doi:10.11887/j.cn.201903024 http://journal.nudt.edu.cn

基于 Kullback-Leibler 距离离散度的加权代理模型*

晏 良,段晓君,刘博文,徐 琎
(国防科技大学 文理学院,湖南 长沙 410073)

摘 要:复杂系统的仿真通常具有高维度、高计算量等特点,代理模型因其明晰的数学表达和良好的计算特性可用于逼近真实系统。加权模型对比单个代理模型来说,其稳定性和适应性更广。不同的代理模型其性能不一,根据特定指标,可以构造最优加权代理模型。基于代理模型预测分布以及 Kullback-Leibler 距离构造各子代理模型之间的离散度,并提出一种新的权函数构造方法。算例表明,该方法与最优子模型的精度相当,同时能提高对真实响应分布的逼近。

关键词:复杂系统;代理模型;Kullback-Leibler 距离
中图分类号:N945.12 **文献标志码**:A **文章编号**:1001-2486(2019)03-159-07

Weighted surrogate models based on Kullback-Leibler divergence

YAN Liang, *DUAN Xiaojun*, *LIU Bowen*, *XU Jin*
(College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410073, China)

Abstract: Surrogate methods (metamodels) are convenient to determine the mathematical relationship underlying the high dimensional complex systems, which are usually computationally expensive. Various stand-alone metamodels have been proposed in literature, and the ensemble of metamodels was being intensively studied recently to utilize the information reveals in construction of different metamodels. Compared with the stand-alone metamodels, the ensembled models were more robust and adaptable. The strategy of the ensemble by comparing the difference of the probability distribution of predictions was considered, where the Kullback-Leibler divergence was introduced to calculate the differences. Experiments show that the strategy has comparable accuracy in predictions with the most accurate stand-alone metamodel, and it can also perform better in recovering the distribution of the true response.

Keywords: complex systems; surrogate models; Kullback-Leibler divergence

在工程研究中,利用计算机通过仿真试验可对真实复杂系统进行高精度的模拟,从而分析该系统的相关特性[1]。然而,单次高精度仿真试验通常面临计算效率低、时间成本大等问题,因此诸如可靠性评估、失效概率验证等需要大规模仿真的任务在实际操作中很难完成。代理模型由于其计算复杂度低、数学表达明晰等特点,在计算机仿真试验中得到了广泛应用[2-3]。实际上,代理模型重点关注输入和输出之间的数学关系。给定试验设计及观测数据,代理模型通过回归或插值等方法拟合观测数据,可给出相应的预测均值、方差等统计量。

常用的代理模型构造方法有很多,例如:多项式响应曲面(Polynomial Response Surface, PRS)[4]、多元自适应回归样条(Multivariate Adaptive Regression Splines, MARS)[5]、高斯过程(Gaussian Process, GP)[6],等等。针对不同的模型特点(如光滑、线性、非线性)和数据特征(噪声、异常值),通常会选用不同的代理模型。

事实上,除了以上文献所提供的经验,基于一些准则可以对代理模型进行筛选。经典的模型选择准则[7]包括:基于拟合优度(Goodness of Fit, GoF)、基于预测误差(Prediction Error, PE)或者模型误差(Model Error, ME)、基于概率分布距离(Distributional Discrepancy, DD)、基于后验概率(Posterior Probability, PP)等。然而,基于上述准则需要对每类候选模型进行计算,会增加较多的计算量;另外,从所有候选模型中仅挑选单个最优模型,不考虑其他模型也在一定程度上造成了信息的浪费。出于上述考虑,一些研究[8-10]也侧重于利用所有的候选模型,即构造加权代理模型。

此外,构造加权代理模型有利于减小不确定

* **收稿日期**:2018-03-25
基金项目:国家自然科学基金资助项目(11771450,61573367)
作者简介:晏良(1990—),男,江西宜春人,博士研究生,E-mail:yanliang@nudt.edu.cn;
段晓君(通信作者),女,教授,博士,博士生导师,E-mail:xjduan@nudt.edu.cn

性。加权模型由于考虑了所有不同类型的代理模型,因此在一定程度上增加了该模型的稳健性。同时,加权模型在构造的同时充分考虑到各模型的优势,因此其精度也能得到保证。

经典的加权模型方法往往从两个指标考虑:一是考虑模型预测的方差大小,通常情况下,预测方差越小,则认为模型越准确;二是考虑模型的预测精度,即预测均值与真实值之间的偏差,加权的主要目的在于降低加权模型与真实模型之间的偏差。然而在实际应用中,预测精度高(即偏差小)的代理模型,其预测方差可能较大,或者预测方差小的代理模型,其与真实值之间的偏差比较大。本文基于以上问题,提出了基于 KL(Kullback-Leibler)距离离散度的加权方法,直接对预测的分布入手,同时考虑预测的均值和方差,在兼顾预测精度的同时,降低预测模型与真实模型之间的分布差异。本文考虑两类子代理模型,一类为线性回归模型,另一类为高斯回归模型。

1 代理模型简介

1.1 线性回归模型

线性回归模型假设响应 y 是一类基函数(basis)的线性组合:

$$\boldsymbol{y}=[\boldsymbol{f}(\boldsymbol{x})]^{\mathrm{T}}\beta+\boldsymbol{\epsilon} \tag{1}$$

