二乘回归的变量筛选方法，通过对变量投影重要性指标(VIP) 的比较进行自变量的筛选，具有计算简捷、自变量可以多于样本数的优点，比较适合小样本问题. VIP 值很小的自变量可以直接删去，但在 VIP 值相差不大，相互间比较均衡时，变量投影重要性分析法需要结合自变量间的相关关系分析来进行. 如果自变量间具有强相关性，要根据实际情况去掉其中一个自变量.

3) 选择模型的参数.

对于各个输入，需要确定信息扩散所依赖的监测集和扩散函数. 为了描述方便，将监测集表达为向量的形式，称作监测向量. 监测向量的各元素之间是等距的，所以监测向量可以根据其元素的最小值、最大值、向量长度计算. 设输入 x_i 的监测点数量为 $n_i(n_i \geqslant 0)$，当 $n_i = 0$ 时，认为 x_i 不进行扩散，直接输入信息扩散支持向量机的第二隐层，即 $\boldsymbol{\mu}'_i(x_i,\boldsymbol{u}_i) = (x_i)$，监测点的最小值为 mgl_i，监测点的最大值为 mgu_i，则 x_i 的监测向量的第 j 个元素可以表示为式(3). 如果采用 Z-Score 法进行数据的

标准化,因为标准化后的输入的均值为 0,所以扩散范围可以认为是对称的,即 $mgu_i = -mgl_i$.

$$u_{i,j} = \begin{cases} mgl_i + (j-1)\cdot(mgu_i - mgl_i)/(n_i - 1), & n_i \geqslant 2; \\ (mgu_i + mgl_i)/2, & n_i = 1. \end{cases} \tag{3}$$

一般情况下,模型中所有输入的扩散函数可以选择如式(1)所示的正态扩散函数,核函数选择常用的高斯核函数

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{1}{\sigma^2}\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2\right).$$

综上,模型中需要确定的参数有:正则化参数 C、不敏感系数 ε、高斯核半径 σ、各个输入的 n_i、mgl_i、mgu_i,共($3m_1 + 3$)个参数,如果 $mgu_i = -mgl_i$,则为($2m_1 + 3$)个参数. 这些参数之间相互作用,共同影响着信息扩散支持向量机的泛化性能. 因为需要选择的参数数量比较多,不宜采用枚举方法,可以采用粒子群优化算法、遗传算法或者模拟退火等智能优化算法进行参数的选择.

4)训练预测.

将训练样本输入信息扩散支持向量机,训练完成后,将测试样本输入,获得预测结果.

2 基于粒子群优化算法的参数选择

粒子群优化算法是一种实现简单的群集智能优化算法,主要用于求解连续优化问题. 为了使该算法能够适应于离散优化问题,研究人员提出了二进制编码的粒子群优化算法[8]、顺序编码的粒子群优化算法、基于近似取整策略的粒子群优化算法[9]等. 因为信息扩散支持向量机需要优化的参数既有连续变量,又有离散变量,所以本文对标准粒子群优化算法的粒子位置更新方式进行修改,对粒子的各个分量采用不同的策略:连续变量的更新方式和标准粒子群优化算法相同,离散变量则采用近似取整策略.

粒子群优化算法的适应度函数应该反映信息扩散支持向量机的泛化能力. 估计其泛化能力的方法主要有留一法和 k - fold 交叉验证法. 对于小样本问题,常采用留一法,其基本思路是逐次从样本训练集中去掉一个样本,然后采用剩余的样本训练支持向量机,并使用训练好的支持向量机对去掉的样本进行预测,最后取所有训练结果的预测精度的平均值作为估计精度. 当预测精度采用预测误差的标准差表示时,适应度函数可以表示为

$$f_{\text{fitness}} = \left(\frac{1}{m_2}\sum_{i=1}^{m_2}(y_i - \hat{y}_i)\right)^{1/2}.$$

式中:m_2 为训练样本数量,y_i 为实际值,$\hat{y}_i$ 为预测值.

