工程科学学报
Chinese Journal of Engineering

基于变量选择的尖点突变模型的两步构建方法

张明 付冬梅 程学群 杨丙坤 郝文魁 陈云 邵立珍

A two-step method for cusp catastrophe model construction based on the selection of important variables

ZHANG Ming, FU Dong-mei, CHENG Xue-qun, YANG Bing-kun, HAO Wen-kui, CHEN Yun, SHAO Li-zhen

引用本文:
张明, 付冬梅, 程学群, 杨丙坤, 郝文魁, 陈云, 邵立珍. 基于变量选择的尖点突变模型的两步构建方法[J]. 工程科学学报, 2023, 45(1): 128–136. doi: 10.13374/j.issn2095–9389.2021.07.19.006
ZHANG Ming, FU Dong-mei, CHENG Xue-qun, YANG Bing-kun, HAO Wen-kui, CHEN Yun, SHAO Li-zhen. A two-step method for cusp catastrophe model construction based on the selection of important variables[J]. *Chinese Journal of Engineering*, 2023, 45(1): 128–136. doi: 10.13374/j.issn2095–9389.2021.07.19.006

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2021.07.19.006

工程科学学报, 第 45 卷, 第 1 期: 128−136, 2023 年 1 月
Chinese Journal of Engineering, Vol. 45, No. 1: 128−136, January 2023
https://doi.org/10.13374/j.issn2095-9389.2021.07.19.006; http://cje.ustb.edu.cn

基于变量选择的尖点突变模型的两步构建方法

张　明[1)]，付冬梅[1,2,3)✉]，程学群[4,5)✉]，杨丙坤[6)]，郝文魁[6)]，陈　云[6)]，邵立珍[1,2)]

1) 北京科技大学顺德研究生院, 佛山 528300　2) 北京科技大学自动化学院, 北京 100083　3) 北京科技大学北京市工业波谱成像工程技术研究中心, 北京 100083　4) 北京科技大学新材料技术研究院, 北京 100083　5) 北京科技大学国家材料腐蚀与防护科学数据中心, 北京 100083　6) 全球能源互联网研究院有限公司先进输电技术国家重点实验室, 北京 102209

✉通信作者，付冬梅，E-mail: fdm_ustb@ustb.edu.cn; 程学群, chengxuequn@ustb.edu.cn

摘　要　突变是工程实践过程中广泛存在的现象. 当系统的状态发生跳跃性变化时, 基于微积分的传统数学建模方法精度较低, 人工神经网络等机器学习算法无法对突变现象作出合理的解释. 基于突变理论的尖点突变模型可以用来解释系统状态的不连续变化, 然而在输入变量维度较大的情况下, 传统的尖点突变模型复杂度高且精度较差. 为了解决这一问题, 提出了一种基于变量选择的尖点突变模型的两步构建方法. 第一步, 利用多模型集成重要变量选择算法 (MEIVS) 量化待选变量的重要性并提取重要变量; 第二步, 基于极大似然法 (MLE) 利用所提取的重要变量构建尖点突变模型. 仿真结果表明, 在具有突变特征的数据集上, 通过 MEIVS 降维后的尖点突变模型在评价指标上优于线性模型、Logistic 模型和通过其他方法降维的尖点突变模型, 并且可以用来解释研究对象的不连续变化.

关键词　突变理论; 突变特征; 尖点突变模型; 变量选择; 模型集成

分类号　O192;TP181

A two-step method for cusp catastrophe model construction based on the selection of important variables

ZHANG Ming[1)], *FU Dong-mei*[1,2,3)✉], *CHENG Xue-qun*[4,5)✉], *YANG Bing-kun*[6)], *HAO Wen-kui*[6)], *CHEN Yun*[6)], *SHAO Li-zhen*[1,2)]

1) Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China

2) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

3) Beijing Engineering Research Center of Industrial Spectrum Imaging, University of Science and Technology Beijing, Beijing 100083, China

4) Institution for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China

5) National Materials Corrosion and Protection Scientific Data Center, University of Science and Technology Beijing, Beijing 100083, China

6) State Key Laboratory of Advanced Transmission Technology, Global Energy Interconnection Research Institute Limited Company, Beijing 102209, China

✉ Corresponding author, FU Dong-mei, E-mail: fdm_ustb@ustb.edu.cn; CHENG Xue-qun, chengxuequn@ustb.edu.cn

ABSTRACT　Sudden transition is a widely existing phenomenon in engineering practice. When the state of the system experiences sudden abrupt transition, calculus-based traditional mathematical modeling methods has low accuracy. Although theoretically, machine learning algorithms, such as artificial neural networks, can approximate any nonlinear function, this type of black-box method makes no reasonable explanation for the sudden transition phenomenon. The cusp catastrophe model based on the catastrophe theory can be applied to explain the discontinuous changes in the system's state. However, the construction of traditional cusp catastrophe models is often based on large amounts of prior knowledge to select the input variables for modeling. On the condition that there is a lack of prior