其中,$\boldsymbol{x}=(x_1,\cdots,x_d)^{\mathrm{T}}$ 为 d 维变量,$\boldsymbol{f}(\boldsymbol{x})$ 为 p 维向量函数,其中每一项是关于 $\boldsymbol{x}$ 的基函数。通常在一类模型中基函数是给定的,因此构造线性回归模型重点在于求解未知参数 β。一般假设 $\boldsymbol{\epsilon}$ 为零均值的随机误差,因此 $[\boldsymbol{f}(\boldsymbol{x})]^{\mathrm{T}}\beta$ 实际上表示响应 $\boldsymbol{y}$ 的期望值。

给定 N 个观测 $(\boldsymbol{X}_i,\ \boldsymbol{Y}_i)$,$i=1,\cdots,N$,即试验设计 $(\boldsymbol{X},\ \boldsymbol{Y})$,根据式(1)可以得到:

$$\boldsymbol{Y}=\boldsymbol{F}\beta+\ \boldsymbol{\epsilon} \tag{2}$$

其中,$\boldsymbol{F}\in\mathbb{R}^{N\times p}$ 为设计矩阵,$\boldsymbol{\epsilon}=(\epsilon_1,\cdots,\epsilon_N)^{\mathrm{T}}$ 为观测误差。由于 $\boldsymbol{\epsilon}\sim\mathrm{N}(0,\sigma^2\boldsymbol{I})$,可得参数 β 的最优线性无偏估计(Best Linear Unbiased Estimate, BLUE)$\hat{\beta}$ 为:

$$\hat{\beta}=(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

由式(3)可得关于响应估计 $\hat{y}(\boldsymbol{x})$ 的均值和方差分别为:

$$\hat{\boldsymbol{y}}(\boldsymbol{x})=[\boldsymbol{f}(\boldsymbol{x})]^{\mathrm{T}}\hat{\beta} \tag{4}$$

$$VAR[\hat{\boldsymbol{y}}(\boldsymbol{x})]=\sigma^2[\boldsymbol{f}(\boldsymbol{x})]^{\mathrm{T}}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}\boldsymbol{f}(\boldsymbol{x}) \tag{5}$$

1.2 高斯回归模型

高斯回归模型是一类典型的非线性模型。与线性回归模型指定基函数不同,高斯回归模型建立在响应 $\boldsymbol{y}$ 服从某一个确定的高斯过程基础之上,即:$\boldsymbol{y}\sim GP(0,k(\boldsymbol{x},\boldsymbol{x})+\sigma^2)$,其中 $k(\boldsymbol{x},\boldsymbol{x})$ 为核函数。具体来说,给定试验设计 $(\boldsymbol{X},\ \boldsymbol{Y})$,则有如下似然函数:

$$\boldsymbol{Y}|\boldsymbol{X}\sim\mathrm{N}(0,k(\boldsymbol{X},\boldsymbol{X})+\sigma^2\boldsymbol{I}) \tag{6}$$

记 $\boldsymbol{K}=k(\boldsymbol{X},\ \boldsymbol{X})\in\mathbb{R}^{N\times N}$,由高斯过程假设以及式(6),可得预测均值和方差分别为:

$$\hat{\boldsymbol{y}}(\boldsymbol{x})=\boldsymbol{K}_*^{\mathrm{T}}[\boldsymbol{K}+\sigma^2\boldsymbol{I}]^{-1}\boldsymbol{Y} \tag{7}$$

$$VAR[\hat{\boldsymbol{y}}(\boldsymbol{x})]=\boldsymbol{K}_{**}-\boldsymbol{K}_*^{\mathrm{T}}[\boldsymbol{K}+\sigma^2\boldsymbol{I}]\boldsymbol{K}_* \tag{8}$$

其中,$\boldsymbol{K}_*=k(\boldsymbol{X},\boldsymbol{x})\in\mathbb{R}^{N\times 1}$,$\boldsymbol{K}_{**}=k(\boldsymbol{x},\boldsymbol{x})\in\mathbb{R}$。实际上,高斯过程的预测能力由核函数所决定。一般来说,针对似然函数式(6)进行优化可以得到关于核函数参数的相关估计。

2 加权代理模型

不同的代理模型其性能也有所不同,因此在加权代理模型中的权重也会随之改变。加权代理模型的一般表达式为:

$$\begin{cases}\bar{y}(\boldsymbol{x})\ =\ \sum\limits_{i=1}^{M}\omega_i(\boldsymbol{x})\hat{y}_i(\boldsymbol{x})\\ \text{s.t.}\ \sum\limits_{i=1}^{M}\omega_i(\boldsymbol{x})\ =\ 1\end{cases} \tag{9}$$

其中,M 表示子代理模型的个数,$\bar{y}(\boldsymbol{x})$ 表示加权代理模型,$\hat{y}_i(\boldsymbol{x})$ 代表第 i 个子代理模型,$\omega_i(\boldsymbol{x})$ 表示该代理模型的权函数。注意到式(9)中权函数应当同时满足非负以及和为 1 的条件。对于不同的评价指标,指标精度越高的代理模型,其权函数必然要更大一些。因此,权函数与模型指标精度息息相关。经典加权模型通常考虑不同的误差函数。

2.1 基于预测误差的权函数构造

Zerpa[8] 等选择 PE 作为对精度的衡量准则,并根据 PE 逐点构造权函数。实际上,假设对每点的预测值是对真实响应的无偏估计,同时假设各点之间的预测是不相关的,根据式(5)及式(8)可得在每点预测值的方差。为使得加权模型 $\bar{y}(\boldsymbol{x})$ 的预测方差最小,即:

$$\begin{aligned}\bar{\omega}_i(\boldsymbol{x})\ &=\ \arg\min VAR[\bar{y}(\boldsymbol{x})]\\ &=\ \arg\min\sum\omega_i^2(\boldsymbol{x})VAR[\hat{y}_i(\boldsymbol{x})]\end{aligned} \tag{10}$$