算法的流程如下:

Step1 在整个搜索空间内随机初始化粒子群的位置 $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,D})$ 和速度 $\boldsymbol{v}_i = (v_{i,1}, v_{i,2}, \cdots, v_{i,D})$,$1 \leqslant i \leqslant m_3$,$D$ 为需要进行优化的参数数量,m_3 为群体大小;

Step2 计算每个粒子的适应值 f_{fitness_i};

Step3 对于每个粒子,将其适应值 f_{fitness_i} 与所经历过的最好位置 $\boldsymbol{p}_i$ 的适应值 p_{best_i} 进行比较,如果 $f_{\text{fitness}_i} < p_{\text{best}_i}$,那么 $p_{\text{best}_i} = f_{\text{fitness}_i}$,$\boldsymbol{p}_i = \boldsymbol{x}_i$;

Step4 将每个粒子历史最优适应值 p_{best_i} 与群体内所经历的最好位置 $\boldsymbol{g}_i$ 的适应值 g_{best} 进行比较,如果 $p_{\text{best}_i} < g_{\text{best}}$,那么 $g_{\text{best}} = p_{\text{best}_i}$,$\boldsymbol{g}_i = \boldsymbol{p}_i$;

Step5 对粒子的速度和位置进行更新:

$$v_{i,d}^{k+1} = \omega v_{i,d}^k + c_1\xi(p_{i,d}^k - z_{i,d}^k) + c_2\eta(p_{g,d}^k - z_{i,d}^k);$$

$$z_{i,d}^{k+1} = \begin{cases} z_{i,d}^k + v_{i,d}^k, & \text{当第 } d \text{ 个参数是连续变量时;} \\ [z_{i,d}^k + v_{i,d}^k], & \text{当第 } d \text{ 个参数是离散变量时.} \end{cases}$$

式中,ω 是惯性权重,c_1 和 c_2 为学习因子,ξ 和 η 是在[0,1]区间内均匀分布的伪随机数,$[z_{i,d}^k + v_{i,d}^k]$ 表示对 $z_{i,d}^k + v_{i,d}^k$ 进行向下取整,即取不大于 $z_{i,d}^k + v_{i,d}^k$ 的最大整数;

Step6 若未达到终止条件,则转 Step2. 一般将终止条件设置为一个足够好的适应值或达到一个预设的最大迭代次数.

3 应用

现代作战飞机的机载电子设备生产费用增长很快,在机载电子设备研制的初期阶段对其费用进行预测,可以为国防预算提供一个可靠的依据. 美国的兰德公司主要采用回归分析方法,立足于首飞时间这样一个时间变量,结合机载电子设备的特征(质量、体积、功率),建立了机载电子设备生产费用的回归方程,但是精度不高. 为了提高预测精度,文献[10]建立了基于最小二乘支持向量机的费用预测模型,预测精度有了较大提高. 本文采用信息扩散支持向量机进行费用预测. 表 1 是原始样本数据.

首先采用 Z - Score 法对样本数据进行标准化处理. 考虑到用于建模的自变量只有 4 个,所以将其全部都作为模型的输入. 所有输入的扩散函数均选为正态扩散函数,核函数选为高斯核函数. 将表 1 中的前 11 条数据作为训练样本,后 3 条数据作为测试样本. 信息扩散支持向量机的初始参数选为:$C = 10^7$,$\varepsilon = 10^{-8}$,$\sigma = 400$,$n_1 = 3$,$n_2 = 3$,$n_3 = 3$,$mgu_1 = -mgl_1 = 2.5$,$mgu_2 = -mgl_2 =$

$2.5, mgu_3 = -mgl_3 = 2.5, mgu_4 = -mgl_4 = 2.5$. 然后采用粒子群优化算法对参数进行选择. 种群规模为 80，最大迭代次数为 1 000，学习因子均为 2，惯性权重为 0.6. 迭代完成后，选定的参数为：$C = 489\ 050\ 000, \varepsilon = 0.12\ 973, \sigma = 608.97, n_1 = 3, n_2 = 3, n_3 = 3, n_4 = 2, mgu_1 = -mgl_1 = 8.282\ 6, mgu_2 = -mgl_2 = 3.177\ 5, mgu_3 = -mgl_3 = 2.613\ 3, mgu_4 = -mgl_4 = 4.856\ 1$. 将训练样本输入信息扩散支持向量机，训练完成后，将测试样本输入，预测结果如表 2 所示. 为了进行对比，普通支持向量机的核函数也选为高斯核函数，同样采用粒子群优化算法对参数进行选择. 迭代完成后，选定的参数为：$C = 4\ 411\ 500\ 000, \varepsilon = 0.609\ 83, \sigma = 1\ 110.3$. 预测结果也列在表 2 中.