收稿日期: 2021−07−19

基金项目: 科技部科技基础资源调查专项资助项目(2019FY101404); 国家电网公司总部科技资助项目(5200-202058470A-0-0-00); 北京科技大学顺德研究生院科技创新基金资助项目(BK20AE004)

knowledge and comparatively large dimensions of input variables, the model has high complexity and poor accuracy. In this paper we have put forward a two-step method for constructing a cusp catastrophe model based on the selection of variables to solve the abovementioned problems. The first step was to apply multimodel ensemble important variable selection (MEIVS) to quantify the importance of the variables to be selected and extract important variables. The second step was to use the extracted important variables to construct a cusp catastrophe model based on the framework of maximum likelihood estimation (MLE). Results indicate that on a dataset with characteristics of catastrophe, the cusp catastrophe model is simple in form using the MEIVS dimensionality reduction algorithm and outperforms the unreduced cusp catastrophe model and reduced cusp catastrophe model using other dimensionality reduction algorithms in terms of evaluation indicators. This shows that the algorithm proposed in this paper have improved the accuracy and reduced the complexity of the cusp catastrophe model. At the same time, the cusp catastrophe model exhibits higher accuracy compared with the linear and logistic models. Thus, it can be used to explain the discontinuous changes of the research object, and it has a practical engineering significance.

KEY WORDS　catastrophe theory; catastrophe flag; cusp catastrophe model; variable selection; model integration

在复杂的系统中，外界因素的变化可能会导致系统状态的跳跃式变化，称为突变．Qiao 等[1] 认为深埋隧道围岩的失稳是一种突变现象，给矿井的安全生产带来极大威胁，并通过分析得出围岩失稳发生在第 3 至 4 步开挖过程中．Zhi 等[2] 通过分析得到使腐蚀速率急剧变化的环境变量的阈值，当环境变量超过该阈值时，腐蚀速率会发生突变．裴甲坤等[3] 认为化工事故是由危险源和不安全因素引起的突发事件，二者的综合影响导致系统的安全状态发生突变．其他诸如股市崩盘[4−5]、人的心理状态变化[6]、电力系统故障[7] 等也属于突变现象．此类现象包含复杂的系统行为，既有连续性变化又有突发的不连续性变化，且影响因素往往众多，给实际工程问题的建模和解释带来困难．

对于小样本工程数据，线性模型、灰色模型[8] 是常用的方法；对于大样本工程数据，人工神经网络[9]、随机森林 (Random forest, RF)[10] 等机器学习模型通常可以获得较好的建模效果．虽然机器学习模型具有强大的非线性映射能力，但无法解释研究对象的突变现象．突变理论是用以解释复杂系统中不连续性和质变现象的数学理论，由法国数学家 Thom[11] 提出．假定一个系统的动力学方程可以由一个光滑的势函数导出，根据控制因子和状态因子个数的不同，Thom 定义了 7 种基本的突变模型，并推导出每一种模型势函数的解析形式．由于形式简单、直观，具有两个控制因子和一个状态因子的尖点突变模型应用最为广泛，模型参数的估计可由 Cobb 提出的极大似然估计法 (Maximum likelihood estimation, MLE) 实现[12−13]．作为解决工程领域中不连续性复杂问题的一种数学工具，突变理论在经济学、生物学、物理学、心理学等领域应用广泛．

在以往的尖点突变模型中，组成控制因子的输入变量往往依据经验或已有的结论来确定．如，文献 [4−5] 基于 Zeeman[14] 的理论基础，将股票市场中基本面交易者和技术分析交易者的多维数据作为输入变量构建股票市场的尖点突变模型．这种建模方式受限于特定学科，不利于推广，且输入变量的实际价值难以判断．在待选的输入变量较多且突变机理不明确的情况下，如何利用少量的重要变量构建尖点突变模型依然是一个难点．常见的变量选择方法分为过滤法、嵌入法、封装法．其中，过滤法依据待选变量统计特性的各项指标来选择重要变量，如皮尔逊相关系数法、方差过滤法；嵌入法依据机器学习算法本身来分析待选变量的重要性，如 RF 的排列变量重要性算法[10,15]；封装法基于构造的最终模型来选择使模型性能达到最优的变量子集，最终模型可以是支持向量机 (Support vector regression, SVR)[16]、梯度提升回归树 (Gradient boosted regression trees, GBRT)[17] 等机器学习算法．过滤法的评价标准独立于特定的学习算法，具有较好的通用性，但难以取得很好的建模效果；嵌入法、封装法虽然可以取得较好的建模效果，但是这种基于单一模型的变量选择算法存在特定的偏差，变量子集的选取依赖于特定模型，容易产生过拟合现象．采用集成方法，即通过组合不同方法的变量选择结果来产生变量子集，既减轻了对特定模型的依赖性，又可以很好地提高结果的准确性和稳定性[18−20]．

针对传统尖点突变模型依据经验建模的问题，提出基于变量选择的尖点突变模型的两步构建方法．该方法的通用性较强，可广泛应用于具有突变特征的系统的建模并能得到模型的数学解析式．建模过程分为两步．第一步，以 RF、GBRT、

SVR 作为基学习器, 利用多模型集成重要变量选择算法 (Multi-model ensemble important variable selection, MEIVS) 来量化待选变量的重要性, 提取得分之和超过总分 90% 的前 n 个待选变量作为后续建模的输入变量; 第二步, 基于 MLE 算法构建尖点突变模型. 本文首先介绍了尖点突变模型的原理、数据拟合方法以及突变特征, 其次介绍了 MEIVS 算法的实现流程, 最后结合工程实例, 验证了该方法的有效性.