由此可求解权函数如下:

$$\bar{\omega}_i(\boldsymbol{x})\ =\ \frac{1}{VAR[\hat{y}_i(\boldsymbol{x})]}\Big/\sum_{i=1}^{M}\frac{1}{VAR[\hat{y}_i(\boldsymbol{x})]} \tag{11}$$

2.2 基于交叉验证均方误差和均方根误差的权函数构造

Goel[9] 等选择泛化交叉验证均方误差

(Generalized Mean Square cross validation Error, GMSE) 作为衡量各子代理模型精度的重要参数,构造权函数如下:

$$\begin{cases}\bar{\omega}_i = (G_i + \alpha\bar{G})^{\beta} \Big/ \sum_{j=1}^{M} (G_j + \alpha\bar{G})^{\beta} \\ \bar{G} = \sum_{j=1}^{M} G_j / M\end{cases} \tag{12}$$

其中,$\alpha < 1$、$\beta < 0$ 为引入的尺度变换参数,G_i 为第 i 个子代理模型的 GMSE,即:

$$G_i = \frac{1}{N}\sum_{n=1}^{N} [\boldsymbol{Y}_n - \hat{y}_i^{(n)}(\boldsymbol{X}_n)]^2 \tag{13}$$

式中,$\hat{y}_i^{(n)}$ 表示基于试验设计$(\boldsymbol{X}, \boldsymbol{Y}) - (\boldsymbol{X}_n, \boldsymbol{Y}_n)$所构造的第 i 个子代理模型。同理,均方根误差(Root Mean Square Error, RMSE) 也可以用于式(12) 计算各子代理模型的权重。此时,只需将 G_i,$\bar{G}$ 替换为 R_i,$\bar{R}$ 即可。给定测试集$(\boldsymbol{X}^v, \boldsymbol{Y}^v)$,则 R_i 计算如下:

$$R_i = \sqrt{\sum_{n=1}^{N_v} [\boldsymbol{Y}_n^v - \hat{y}_i(\boldsymbol{X}_n^v)]^2 / N_v} \tag{14}$$

其中,N_v 为测试集中的观测数据个数,$\hat{y}_i$ 为基于试验设计$(\boldsymbol{X}, \boldsymbol{Y})$所构造的第 i 个代理模型。根据式(12)可以看出,子模型的 GMSE(或 RMSE)与其在加权模型中所占权重成反比。

GMSE 反映了代理模型在所对应试验设计点上的平均误差,GMSE 越小,说明模型在该试验设计上越稳定,即模型与试验设计相适应;而 RMSE 反映了代理模型在测试集上的预测与真实值的误差,与 GMSE 相比,其在一定程度上更关注模型在全局上的预测性能,RMSE 越小,其预测精度越高。同时注意到,与式(11)不同,基于 GMSE(RMSE)的权函数在模型定义域上是常值,因此权函数代表各子模型整体预测性能的比重。

此外,由于 GMSE 对每个子模型都需要计算 N 次,因此会增加较大的计算量。一般情况下,可以进行并行处理或者进行 k ($\leqslant$10)折交叉验证。然而,针对线性回归模型,GMSE 的计算可以直接进行简化,参见文献[7]第 12 章。对于 RMSE 来说,则需要增加 N_v 个观测,对于复杂系统来说,这会成为主要制约因素。

2.3　基于优化加权模型误差的权函数构造

Acar[10] 等在前述误差的基础上,提出了基于优化误差的权函数构造方法,即求解如下最优化问题:

$$\begin{cases}\min\ Err[\bar{y}(\boldsymbol{X}), \boldsymbol{Y}] \\ \text{s.t.}\ \sum_{i=1}^{M} \boldsymbol{\omega}_i = 1,\ \boldsymbol{\omega}_i \in [0,1]\end{cases} \tag{15}$$

其中,$Err[\cdot]$表示误差函数,如 RMSE 或 GMSE。对于式(15)的求解分为两步,首先针对试验设计构造各子代理模型,其次将各子代理模型代入式(15)求解该约束优化问题。与基于 PE(RMSE, GMSE) 的方法相比,其计算量主要增加在求解最优化问题式(15),当子模型数量 M 较大时,很难求得全局最优解。

3　Kullback-Leibler 距离离散度加权方法

以上三类经典加权模型方法基于误差函数,往往只考虑代理模型预测的某一方面,如预测方差或偏差。本文从代理模型的预测分布入手,同时考虑这两类因素,利用 Kullback-Leibler 距离[11] 构造权函数,从而降低预测总体分布的不确定性。

与式(9)直接假设加权模型为各模型预测值的加权不同,本文假设加权模型在各点的预测概率密度函数为各子代理模型预测概率密度函数的加权,即:

$$\begin{cases}\bar{p}(\boldsymbol{x}) = \sum_{i=1}^{M} \boldsymbol{\omega}_i(\boldsymbol{x})\hat{p}_i(\boldsymbol{x}) \\ \text{s.t.}\ \sum_{i=1}^{M} \boldsymbol{\omega}_i(\boldsymbol{x}) = 1\end{cases} \tag{16}$$

