

基于高斯间距核回归的产品设计时间预测

商志根

(盐城工学院 电气工程学院, 江苏 盐城 224051)

摘要:为克服产品设计时间预测中的小样本和异方差噪音问题,建立一种基于高斯间距核回归(Gaussian margin kernel regression, GMKR)预测模型。首先,假定核函数回归模型的权重向量服从高斯分布,利用相对熵与输出概率密度的自然对数和设计优化目标,构建GMKR模型;然后,假设高斯分布的协方差阵为对角矩阵以简化GMKR模型,并利用粒子群算法求解相应优化问题。最后,以注塑模具设计的实例进行分析,结果表明基于GMKR的时间预测模型可行有效。

关键词:设计时间;预测;核函数;相对熵;异方差

中图分类号:TP391 **文献标识码:**A **文章编号:**1671-5322(2013)02-0009-04

产品设计是复杂、动态的过程,受多种混合噪音的影响,其所需时间在设计任务完成之前是不可预知的。很多因素影响产品设计时间,如设计团队、设计环境、产品特征和设计过程等,且这些因素与设计时间之间的关系是高度非线性的。产品设计活动时间是产品开发过程的分析和控制的重要信息。

产品开发过程的分析和控制依赖于设计任务所需时间的预先设定。文献[1]假定产品设计活动时间服从三角分布,利用仿真的方法建立资源约束下的产品开发时间模型;文献[2]假定产品开发项目中各任务的时间分布,通过优化任务执行顺序估算产品开发时间。给定的产品设计时间分布是否合理是在决定上述方法的效果的重要因素。

近年来已有学者对产品设计时间的智能预测进行了研究。文献[3]提出一种模糊神经网络模型,但神经网络的训练需要较多样本,且泛化能力较差。基于核函数的模型^[4-5]将样本数据映射到某一高维空间里,将数据间的关系用一种确定的方式表达,并已在很多领域受到重视与应用^[6-9]。文献[7]在 ν -SVM方法中融入模糊回归理论,提出了 $F\nu$ -SVM。文献[8]将高斯过程元模型(Gaussian Process Meta-model, GPM)应用于产品设计时间预测。但 $F\nu$ -SVM和GPM没有考虑产

品设计时间的异方差性。虽然产品设计时间具有随机误差,但是在实际问题中,连续变量可用高斯分布来刻画^[10]。文献[11]将线性分类模型的系数设置为高斯分布的未知量,以样本获得较大正确分类概率为条件设置概率约束,以最小化相对熵为目标,提出高斯间距机,获得一种新的分类模型。

为使产品设计时间预测模型克服小样本、异方差噪音问题,以高斯间距机为基础,假定基于核函数的回归模型的权重向量服从高斯分布,以使高斯分布含有最小信息和输出样本概率密度最大为优化目标,将高斯分布的协方差矩阵设置为对角型以简化问题,提出高斯间距核回归模型,并利用粒子群算法求解相应的优化问题。与现有的设计时间预测模型相比,高斯间距核回归模型可给出较好的设计时间预测值。

1 高斯间距机

高斯间距机假定分类模型的权重向量服从多元高斯分布,设置约束条件以使训练样本获得较大概率的正确分类,根据最大熵原理,选择具有最小信息的分布作为向量的后验分布。

由 l 个样本构成训练集 $\{(x_i, y_i)\}_{i=1}^l$,其中 $x_i \in R^m$ 为列向量, $y_i \in \{-1, 1\}$,假定权重向量为列向量 $w, w \in R^m$ 且 $w \sim N_m(\mu_1, \Sigma_1)$,其中 $\mu_1 \in$

收稿日期:2013-05-10

作者简介:商志根(1979-),男,江苏盐城人,讲师,博士,主要研究方向为预测、进化算法。

R^m 为列向量, $\Sigma_1 \in R^{m \times m}$ 为正定阵。对于样本 x_i , 有

$$x_i^T w \sim N(x_i^T \mu_1, x_i^T \Sigma_1 x_i) \quad (1)$$

要求模型对训练样本有较大概率的正确分类, 得到

$$\Pr(y_i x_i^T w \geq 0) \geq \rho \quad (2)$$

其中 ρ 为置信参数, $\rho \in (0.5, 1]$ 。

由式(1)和式(2)得到

$$\Pr\left(\frac{y_i x_i^T w - y_i x_i^T \mu_1}{\sqrt{x_i^T \Sigma_1 x_i}} \leq \frac{-y_i x_i^T \mu_1}{\sqrt{x_i^T \Sigma_1 x_i}}\right) \leq 1 - \rho \quad (3)$$