1　基本原理

1.1　尖点突变模型与突变特征

1.1.1　尖点突变模型

突变理论描述了动力学系统中控制因子和状态因子之间的关系, 在控制因子固定的情况下, 系统始终寻求平衡状态, 直到达到势函数的极小值或极大值为止. 以动力学系统表达式来描述系统的状态因子 z 在控制因子 a 的影响下随时间 t 的变化:

$$\frac{\mathrm{d}z}{\mathrm{d}t}=-\frac{\partial V(z;a)}{\partial z} \tag{1}$$

$V(z;a)$ 是系统的势函数. 应用最广泛的尖点突变模型由 2 个控制因子 α、β 和一个状态因子 z 组成, 其势函数的规范形式是:

$$V(z;\alpha,\beta)=\frac{1}{4}z^4-\frac{1}{2}\beta z^2-\alpha z \tag{2}$$

系统的平衡方程由 (3) 式确定, 在无扰动的情况下, 系统的状态不随时间变化:

$$\frac{V(z;\alpha,\beta)}{\partial z}=z^3-\beta z-\alpha=0 \tag{3}$$

当平衡点的势函数 $V(z;\alpha,\beta)$ 是关于 z 的极小值时, 平衡点是稳定的, 系统即使受到扰动的影响, 也会随着时间 t 回到稳定状态; 当平衡点的势函数 $V(z;\alpha,\beta)$ 是关于 z 的极大值时, 平衡点是不稳定的, 系统在扰动的影响下会偏离此平衡点, 从而被稳定的平衡点吸引. 在不同的 α 和 β 值下系统平衡点的数目和性质可以由 Cardan 判别式 δ 判断, 表示为:

$$\delta=27\alpha^2-4\beta^3 \tag{4}$$

当 $\delta>0$ 时, 存在一个稳定的平衡点; 当 $\delta<0$ 时, 存在两个稳定的平衡点和一个不稳定的平衡点; 当 $\delta=0$ 时, 存在一个稳定的平衡点和一个不稳定的平衡点. 图 1 给出了由平衡点的集合构成的平衡曲面和由控制因子构成的控制平面. 平衡曲面的形状像一个有“褶皱”的连续曲面, 并且由上叶、中叶、下叶 3 部分构成, 上叶和下叶部分对应的平衡点是稳定的, 中叶部分对应的平衡点是不稳定的. 控制平面是平衡曲面在 z 轴方向上的投影, 中叶区域在控制平面上的投影称为尖点突变模型的分叉集. 图 1 中, 若控制因子 α、β 沿红色轨迹 A 变化, 状态因子 z 会在分叉集内发生突变, 从平衡曲面的下叶直接跳变到上叶而不经过中叶; 若控制因子 α、β 沿蓝色轨迹 B 变化, 则状态因子 z 不会发生突变.

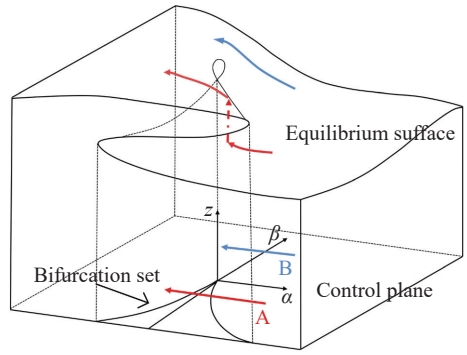

图 1　尖点突变模型的平衡曲面和控制平面

Fig.1　Equilibrium surface and control plane of the cusp catastrophe model

在实际应用中, 数据难免受到随机噪声影响, Cobb 和 Zacks[12−13] 通过引入随机微分方程, 以概率密度函数的形式描述了系统在 α、β 固定的条件下 z 的分布, 表示为:

$$f(z|\alpha,\beta)=\psi \mathrm{e}^{\left(\alpha z+\frac{1}{2}\beta z^2-\frac{1}{4}z^4\right)} \tag{5}$$

式中, ψ 是归一化常数, α、β 分别为 n 维输入变量 $\{X_1,\cdots,X_n\}$的线性组合, z 为输出变量 Y 的线性变换, 表示为:

$$\begin{cases}\alpha=a_0+\sum\limits_{i=1}^{n}a_iX_i\\ \beta=b_0+\sum\limits_{i=1}^{n}b_iX_i\\ z=w_0+w_1Y\end{cases} \tag{6}$$

参数 $\theta=\{w_0,w_1,a_0,\cdots,a_n,b_0,\cdots,b_n\}$由 MLE 方法估计. 在给定 N 个观测样本的情况下, 参数 θ 的对数似然函数如下:

$$L(\theta)=\prod_{i=1}^{N}f(z_i|\theta,\alpha_i,\beta_i) \tag{7}$$

最大化观测样本的对数似然函数得到参数 θ 的估计值, 优化搜索算法为带边界约束的 Bro-

yden–Fletcher–Goldfarb–Shanno 算法：

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \tag{8}$$

1.1.2　评价指标

为了验证尖点突变模型的性能，将线性模型和非线性的 Logistic 模型与尖点突变模型作对比[21]. 其中，Logistic 模型可以模拟研究对象的急剧变化，但没有考虑不连续性变化. 评价指标采用可决系数，R^2、赤池信息准则 (Akaike information criterion, AIC) 和贝叶斯信息准则 (Bayesian information criterion, BIC). 当 R^2 越大时，模型的精度越高；AIC 和 BIC 考虑了模型复杂度，当 AIC 和 BIC 越小时，模型越好.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2} \tag{9}$$