其中,$\bar{p}(\boldsymbol{x})$ 和 $\hat{p}_i(\boldsymbol{x})$ 分别为 $\bar{p}(\boldsymbol{y}(\boldsymbol{x}))$ 和 $\hat{p}_i(\boldsymbol{y}(\boldsymbol{x}))$ 的简化形式。此时,加权模型在各点的预测均值和方差为:

$$\bar{y}(\boldsymbol{x}) = \sum_{i=1}^{M} \boldsymbol{\omega}_i(\boldsymbol{x})\hat{y}_i(\boldsymbol{x}) \tag{17}$$

$$VAR[\boldsymbol{y}(\boldsymbol{x})] = \sum_{i=1}^{M} \boldsymbol{\omega}_i(\boldsymbol{x}) VAR[\hat{y}_i] \tag{18}$$

一般地,若 $\hat{p}_i$ 与真实概率 p 越接近,则其权重函数越大。因此可以利用距离来表征两个概率密度函数之间的相似度。由于 Kullback-Leibler 距离是不对称的,因此首先需要将其对称化。给定概率密度函数 p 和 q,则 SKL (symmetrised Kullback-Leibler)距离定义如下:

$$D_{\mathrm{SKL}}(p,q) = \frac{1}{2}D_{\mathrm{KL}}(p|q) + \frac{1}{2}D_{\mathrm{KL}}(q|p) \tag{19}$$

其中,D_{KL} 为通常意义下的 Kullback-Leibler 距离。对于本文中的两类模型,其在各点的分布服从高斯分布,同时各模型之间相互独立,因此概率密度函数 $p = N(\mu_p, \sigma_p)$ 和 $q = N(\mu_q, \sigma_q)$ 的 Kullback-Leibler 距离可以计算如下:

$$D_{\mathrm{KL}}(p,q) = \ln\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \tag{20}$$

从式(20)可以看出,该距离包含预测均值和方差两种参数。基于距离的权重函数构造方法有多种,如基于反函数和减函数等特殊函数。本文拟采用高斯核函数,基于 D_{SKL} 可定义权函数如下:

$$\omega_i(\boldsymbol{x}) = \exp\left[-\frac{D^2_{\mathrm{SKL}}(\hat{p}_i, \bar{p})}{\rho^2}\right] \tag{21}$$

其中,ρ 为尺度参数。注意到在实际应用中,真实的分布 $\bar{p}$ 是无法预先得到的,即式(21)是不可计算的。实际上,由式(16)可以得到 $\bar{p}$ 为各子代理模型和 $\hat{p}_i$ 的凸组合,即可以假设 $\bar{p}$ 落在 $\hat{p}_i$ 所构成的凸集当中。不失一般性,各 $\hat{p}_i$ 之间的距离越小,则其构成的凸集也越小,则得到的 $\bar{p}$ 可靠性越高。因此,各子代理模型的分布在全体分布中的离散度 d_i 可以代替式(21)中的 SKL 距离 $D_{\mathrm{SKL}}(\hat{p}_i, \bar{p})$。与距离类似,离散程度越高,则其权重越小。离散度 d_i 定义如下:

$$d_i = \frac{1}{M-1}\sum_{j\neq i} D_{\mathrm{SKL}}(\hat{p}_j, \hat{p}_i) \tag{22}$$

此外,尺度参数 ρ 可以选取为全体分布的平均离散度,即各分布之间的平均距离为:

$$\rho = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j\neq i} D_{\mathrm{SKL}}(\hat{p}_j, \hat{p}_i) \tag{23}$$

将式(22)和式(23)代入式(21)中,并对其进行归一化可以得到最终的权函数如下:

$$\bar{\omega}_i(\boldsymbol{x}) = \exp(-d_i/\rho^2) / \sum_{j=1}^{M}\exp(-d_j/\rho^2) \tag{24}$$

4 算例分析

本节首先针对两个测试函数对基于 Kullback-Leibler 距离离散度的加权代理模型进行分析,其函数表达式如下:

$$f_1(\boldsymbol{x}) = (10\cos 2\boldsymbol{x} + 15 - 5\boldsymbol{x} + \boldsymbol{x}^2)/50 \tag{25}$$

$$f_2(\boldsymbol{x}) = 10 - \sum_{i=1}^{2}\left[x_i^2 - 5\cos(2\pi x_i)\right] \tag{26}$$

其中,f_1 是 Viana 函数,f_2 为 Rastrigin 函数,且观测误差均为 $\boldsymbol{\epsilon}\sim \mathrm{N}(0, 0.1)$。

其次,本文考虑 5 类子代理模型及其加权模型,其中 2 类属于线性回归模型,3 类属于高斯回归模型。子代理模型的具体形式如表 1 所示。

因此,给定试验设计($\boldsymbol{X}$, $\boldsymbol{Y}$),本文比较了 9 类代理模型,即 5 类子代理模型 $M_1 \sim M_5$(如表 1 所示);基于 PE 的加权模型 M_6;基于 GMSE 的加权模型 M_7;基于优化 GMSE 的加权模型 M_8;本文方法 M_9。同时,对于每个算例,均随机生成包含 10 000 个样本点的测试集($\boldsymbol{X}^v$, $\boldsymbol{Y}^v$),并基于测试集比较两种指标:RMSE 和经验累积分布函数(Empirical Cumulative Distribution Function, ECDF)。具体计算结果如下。

表 1 5 类子代理模型

Tab. 1 Five different stand-alone surrogate models

模型分类	方法	数学表达式
线性回归模型	多项式基	$\sum \beta_i x^i$
	MARS 模型	$\sum \beta_i B_i(x)$
高斯回归模型	多项式核函数	$\sigma_{\mathrm{f}}^2 (c + xz)^d$
	高斯核函数	$\exp\left(\frac{(x-z)^2}{\sigma_{\mathrm{f}}^2}\right)$
	Matérn 3/2 核函数	$\sigma_{\mathrm{f}}^2\left[1 + \frac{\sqrt{3}(x-z)}{\sigma_1}\right]\times \exp\left[-\frac{\sqrt{3}(x-z)}{\sigma_1}\right]$