在满足式(3)的条件下, 将含有最小信息的高斯分布作为后验分布, 也就是以各向均匀高斯分布 $N_m(0, aI_m)$ 为先验分布, 选择与其差异最小的高斯分布作为后验分布, 其中 $a \in R$ 且 $a > 0$ 。高斯间距机的优化目标为

$$\min_{\mu_1, \Sigma_1} D_{KL}(N_m(\mu_1, \Sigma_1) \| N_m(0, aI_m)) \quad (4)$$

其中 $D_{KL}(N_m(\mu_1, \Sigma_1) \| N_m(0, aI_m))$ 为相对熵^[10]。相对熵的计算表达式为

$$\begin{aligned} & \frac{1}{2} \ln \det(aI_m \Sigma_1^{-1}) + \\ & \frac{1}{2} \text{tr}((aI_m)^{-1}(\mu\mu_1^T + \Sigma_1 - aI_m)) \end{aligned} \quad (5)$$

$D_{KL}(N_m(\mu_1, \Sigma_1) \| N_m(0, aI_m))$ 可衡量 $N_m(\mu_1, \Sigma_1)$ 与 $N_m(0, aI_m)$ 之间的差异。当 $\mu = 0$ 、 $\Sigma_1 = aI_m$ 时, 相对熵为 0。

消除目标函数中的常数项, 并对式(3)进行适当变形, 得到高斯间距机的优化问题:

$$\begin{aligned} \min_{\mu_1, \Sigma_1} & \frac{1}{2} \left(-\ln \det \Sigma_1 + \frac{1}{a} \text{tr}(\Sigma_1) + \frac{1}{a} \|\mu_1\|^2 \right) \\ \text{s.t.} & y_i x_i^T \mu_1 \geq \phi \sqrt{x_i^T \Sigma_1 x_i} \\ & \Sigma_1 > 0, i = 1, 2, \dots, l \end{aligned} \quad (6)$$

其中 $\phi = \Phi^{-1}(\rho)$, Φ 是标准正态分布的分布函数。

2 高斯间距核回归

2.1 模型构造

给定 l 个独立的、无重复的样本集 $\{(x_i, t_i)\}_{i=1}^l$, 其中 $x_i \in R^m$, $t_i \in R$ 。基于核函数的回归模型表现形式如下:

$$f(x) = \sum_{j=1}^l w_j k(x, x_j) \quad (7)$$

其中 $k(x, x_j)$ 为核函数, 本文选用的核函数为高斯径向基函数。

假定权重向量 w 服从 $N_l(\mu, \Sigma)$ 分布, 其中 $w = (w_1, \dots, w_l)^T$, $\mu \in R^l$, $\Sigma \in R^{l \times l}$ 且为正定阵。对于样本 x_i 而言, 其预测值用 y_i 表示,

$$y_i \sim N(K_{i \cdot} \mu, K_{i \cdot} \Sigma K_{i \cdot}^T) \quad (8)$$

其中 $K_{i \cdot}$ 为对称核矩阵 K 的第 i 行, 且 $K_{ij} = k(x_i, x_j)$ 。

高斯间距核回归的目标是最小化高斯分布的相对熵, 同时最大化训练样本的输出概率密度的自然对数和, 其表现形式如下:

$$\begin{aligned} \min_{\mu, \Sigma, \sigma} & \frac{n}{\sigma} D_{KL}(N(\mu, \Sigma) \| N(0, aI_l)) - \\ & C \sum_{i=1}^l \ln p(t_i | x_i, w, b) \\ \text{s.t.} & \Sigma > 0, \sigma > 0 \end{aligned} \quad (9)$$

结合式(8), 去除目标函数中的常数项, 得到

$$\begin{aligned} \min_{\mu, \Sigma} & -\frac{1}{2} \ln \det \Sigma + \frac{1}{2a} \mu^T \mu + \frac{1}{2a} \text{tr}(\Sigma) + \\ & C \left(\frac{1}{2} \sum_{i=1}^l \ln K_{i \cdot} \Sigma K_{i \cdot}^T + \sum_{i=1}^l \frac{(t_i - K_{i \cdot} \mu)^2}{2 K_{i \cdot} \Sigma K_{i \cdot}^T} \right) \\ \text{s.t.} & \Sigma > 0 \end{aligned} \quad (10)$$