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{10}$$

$$\mathrm{BIC} = k\ln(N) - 2\ln(L) \tag{11}$$

式中，$\hat{y}_i$为第 i 个样本的预测值，y_i为第 i 个样本的实际值，$\bar{y}$为所有样本实际值的均值，k 是模型参数个数，L 是最大对数似然值. 模型选择标准通常基于最低的 AIC 和 BIC 值，并以 R^2 作为参考. 对于尖点突变模型，当 α、β 位于分叉集内部时，根据延迟约定，z 的预测值在离实际值最近的平衡曲面上[21−22].

1.1.3　突变特征

在系统的势函数未知的情况下，常常根据系统表现的外部性态来判断系统是否存在突变，这些性态被称为突变特征[14,21]. 尖点突变有 5 个特征：(1) 多模态：系统中可能出现两个不同的状态；(2) 不可达性：系统存在不稳定的平衡态；(3) 突跳：系统从一个势函数极小值跳到另一个极小值；(4) 发散：控制因子的微小变化可以导致状态因子的质变；(5) 滞后：当物理过程可逆时，发生突变时对应的控制参数位置可能不同. 当系统存在突变现象时，对外往往表现为其中的一个或几个的组合. 在实际应用中，针对截面数据，应首先检查研究对象概率密度的双峰性，双峰性意味着系统可能存在多个状态；针对时序数据，则应首先检查时间序列中的跳变现象[21].

1.2　多模型集成重要变量选择算法

而在传统的尖点突变模型的建模过程中，输入变量的选取往往依赖于已有的实践或经验，这与目前数据规模的爆发式增长相矛盾，不利于尖点突变模型的普及应用. 为了解决上述问题，同时提高模型的精度、降低模型的复杂度，本文基于排列[23] 的思想提出 MEIVS 算法.

排列的思想借鉴于随机森林的变量重要性度量方法，认为模型会更依赖于重要的输入变量做预测. 当打乱某一变量在测试集上的观测序列后，用新生成的数据做预测，更重要的输入变量会使模型的精度损失更大. MEIVS 算法组合了 RF、GBRT、SVR 3 种常用的机器学习算法，其中 RF 和 GBRT 都属于决策树的集成学习算法，但它们采用的计算策略不同；SVR 采用高斯核函数. 文献 [24−26] 中对每种方法的机理都作了解释. 本文的损失函数采用的是均方根误差 (Root mean squared error, RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{12}$$

以样本的 80% 作为训练集，20% 作为测试集，使用 Z-Score 标准化方法对输入变量进行处理，经过处理的数据的均值为 0，标准差为 1. 记 m 个待选变量的集合为$\{S_1,\cdots,S_m\}$，目标是得到 n 个重要变量的集合$\{X_1,\cdots,X_n\}$作为尖点突变模型的输入变量. 算法步骤如下，流程图如图 2 所示.

步骤 1 利用训练集训练 RF、GBRT、SVR 模型，记为 M_1、M_2、M_3，对于所建立的每个模型 M_i，分别基于置换算法计算变量重要性，即执行步骤 2、步骤 3；

步骤 2 计算模型 M_i 在测试集上的均方根误差并记为L^{M_i}，对$\{S_1,\cdots,S_m\}$，依次执行 (1) ~ (3)：

(1) 打乱 S_j 在测试集上的观测序列并重新计算模型的均方根误差，由于涉及随机性，此过程重复 10 次，分别记为$L_{j1}^{M_i},\cdots,L_{j10}^{M_i}$；

(2) 计算 S_j 在测试集上的平均预测精度损失：

$$L_j^{M_i} = \frac{1}{10}\sum_{k=1}^{10}\left(L^{M_i} - L_{jk}^{M_i}\right) \tag{13}$$