4.1 Viana 函数

假设 $\boldsymbol{x}\sim \mathrm{N}(0,1)$,则利用拉丁超立方体抽样方法(Latin Hypercube Sampling, LHS)生成包含 7 个样本的试验设计($\boldsymbol{X}$, $\boldsymbol{Y}$)。对每类代理模型总共进行 50 次独立试验,则其 RMSE 对比如图 1 所示。

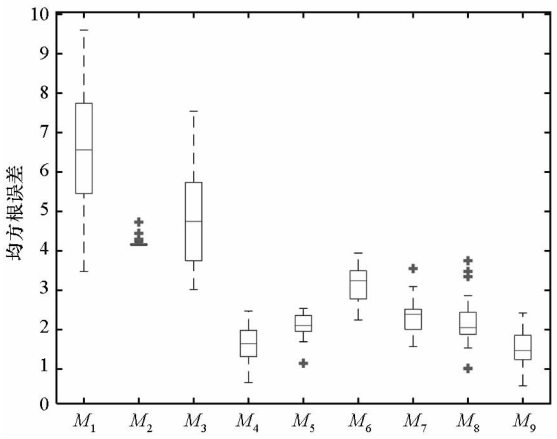

图 1 9 类代理模型 RMSE 对比

Fig. 1 Comparison of RMSE of 9 surrogate models

由图 1 可以看出,对于 5 类子代理模型来说,M_4 的逼近精度最高;对于 4 类加权模型来说,本文方法,即 M_9 的逼近精度也是最高的;总体上,本文方法的精度与 M_4 精度相近。图 2 利用 1000 个样本的测试集生成经验累积分布函数。

图 2 中可以看到,对比于各子模型,加权模型的 ECDF 都能够较好地反映真实响应的分布情形。同样地,可以看出 M_9 与真实响应之间的差距更小,在 $y>5$ 的区域上基本与真实 ECDF 重叠。

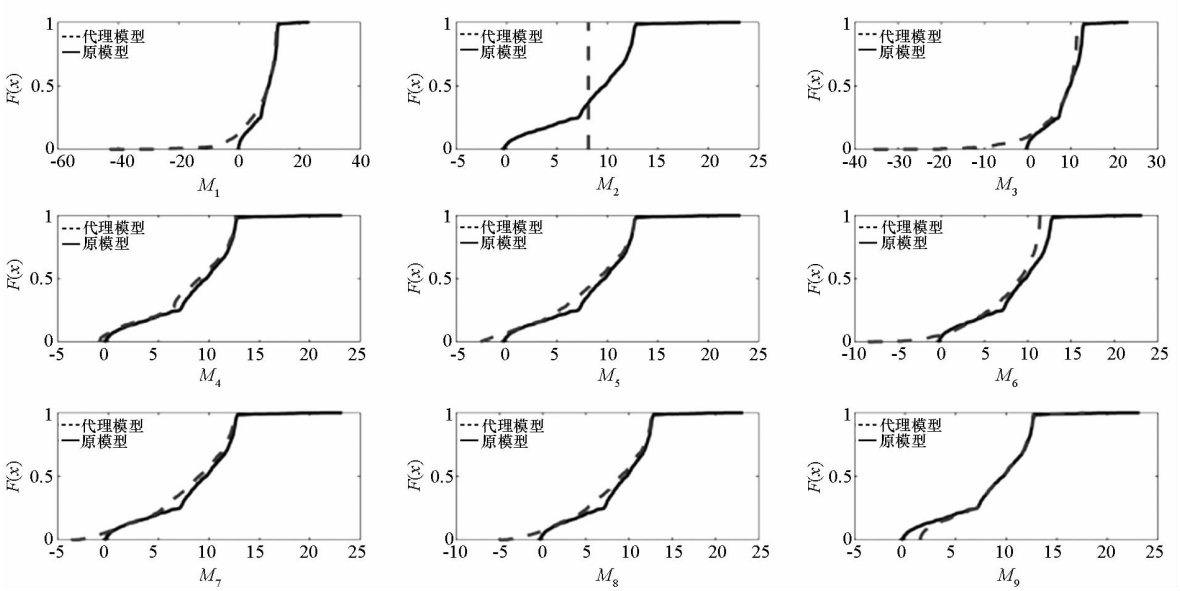

图 2　9 类代理模型 ECDF 对比

Fig. 2　Comparison of ECDF of 9 surrogate models

4.2　Rastrigin 函数

假设 $x_i \sim \mathrm{N}(0,1)$, $i=1,2$,同样利用拉丁超立方体抽样方法(Latin Hypercube Sampling, LHS)生成包含 20 个样本的试验设计($\boldsymbol{X}$, $\boldsymbol{Y}$),并对每类代理模型总共进行 50 次独立试验,则其 RMSE 对比如图 3 所示。

由图 3 可以看出,对于 5 类子代理模型来说,除了 M_1 以外,$M_2 \sim M_5$ 逼近精度类似,其中 M_2, M_3 的整体预测性能更加稳定;对于 4 类加权模型来说,其预测性能与 $M_2 \sim M_5$ 相当,M_9 的逼近精度略高,在某些试验设计条件下具有最高的精度。图 4 与图 2 类似,同样利用 1000 个样本的测试集生成经验累积分布函数,其结果如图 4 所示。