优化问题(10)的未知参数个数为 $l + l(l + 1)/2$, 且该问题复杂, 难以求解。为此, 本文对优化问题(10)进行了简化。

2.2 模型简化和求解

为简化问题(10), 将 Σ 假设为一个对角元素为正的对角阵, 即令 $\Sigma = \Lambda$, 其中 $\Lambda = \text{diag}(\lambda_1^2, \dots, \lambda_l^2)$, $\lambda_i > 0, i = 1, \dots, l$ 。于是优化问题(10)可简化为

$$\begin{aligned} \min_{\mu, \Lambda} & -\sum_{i=1}^l \ln \lambda_i + \frac{1}{2a} \mu^T \mu + \frac{1}{2a} \sum_{i=1}^l \lambda_i^2 + \\ & C \left(\frac{1}{2} \sum_{i=1}^l \ln K_{i \cdot} \Lambda K_{i \cdot}^T + \sum_{i=1}^l \frac{(t_i - K_{i \cdot} \mu)^2}{2 K_{i \cdot} \Lambda K_{i \cdot}^T} \right) \\ \text{s.t.} & \lambda_i > 0, i = 1, \dots, l \end{aligned} \quad (11)$$

假设问题(11)的解为 $(\hat{\mu}, \hat{\Lambda})$, 对于测试样本 (x^*, t^*) 而言, 其预测值为 $s^T \hat{\mu}$, 其中 s 为 $(k(x^*, x_1), \dots, k(x^*, x_l))^T$ 。本文使用粒子群算法 (particle swarm optimization, PSO) 求解优化问题(11)。粒子群算法因其实现容易、精度高、收敛速度快等优点引起了研究者的重视, 并在解决实际问题中表现出了优越性。问题(11)的未知参数个数为 $2l$, 粒子群算法中的每一个个体对应一个可能解。粒子群算法的目的是寻找使目标函数最小的 $(\hat{\mu}, \hat{\Lambda})$ 。

2.3 模型泛化能力

本文基于经验 Rademacher 复杂度证明模型具有较好的泛化能力。问题(11)的目标函数可改写为:

$$\min_{\mu, \Lambda} -a \sum_{i=1}^l \ln \lambda_i + \frac{1}{2a} \mu^T \mu + \frac{1}{2a} \sum_{i=1}^l \lambda_i^2 + aC \left(\frac{1}{2} \sum_{i=1}^l \ln K_{i \cdot} \Lambda K_{i \cdot}^T + \sum_{i=1}^l \frac{(t_i - K_{i \cdot} \mu)^2}{2 K_{i \cdot} \Lambda K_{i \cdot}^T} \right) \quad (12)$$

考虑函数集

$$F_c = \left\{ \sum_{j=1}^l \mu_j k(x, x_j) \mid x \in R^m, \mu \in R^l, \mu^T K \mu \leq c^2 \right\} \quad (13)$$

其中 c 为正实数。

$\hat{S}(F_c)$ 表示经验 Rademacher 复杂度。由文献[12]可知

$$\hat{S}(F_c) \leq 2c \sqrt{\text{tr}(K)} / l \quad (14)$$

由式(13)可知, c 的最小化可通过 $\mu^T K \mu$ 的最小化来实现。根据 Cauchy - Schwarz 不等式, 有

$$\mu^T K \mu = \langle \mu, K \mu \rangle \leq \| \mu \| \cdot \| K \mu \| \leq \| K \| \cdot \| \mu \|^2 \quad (15)$$

所以式(12)中的第2项可减小 $\hat{S}(F_c)$ 。

对于样本 x_i 而言, 其预测值的不确定性会随着 λ_i^2 的增大而增大, 反之, 则减小。若训练集的预测不确定性过大, 则模型会欠拟合, 反之, 则会过拟合。目标函数中的 $\frac{1}{2} \sum_{i=1}^l \lambda_i^2 + \frac{1}{2} aC \sum_{i=1}^l \ln K_{i \cdot}$ 可防止训练集的预测不确定性过大。 $-a \sum_{i=1}^l \ln \lambda_i + aC \sum_{i=1}^l \frac{(y_i - K_{i \cdot} \mu)^2}{2 K_{i \cdot} \Lambda K_{i \cdot}^T}$ 会可防止预测不确定性过小。所以可将

$$-a \sum_{i=1}^l \ln \lambda_i + \frac{1}{2} \sum_{i=1}^l \lambda_i^2 + aC \left(\frac{1}{2} \sum_{i=1}^l \ln K_{i \cdot} \Lambda K_{i \cdot}^T + \sum_{i=1}^l \frac{(t_i - K_{i \cdot} \mu)^2}{2 K_{i \cdot} \Lambda K_{i \cdot}^T} \right)$$