(3) 计算 S_j 在模型 M_i 上的排列重要性得分，当 $L_j^{M_i} \leqslant 0$时，将重要性得分记为 0，该变量无用：

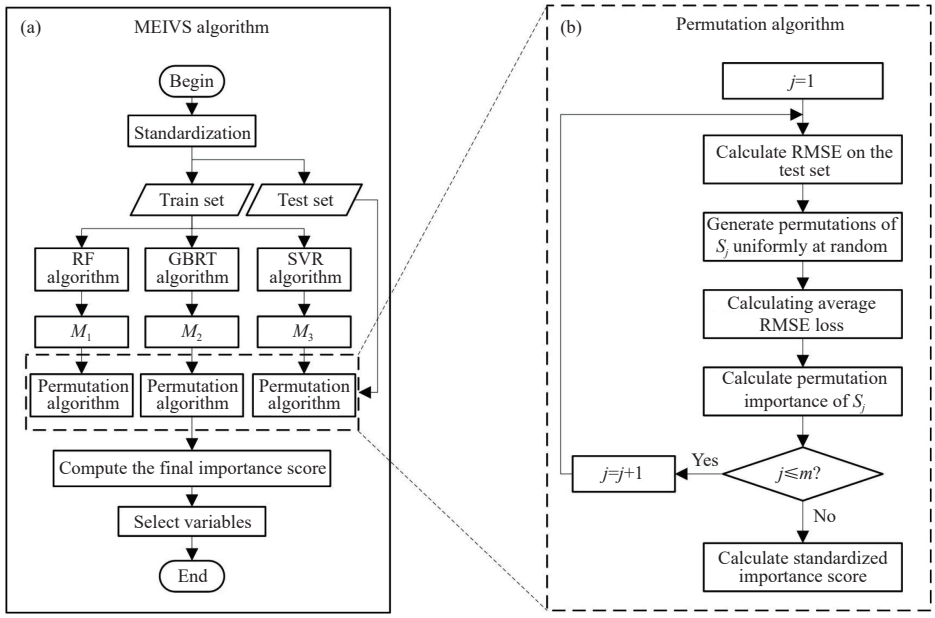

图 2　MEIVS 算法流程图. (a) MEIVS 算法主流程; (b) 排列算法流程

Fig.2　MEIVS algorithm flowchart: (a) main process steps of the MEIVS algorithm; (b) process steps of the permutation algorithm

$$V_j^{M_i}=\begin{cases}L_j^{M_i}, L_j^{M_i}>0\\ 0, L_j^{M_i}\leqslant 0\end{cases} \tag{14}$$

步骤 3 计算 S_j 的标准化排列重要性得分:

$$V'^{M_i}_j=\frac{V_j^{M_i}}{\sum_{k=1}^{m}V_k^{M_i}} \tag{15}$$

步骤 4 计算 S_j 在 M_1、M_2、M_3 上的重要性总得分:

$$V_j=V'^{M_1}_j+V'^{M_2}_j+V'^{M_3}_j \tag{16}$$

步骤 5 按变量重要性得分$\{V_1,\cdots,V_n\}$降序排列待选变量$\{S_1,\cdots,S_m\}$, 提取得分之和超过总分 90% 的前 n 个待选变量作为重要变量, 记为$\{X_1,\cdots,X_n\}$.

1.3　基于变量选择的尖点突变模型的两步构建方法

将 MEIVS 方法与基于 MLE 的尖点突变模型参数估计方法相结合, 分两步构建尖点突变模型. 第一步, 利用 MEIVS 来量化待选变量$\{S_1,\cdots,S_m\}$的重要性, 提取重要变量$\{X_1,\cdots,X_n\}$; 第二步, 利用提取的 n 个重要变量, 基于 MLE 算法构建尖点突变模型.

2　仿真结果和分析

以两个不同领域的、具有突变特征的数据集为例, 验证了所提方法的有效性. 其中, 欧洲旅馆住宿价格数据集[27]为截面数据集, 来源于 Kaggle 平台; 北京大气腐蚀数据集为时序数据集, 来源于北京地区的大气暴露实验.

2.1　在截面数据集上的应用——以欧洲旅馆住宿价格数据集为例

Kaggle 平台的欧洲旅馆住宿价格数据集一共包含 120 个样本, 每个样本包括每日住宿价格 (Price)、星级 (Star)、离市中心距离 (Distance)、评分 (Rating)、房间数目 (Room)、房间面积 (Square) 和所在城市 (City). Price 为输出变量, 单位为每日花费的可兑换马克 (KM·d^{-1}), 其余为输入变量, 其中类别变量 City 以 Price 的类别均值来编码. Price 概率密度的非参数估计如图 3, 非参估计的核函数选用高斯核, 带宽设置为 25, 概率密度的双峰性暗示了 Price 可能会发生突变, 因此适用于建立尖点突变模型.

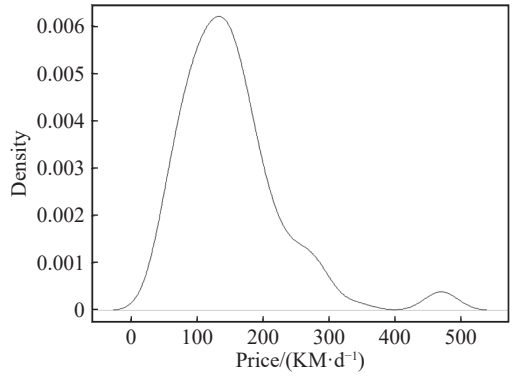

图 3　每日住宿价格的概率密度非参数估计

Fig.3　Nonparametric estimation of the probability density of the daily accommodation price

两步构建方法中第一步为提取重要变量. 利用 MEIVS 得到各个待选变量的重要性得分, 如图 4. 条形图中横轴表示影响每日住宿价格的待选变量, 纵轴表示每个待选变量的重要性总得分, 每个待选变量在各模型上的得分以不同的颜色区分, 并且根据得分降序排列. 依据 MEIVS 算法中步骤 (5), Square、Rating、Star 和 Room 为重要变量,设为 X_1、X_2、X_3、X_4, 每日住宿价格 Price 设为 Y. 算法基于 R 语言中的 DALEX 程序包[28] 实现.