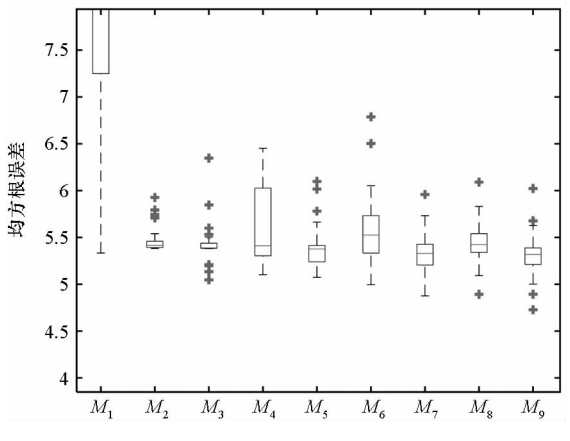

图 3　9 类代理模型 RMSE 对比

Fig. 3　Comparison of RMSE of 9 surrogate models

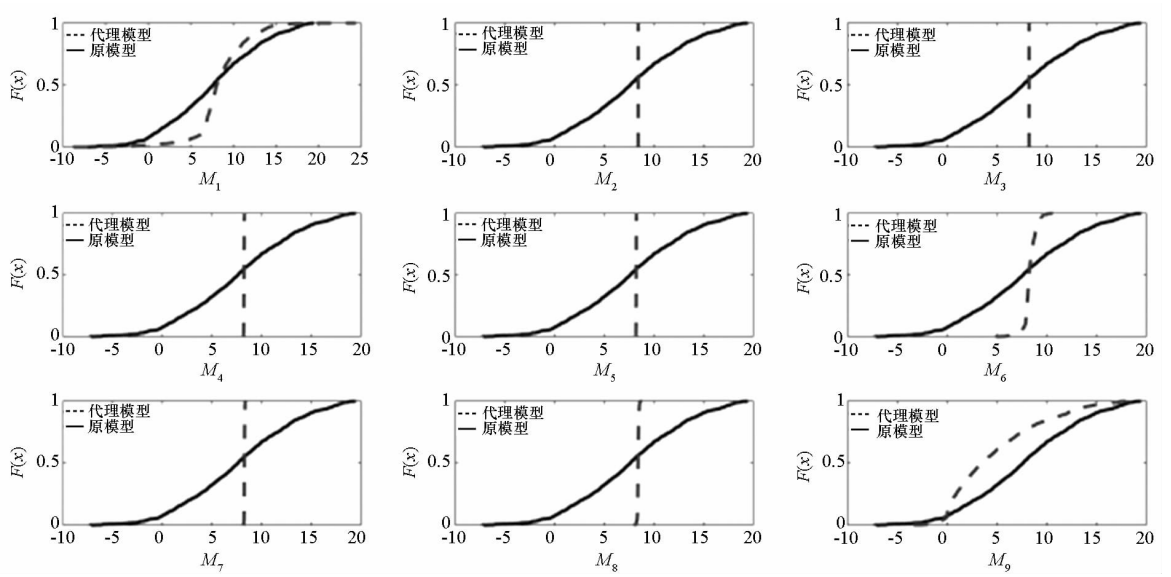

图 4　9 类代理模型 ECDF 对比

Fig. 4　Comparison of ECDF of 9 surrogate models

由图4可以看出，除了 M_1，M_9 以外，其他代理模型对真实响应分布逼近并不理想。同时注意到，对这些模型，其数据分布集中在 $y=8.4$ 附近；作为对比，M_9 仍然能够较好地还原真实响应的分布。

最后对一个具体的案例进行测试，即悬臂梁(cantilever beam)问题[12]，如图5所示。

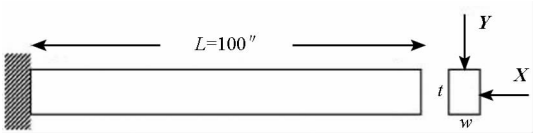

图5 悬梁臂问题

Fig. 5 Cantilever beam problem

其中，w 和 t 分别代表横截面的宽度和厚度，$\boldsymbol{Y}$ 表示垂直载荷，$\boldsymbol{X}$ 表示水平载荷，L 代表长度。对于悬臂梁问题来说，总共有两个响应，即位移 $D(\boldsymbol{x})$ 和应力 $S(\boldsymbol{x})$。具体数学表达式如下：

$$\begin{cases} D(\boldsymbol{x}) = \dfrac{4L^3}{Ewt}\sqrt{\left(\dfrac{\boldsymbol{Y}}{t^2}\right)^2+\left(\dfrac{\boldsymbol{X}}{w^2}\right)^2} \\ S(\boldsymbol{x}) = \dfrac{600\boldsymbol{Y}}{wt^2}+\dfrac{600\boldsymbol{X}}{w^2t} \end{cases} \tag{27}$$

式中总共有三个自变量，即杨氏模量(Young's modulus) $E\sim \mathrm{N}(2.9\times10^7,\ 1.45\times10^6)$，垂直载荷 $\boldsymbol{Y}\sim\mathrm{N}(1000,100)$，水平载荷 $\boldsymbol{X}\sim\mathrm{N}(500,100)$。令 $w=4$，$t=2$，同样利用LHS方法生成包含36个样本的试验设计，并对9类代理模型总共进行50次独立试验，则对于应力响应来说，其RMSE对比如图6所示。