看成一个特殊的损失函数。通过选择合适的 a 来控制经验 Rademacher 复杂度和拟合程度之间的平衡, 所以 GMKR 可具有较好的泛化能力。

3 产品设计时间预测

本文以注塑模具设计活动时间的数据验证 GMKR 的有效性。家电、汽车等产品的开发项目通常涉及到注塑模具设计。模具的开发过程包括塑件设计、模具设计、工程分析、制造过程等环节, 其中最关键的是模具设计^[3]。

对若干相关企业的设计案例进行分析和研究, 考虑6个产品特征因素(造型难度、壁厚变化、结构复杂度、长径比尺寸、形状特征数和型腔数), 忽略设计环境、设计团队和设计过程的影响。收集到的数据有72个样本, 前3个产品特征因素为模糊语言数据, 后3个因素为数值数据^[13]。60个样本构成训练数据, 后12个构成测试数据。依据相关经验将模糊语言数据转换成数值数据, 将分别用0.1、0.25、0.5、0.75和0.95来代替语言变量中的 VL、L、M、H 和 VH。为克服数值数据的量纲的影响, 在模型训练之前将数值数据进行归一化处理。

对于 a 、 C 和核参数 σ 的选择, 可用基于网格搜索的交叉验证方法^[14-15]。本文选用智能优化算法中的粒子群算法以代替网格搜索, 算法以均方根误差最小化为目标, 确定目标最小值时对应 a 、 C 与 σ 的值。 a 、 C 和核参数 σ 的选取范围分别为 $[10^{-3}, 10^3]$ 、 $[10^{-3}, 10^3]$ 和 $[10^{-3}, 5]$ 。粒子群算法的相关参数设置为: $w = 0.6$, $c_1 = c_2 = 1.7$ 。参数辨识的结果为 $a = 76.358$ 、 $C = 244.201$ 和 $\sigma = 4.846$ 。

图1给出了 GMKR 模型的训练和测试效果, 由图1可知, GMKR 模型在测试集上具有较好的预测精度。表1给出了 Fv-SVM、GPM 和 GMKR 在测试集上还原后的预测值。表2给出了3种模型在测试数据集的3项误差指标: 均方误差(Root Mean Square Error, RMSE)、平均绝对误差率(Mean Absolute Percentage Error, MAPE)、平均绝对误差(Mean Absolute Error, MAE)。Fv-SVM 需利用相关经验进行确定三角模糊数的繁琐过程, 而 GMKR 可减少对使用者的要求。由表2的3项指标可看出, GMKR 的性能略优于 Fv-SVM 和 GPM。

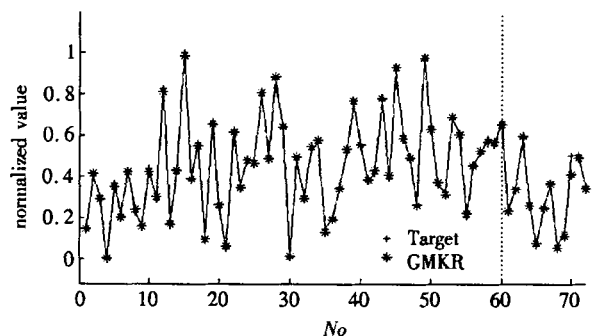


图1 GMKR 的训练和测试结果

Fig.1 Training and testing results of GMKR

表 1 3 种不同预测模型的测试结果比较
Table 1 Comparison of forecast results from three different models

序号	实际值/h	预测值/h		
		Fv-SVM	GPM	GMKR
T1	31	35.315	31.322	30.614
T2	41	40.672	40.537	40.161
T3	62	62.029	63.905	62.619
T4	34.5	33.313	31.202	33.024
T5	16	16.424	15.469	16.967
T6	32.5	32.965	31.813	32.223
T7	42.5	40.243	38.390	42.455
T8	16.5	15.394	14.044	14.895
T9	22	21.521	20.761	20.158
T10	54.5	46.391	44.446	46.569
T11	55	54.149	52.917	53.504
T12	41.5	41.752	38.775	40.381

表 2 3 种预测模型的误差统计
Table 2 Error statistics of three forecast models

模型	测试结果		
	RMSE	MAPE	MAE
Fv-SVM	2.793	0.045	1.650
GPM	3.569	0.066	2.489
GMKR	2.530	0.045	1.550