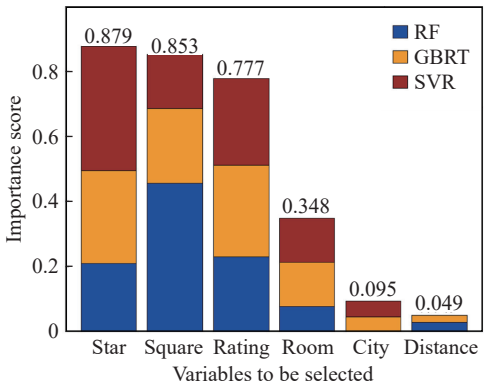

图 4　欧洲旅馆住宿价格数据集待选变量重要性得分

Fig.4　Importance score of the variables to be selected in the European hotel accommodation price dataset

将 MEIVS 提取的重要变量 X_1、X_2、X_3、X_4 作为输入变量、每日住宿价格 Y 作为输出变量建立尖点突变模型, 为了消除变量间量纲的影响, 用 Z-Score 标准化方法对原始输入变量进行处理. 算法基于 R 语言 Cusp 程序包[22] 实现. 利用 MLE 算法和 120 条样本对参数 $\theta=\{w_0,w_1,a_0,a_1,a_2,a_3,a_4,b_0,b_1,b_2,b_3,b_4\}$ 进行估计, 代入式 (6) 中, 得到如下形式的尖点突变模型的平衡方程:

$$\begin{cases}\alpha=-36.548+0.119X_1+2.670X_2+1.251X_3+0.339X_4\\ \beta=-2.811-0.018X_1+0.845X_2-0.357X_3-0.727X_4\\ z=-3.138+0.009Y\\ z^3-\beta z-\alpha=0\end{cases}\tag{17}$$

表 1 展示了采用两步构建法建立的尖点突变模型与经 MEIVS 降维后构建的线性模型、Logistic 模型的评价指标, 同时与传统的直接建模方法、经斯皮尔曼相关系数 (Spearman's correlation coefficient, SCC)、最大互信息系数 (Maximal information coefficient, MIC)、随机森林变量重要性算法 (Random forest variable importance measure, RFVIM) 降维的建模方法作比较. 其中, SCC 和 MIC 剔除系数小于 0.3 的弱相关变量, RFVIM 提取累计变量重要性达到 90% 的前 n 个变量. 结果显示, 在考虑样本量的情况下, 更高的 R^2 和更低的 BIC 说明基于两步构建法所构建的尖点突变模型优于未降维的传统尖点突变模型以及经 SCC、MIC、RFVIM 降维后所构建的尖点突变模型.

表 1　欧洲旅馆住宿价格数据集建模结果评价

Table 1　Evaluation of the modeling results of the European hotel accommodation price dataset

Model	Number of parameters	R^2	AIC	BIC
Linear		0.549	1306	1323
Logistic		0.626	1294	1324
Cusp (based on the two-step method)	12	0.727	195	228
Cusp (based on the traditional method)	16	0.697	190	235
Cusp (based on SCC)	10	0.572	204	232
Cusp (based on MIC)	6	0.421	234	251
Cusp (based on RFVIM)	12	0.565	210	243

图 5(a) 展示了样本在控制平面上的分布, 其中散点的颜色代表经过式 (6) 线性变换后旅馆价格的数值大小. 影响旅馆价格的控制因子的变化轨迹从左到右穿过了分叉集, 表明旅馆的价格发生了突变. 图 5(b) 展示了样本在平衡曲面上的分布, 平衡曲面设置为半透明状态, 颜色较暗的散点位于平衡曲面下方. 易观察到在较低的价格范围内旅馆价格的变化具有连续性, 而从低价到高价的变化并不连续.

2.2　在时序数据集上的应用——以北京大气腐蚀数据为例

北京大气腐蚀数据集一共包含 719 个样本, 采集时间为 2018 年 8 月 5 日 16 时至 9 月 6 日 14 时, 采集地点为北京, 每个样本包括大气腐蚀监测仪 (Atmospheric corrosion monitor, ACM) 采集得到的早期大气腐蚀电偶电流 (Galvanic current)、温度 (T)、相对湿度 (RH)、降雨状态 (Rainfall) 以及大气环境中 PM2.5、PM10、SO_2、NO_2、O_3 的浓度. 电偶电流与腐蚀速率成正相关关系[29], 为了便于分析, 取电偶电流的自然对数作为输出变量. 图 6 展示了电偶电流的时间序列表明腐蚀电偶电流波动较大, 表明时间序列中具有突变的特性, 因此适用于建立尖点突变模型.

通过 MEIVS 算法得到待选变量重要性得分如图 7,可见 T、RH 和 Rainfall 为影响早期大气腐蚀的重要变量, 设为 X_1、X_2、X_3, 对数化腐蚀电偶电流设为 Y. 其他污染物浓度的影响是微弱的.