由图6可以看出，对于5类子代理模型来说，M_4、M_5 的预测精度较低，其RMSE远远大于200，

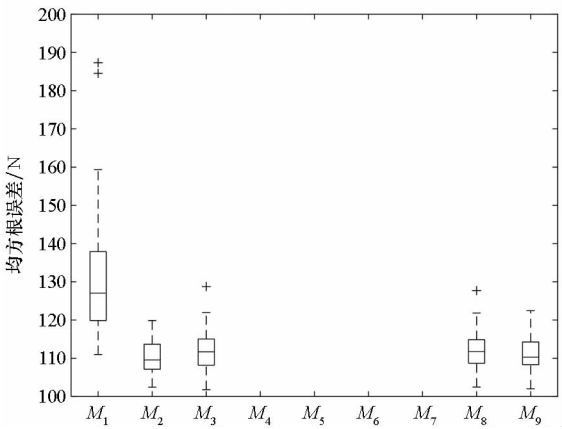

图6 应力RMSE对比

Fig. 6 Comparison of RMSE of stress

因此未在图中展示。对于加权模型来说，M_6、M_7 的效果同样较差，M_8、M_9 的逼近性能相当，与最佳子模型 M_2 的精度也在同一量级。对于ECDF来说，同样利用1000个样本的测试集生成经验累积分布函数，其结果如图7所示。

事实上，由图7可以看出，子模型 M_4、M_5 对本问题及试验设计的应用效果较差，而 $M_1\sim M_3$ 均能够较好地估计应力的分布。对于加权模型来说，其对各子模型的"平均"作用都能使得模型更加稳健，如 M_6、M_7，一定程度上还原了真实分布，但在性能上有着较大差异。M_8、M_9 能够极大地利用"好"模型($M_1\sim M_3$)的信息，从而其对ECDF的逼近效果也是最好的。对于位移 $D(\boldsymbol{x})$ 来说，由于 M_3 不能应用在该模型之上，因此本文仅比较其余8类代理模型。注意到此时加权模型为 M_1、M_2、M_4、M_5 四类模型的加权。使用与应力相同的试验设计，得到关于位移的RMSE如图8所示。

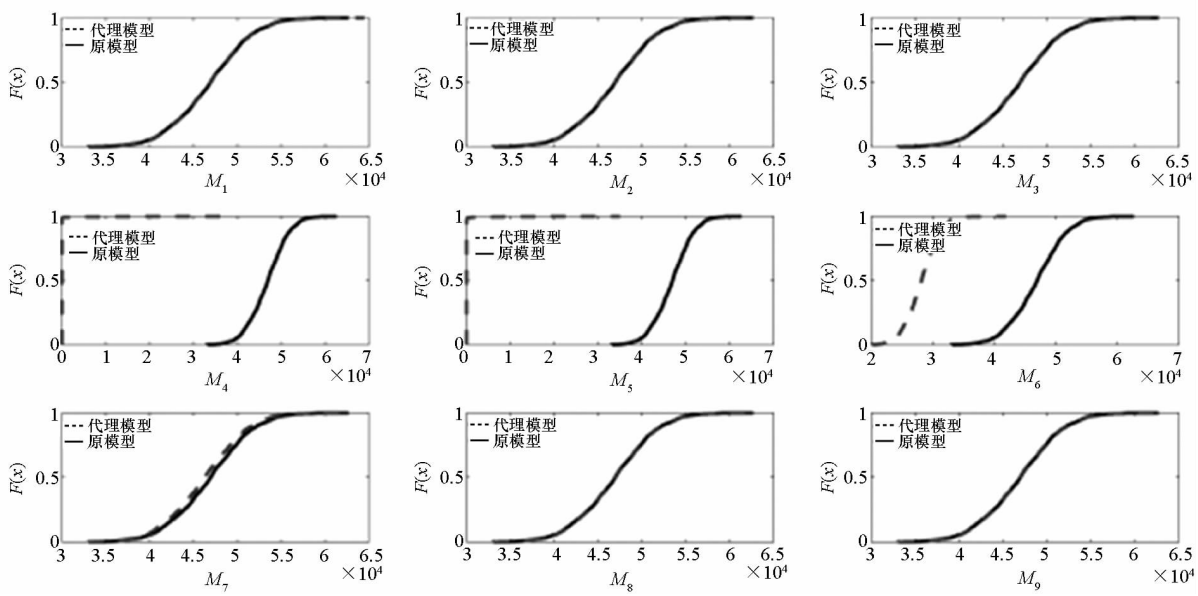

图7 应力ECDF对比

Fig. 7 Comparison of ECDF of stress

由图 8 可以看出,对于 4 类子代理模型来说,M_4 ~ M_5 其预测精度同样较低,因此也未在图中展示。对于加权模型来说,M_6 的效果最差,M_7、M_8、M_9 的逼近性能依次递增,一般情况下,M_9 的效果要比单独的代理模型都要好。利用 1000 个测试样本生成相关的 ECDF,其对比如图 9 所示。

图 9 与图 7 类似,对比于各子模型,加权模型至少避免了类似 M_4、M_5 的模型选择错误,其在一定程度上能够较为稳健地逼近原模型。在所有加权类型中,本文方法与 M_8 对 ECDF 的逼近是最准确的。