4 结论

为了使产品设计时间预测模型克服小样本和异方差噪音问题,在高斯间距机的基础上,假定核函数回归模型的系数服从高斯分布,最小化高斯分布的相对熵,同时最大化训练样本的输出概率密度的自然对数和,对所得的优化问题进行简化,并利用粒子群算法进行求解。通过对注塑模具设计进行实例分析,证明基于高斯间距核回归的设计时间预测模型可获得较好的精度。

参考文献:

- [1] Cho S H, Eppinger S D. A simulation - based process model for managing complex design projects[J]. IEEE Transactions on Engineering Management, 2005, 52(3): 316 - 328.
- [2] 陈冬宇,邱苑华,杨青,等. 基于 DSM 的复杂产品开发流程优化遗传算法[J]. 控制与决策, 2008, 23(8): 910 - 914.
- [3] 许多,严洪森. 产品设计时间智能预测方法的研究[J]. 控制与决策, 2004, 19(2): 143 - 146.
- [4] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer - Verlag, 1995.
- [5] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning[M]. Cambridge: MIT Press, 2006.
- [6] Hong W C. Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model[J]. Energy Conversion and Management, 2009, 50(1): 105 - 117.
- [7] 许多,严洪森. 基于模糊支持向量机的产品设计时间估计方法[J]. 中国机械工程, 2004, 16(3): 533 - 537.
- [8] 张昆仑,刘新亮,郭波. 基于高斯过程元模型的产品设计时间估计方法[J]. 计算机集成制造系统, 2011, 17(1): 18 - 22.
- [9] 周金柱,黄进. 集成先验知识的多核线性规划支持向量回归[J]. 自动化学报, 2011, 37(3): 360 - 370.
- [10] Jøbsang A. A logic for uncertain probabilities[J]. International Journal of Uncertainty, Fuzziness and Knowledge - Based Systems, 2001, 9(3): 279 - 311.
- [11] Crammer K, Mohri M, Pereira F. Gaussian margin machines[C]//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Florida: Journal of Machine Learning Research, 2009, 5: 105 - 112.
- [12] Robin C G, Theodore B T. Quadratic programming formulations for classification and regression[J]. Optimization Methods and Software, 2009, 24(2): 175 - 185.
- [13] 许多. 产品开发时间分析与估计方法的研究[D]. 南京: 东南大学, 2004.
- [14] Joachims T. Estimating the generalization performance of a SVM efficiently[C]. Proceedings of 17th International Conference on Machine Learning, 2000, 431 - 438.
- [15] 车红昆,吕福在,项占琴. 多特征 SVM - DS 融合决策的缺陷识别[J]. 机械工程学报, 2010, 46(16): 101 - 105.

(下转第 34 页)

Fingdings of Mixed Exponential Distribution

WEI Biao

(Department of Basic Science, Nanjing Political College, Nanjing Jiangsu 210003, China)

Abstract: In the statistical analysis, multivariate t distribution is an important distribution and exponential distribution family is a family of distributions is important. This paper imitated the multivariate normal distribution to obtain a multivariate distribution and mixed exponential distribution and distribution, a new family of distributions obtained, known as the mixed exponential distribution family. The function of parameters in the mixed exponential distribution family under the unbiased was discussed. This paper pointed out that the exponential distribution family and exponential mixture distribution family had the same complete statistic and the sufficient statistics.

Keywords: multivariate t - distribution; mixed exponential distribution; sufficient statistics

(责任编辑:张英健)

(上接第 12 页)

Product Design Time Forecast by Using Gaussian Margin Kernel Regression

SHANG Zhi-gen

(School of Electrical Engineering, Yancheng Institute of Technology, Yancheng Jiangsu 224051, China)

Abstract: There exist problems of small samples and heteroscedastic noise in design time forecast. To solve them, Gaussian margin kernel regression (GMKR) is proposed. First, the Gaussian distribution over weight vectors for the kernel - based regression is assumed for GMKR, and the optimization objective function of GMKR is designed by considering both the relative entropy and the sum of the natural log of the output probability densities. Then, the optimization problem of GMKR is simplified by assuming the covariance matrix of the Gaussian distribution to be a diagonal matrix, and its relevant optimization problem is solved based on particle swarm optimization algorithm. Finally, the effectiveness of GMKR is verified by our experiment results on the time forecast of plastic injection mold design.

Keywords: Design time; Forecast; Kernel function; Relative entropy; Heteroscedasticity

(责任编辑:张振华)