表 1　机载电子设备费用样本

机型	首飞时间/年	质量/kg	体积/m^3	功率/kW	实际平均费用/k $
A－6E	1970	624.25	0.569 1	6.4	1 069
A－7D	1968	508.93	0.710 1	10.5	669
F－4D	1965	790.41	0.843 0	8.2	582
A－10A	1972	265.14	0.239 4	3.1	315
E－4E	1967	566.14	0.677 3	5.3	662
F－4J	1966	1 021.05	0.982 4	19.4	1 329
F－15A	1972	717.32	0.833 1	22.5	2 488
F－111A	1964	805.40	0.877 4	5.6	1 267
F－111D	1968	1 068.72	0.910 2	12.5	2 392
F－111F	1971	933.88	1.061 1	8.9	1 577
FB－111A	1970	1 136.36	1.343 2	7.9	1 965
F－14A	1970	998.35	1.062 8	29.4	2 383
A－7E	1968	653.76	0.841 3	8.3	828
F－111E	1969	987.00	1.105 4	8.9	1 564

表 2　预测结果比较

机型	实际平均费用/k $	预测平均费用/k $					平均相对误差绝对值/%				
		回归[a]	支持向量机	最小二乘[a]	信息扩散		回归[a]	支持向量机	最小二乘[a]	信息扩散	
					优化前	优化后				优化前	优化后
F－14A	2 383	2 182	2 465	2 362	2 620	2 166					
A－7E	828	708	1 114	905	1 020	827	24.9	19.1	5.8	14.0	3.3
F－111E	1 564	753	1 867	1 678	1 424	1 551					

注：(a)数据来源于文献[10]，回归预测值为首飞时间与设备功率和实际平均费用的回归值.

从表 2 可看出，信息扩散支持向量机在参数优化前、后预测的平均相对误差绝对值分别为 14.0%、3.3%，远远低于采用回归方法和普通支持向量机的预测结果，也低于最小二乘支持向量机的预测结果，最大相对误差为 9.1%，能够满足实际工程的需要.

4　结　论

1)本文提出的信息扩散支持向量机将信息扩散技术和支持向量机结合起来，既保证了预测模型有好的泛化能力，又能够充分学习小样本的模糊信息.

2)采用粒子群优化算法进行模型的参数选择时，对粒子位置的各个分量采用不同的更新策略，能够适应既有连续变量又有离散变量的情况.

3)采用信息扩散支持向量机进行小样本条件下的航空装备费用的预测，可以获得比回归方法、普通支持向量机、最小二乘支持向量机更高的精度，可以为航空装备的改进设计、采购、维修厂家选择等提供决策支持.

参考文献：

[1] 白暴力，杨琳，陈云翔. 飞机维修费用估算的分析[J]. 空军工程大学学报：自然科学版，2005，6(5)：8－10.

[2] SEO K K，AHN B J. A learning algorithm based estimation method for maintenance cost of product concepts [J]. Computers & Industrial Engineering，2006，50(1)：66－75.

[3] FICKO M，DRSTVENSEK I，BREZOCNIK M，*et al*. Prediction of total manufacturing costs for stamping tool on the basis of CAD－model of finished product [J]. Journal of Materials Processing Technology，2005，164－165：1327－1335.

[4] 张恒喜，郭基联，朱家元，等. 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，2002：1－3.

[5] HUANG C F. Information diffusion techniques and small-sample problem [J]. International Journal of Information technology & Decision Making，2002，1(2)：229－249.

[6] 吴静敏. 民用飞机全寿命维修成本控制与分析关键问题研究[D]. 南京：南京航空航天大学民航学院，2006.

[7] HUANG C F. Principle of information diffusion [J]. Fuzzy Sets and Systems，1997，91(1)：69－90.

[8] KENNEDY J，EBERHART R C. A discrete binary version of the particle swarm algorithm[C]//Proceedings of the IEEE International Conference on Systems，Man and Cybernetics. Orlando：IEEE，1997：4104－4109.

[9] SALMAN A，AHMAD I，AI-MADANI S. Particle swarm optimization for task assignment problem [J]. Microprocessors and Microsystems，2002，26(8)：363－371.

[10] 张晓晖，朱家元，张恒喜. 基于 LS_SVM 的小样本费用智能预测[J]. 计算机工程与应用，2004，40(27)：203－204.

（编辑　杨　波）