以 Z-Score 标准化后的 X_1、X_2、X_3 作为输入变

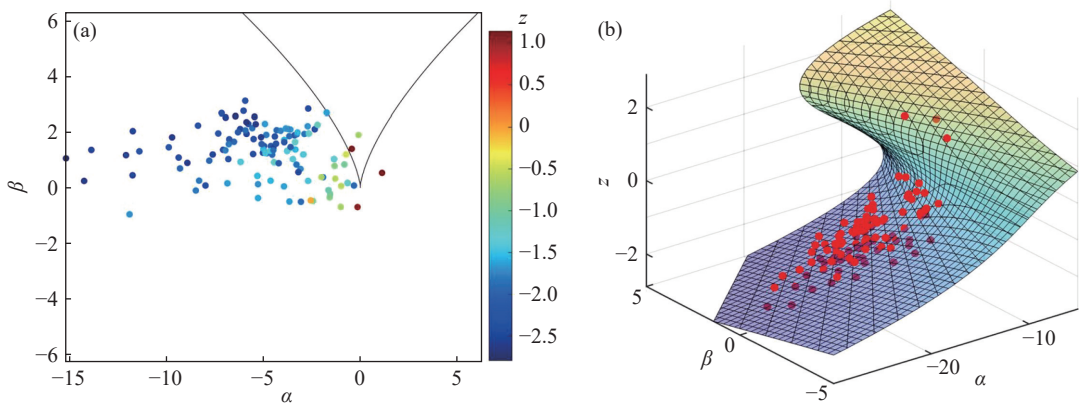

图 5　欧洲旅馆住宿价格数据在控制平面 (a) 和平衡曲面 (b) 上的分布

Fig.5　Distribution of the European hotel accommodation price dataset on the control plane (a) and equilibrium surface (b)

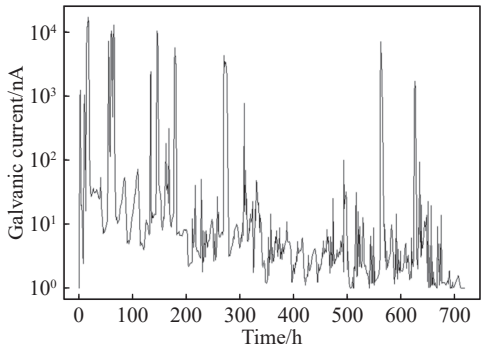

图 6　ACM 采集到的电偶电流时间序列

Fig.6　Time series of the galvanic current collected by ACM

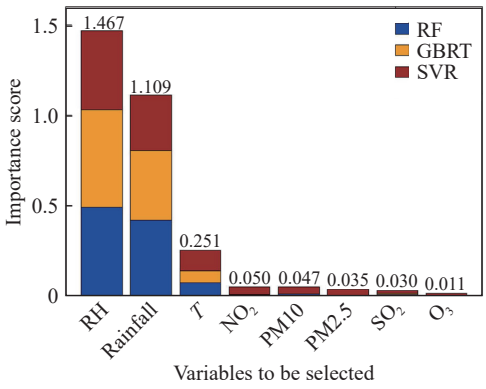

图 7　北京大气腐蚀数据集待选变量重要性得分

Fig.7　Importance score of the variables to be selected in the Beijing atmospheric corrosion dataset

量、Y 作为输出变量构建尖点突变模型, 利用 MLE 算法和 719 条样本对参数 $\theta=\{w_0,w_1,a_0,a_1,a_2,a_3,b_0,b_1,b_2,b_3\}$ 进行估计, 得到的平衡方程如下:

$$\begin{cases}\alpha=-4.978+0.130X_1+1.539X_2+1.421X_3\\ \beta=1.929-0.996X_1-1.697X_2+0.262X_3\\ z=-2.802+0.476Y\\ z^3-\beta z-\alpha=0\end{cases}\tag{18}$$

此外, 采用上文所述方法, 表 2 的模型评估结果显示了两步构建法的优越性.

表 2　北京大气腐蚀数据集建模结果评价

Table 2　Evaluation of the modeling results of the Beijing atmospheric corrosion dataset

Model	Number of parameters	R^2	AIC	BIC
Linear		0.668	2180	2203
Logistic		0.755	1970	2011
Cusp (based on the two-step method)	10	0.778	670	716
Cusp (based on the traditional method)	20	0.775	672	764
Cusp (based on SCC)	10	0.719	816	862
Cusp (based on MIC)	8	0.725	820	857
Cusp (based on RFVIM)	18	0.765	673	755

样本在控制平面和平衡曲面的分布情况如图 8(a)、8(b), 圆点代表未降雨时的样本, 三角形代表降雨时的样本. 当由温度、相对湿度、降雨组成的控制因子进入分叉集时, 腐蚀电偶电流在平衡曲面的下叶和上叶之间跳跃. 从图 8(a) 观测到, 降雨会促使腐蚀系统中的电偶电流不能沿着原有的轨迹运动, 而是突变到新的演变轨迹上.

3　结论

(1) 对于存在突变现象的系统, 通过理论和数据相结合的方式建立尖点突变模型是一种有效的建模手段.

(2) 提出了基于变量选择的尖点突变模型的两步构建方法. 在具有突变特征的数据集上, 相比于其他模型, 利用本文所提方法构建的尖点突变模型拟合效果更优.

(3) 结合样本在控制平面和平衡曲面的分布图, 尖点突变模型可以解释系统的突变行为.