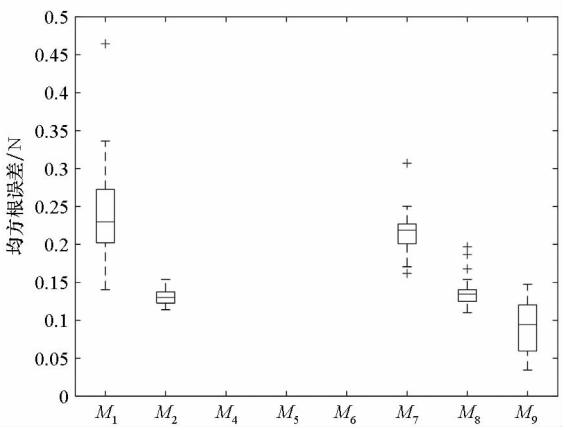

图 8　位移 RMSE 对比图

Fig. 8　Comparison of RMSE of Displacement

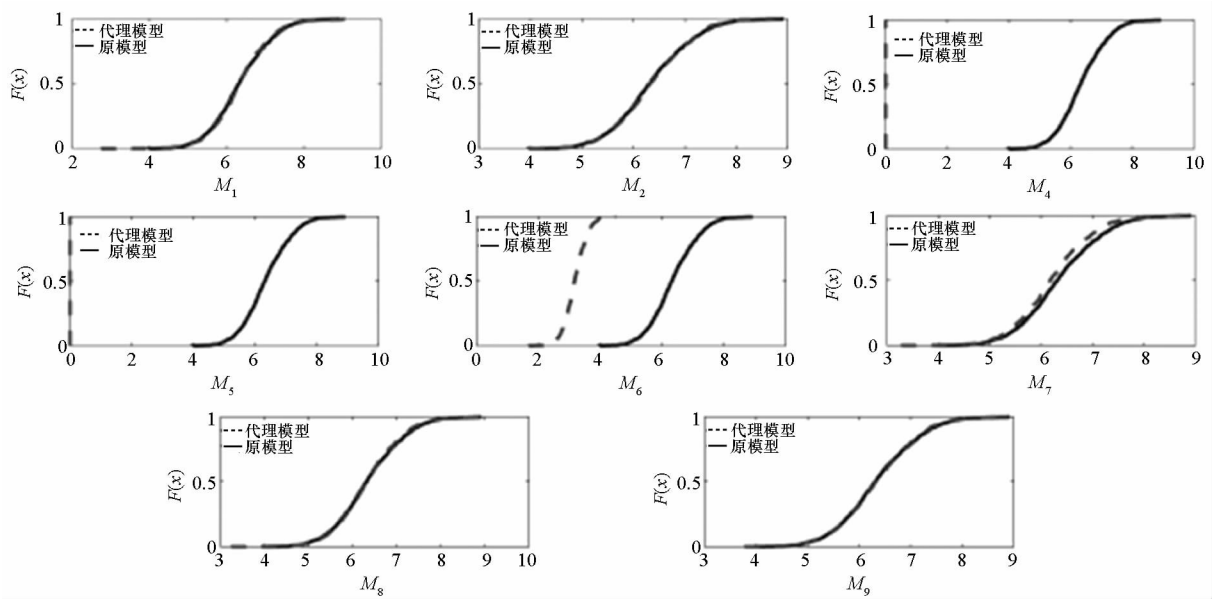

图 9　位移 ECDF 对比

Fig. 9　Comparison of ECDF of displacement

5　结论

与通常加权模型选择降低加权模型精度不同,本文针对预测分布构造权函数,引入 Kullback-Leibler 距离刻画模型与模型之间预测分布的差异性。为了克服真实分布无法获得的问题,本文利用各模型之间的离散度来表征各模型的预测能力。算例分析表明,本文方法在兼顾模型预测精度的同时,也能够较好地还原真实响应的分布。

参考文献(References)

[1] Law A M, Kelton W D. Simulation modelling and analysis[M]. USA: McGraw-Hill Education, 2007.

[2] Koziel S, Leifsson L. Surrogate based modelling and optimization[M]. USA: Springer-Verlag New York, 2013.

[3] Meckesheimer M, Booker A J, Barton R R, et al. Computationally inexpensive metamodel assessment strategies[J]. AIAA Journal, 2002, 40(10): 2053 – 2060.

[4] Myers R H, Montgomery D C, Vining G G, et al. Response surface methodology: a retrospective and literature survey[J]. Journal of Quality Technology, 2004, 36(1): 53 – 77.

[5] Friedman J H. Multivariate adaptive regression splines[J]. The Annals of Statistics, 1991, 19(1): 1 – 67.

[6] Rasmussen C E. Gaussian processes in machine learning[C]// Proceedings of Advanced Lectures on Machine Learning, 2004: 63 – 71.

[7] Seber G A F, Lee A J. Linear regression analysis[M]. USA: John Wiley & Sons, 2012.

[8] Zerpa L E, Queipo N V, Pintos S, et al. An optimization methodology of alkaline surfactant polymer flooding processes using field scale numerical simulation and multiple surrogates[J]. Journal of Petroleum Science and Engineering, 2005, 47(3/4): 197 – 208.

[9] Goel T, Haftka R T, Shyy W, et al. Ensemble of surrogates[J]. Structural and Multidisciplinary Optimization, 2007, 33(3): 199 – 216.

[10] Acar E, Rais-Rohani M. Ensemble of metamodels with optimized weight factors[J]. Structural and Multidisciplinary Optimization, 2009, 37(3): 279 – 294.

[11] Hershey J R, Olsen P A. Approximating the Kullback Leibler divergence between Gaussian mixture models [C]// Proceedings of Acoustics, Speech and Signal Processing, 2007: Ⅳ – 317 – Ⅳ – 320.

[12] Wu Y T, Shin Y, Sue R H, et al. Safety-factor based approach for probability-based design optimization [C]// Proceedings of 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 2001: AIAA – 2001 – 1522.