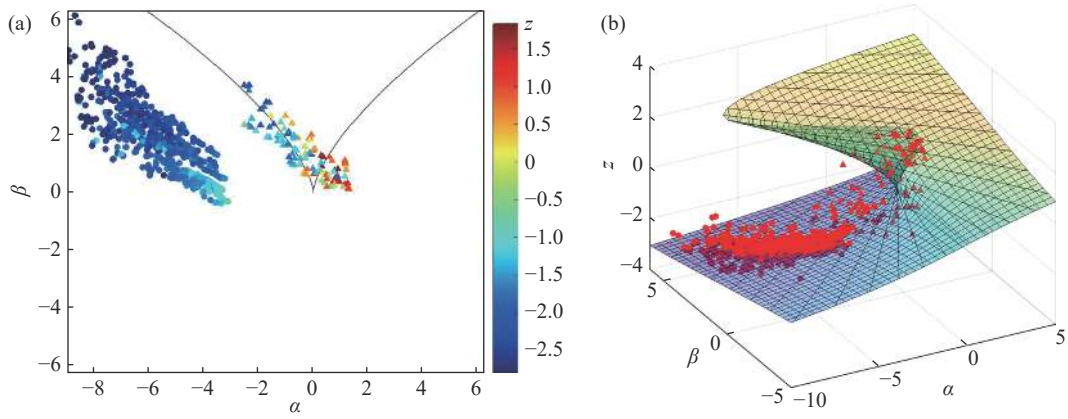

图 8　北京大气腐蚀数据在控制平面 (a) 和平衡曲面 (b) 上的分布

Fig.8　Distribution of the Beijing atmospheric corrosion dataset on the control plane (a) and equilibrium surface (b)

参　考　文　献

[1] Qiao C, Guo Y H, Li C H. Study on rock burst prediction of deep buried tunnel based on cusp catastrophe theory. *Geotech Geol Eng*, 2021, 39(2): 1101

[2] Zhi Y J, Yang T, Fu D M. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. *J Mater Sci Technol*, 2020, 49: 202

[3] Pei J K, Wang F Y, Guo H H, et al. Cause analysis of chemical accidents based on improved cusp catastrophe model. *China Saf Sci J*, 2019, 29(7): 20
(裴甲坤, 王飞跃, 郭焕焕, 等. 基于改进尖点突变模型的化工事故致因分析. 中国安全科学学报, 2019, 29(7): 20)

[4] Lin L. Stochastic cusp catastrophe model for Chinese stock market. *J Syst Eng*, 2016, 31(1): 55
(林黎. 中国股票市场的随机尖点突变模型. 系统工程学报, 2016, 31(1): 55)

[5] Barunik J, Kukacka J. Realizing stock market crashes: Stochastic cusp catastrophe model of returns under time-varying volatility. *Quant Finance*, 2015, 15(6): 959

[6] Ma Y R, Yi D, Hu B. Analysis of stochastic catastrophe mechanism of occupational well-being of nursing practitioner servicing for the elderly. *J Syst Manag*, 2021, 30(3): 526
(马跃如, 易丹, 胡斌. 养老护理员工作幸福感的随机突变机理. 系统管理学报, 2021, 30(3): 526)

[7] Eladany M M, Eldesouky A A, Sallam A A. Power system transient stability: An algorithm for assessment and enhancement based on catastrophe theory and FACTS devices. *IEEE Access*, 2018, 6: 26424

[8] Xiao X P, Duan H M. A new grey model for traffic flow mechanics. *Eng Appl Artif Intell*, 2020, 88: 103350

[9] Wei X, Fu D M, Chen M D, et al. Data mining to effect of key alloying elements on corrosion resistance of low alloy steels in Sanya seawater environmentAlloying Elements. *J Mater Sci Technol*, 2021, 64: 222

[10] Pei Z B, Zhang D W, Zhi Y J, et al. Towards understanding and prediction of atmospheric corrosion of an Fe/Cu corrosion sensor via machine learning. *Corros Sci*, 2020, 170: 108697

[11] Thom R. *Structural Stability and Morphogenesis*: *An Outline of a General Theory of Models*. London: Benjamin W A, 1975

[12] Cobb L. Stochastic catastrophe models and multimodal distributions. *Syst Res*, 1978, 23(4): 360

[13] Cobb L, Zacks S. Applications of catastrophe theory for statistical modeling in the biosciences. *J Am Stat Assoc*, 1985, 80(392): 793

[14] Zeeman E C. Catastrophe theory. *Sci Am*, 1976, 234(4): 65

[15] Niu D X, Wang K K, Sun L J, et al. Short-term photovoltaic power generation forecasting based on random forest feature selection and CEEMD: A case study. *Appl Soft Comput*, 2020, 93: 106389

[16] Al-Fugara A, Ahmadlou M, Shatnawi R, et al. Novel hybrid models combining meta-heuristic algorithms with support vector regression (SVR) for groundwater potential mapping. *Geocarto Int*, 2020: 1

[17] Zhang J L, da Xu, Hao K J, et al. FS-GBDT: Identification multicancer-risk module via a feature selection algorithm by integrating Fisher score and GBDT. *Brief Bioinform*, 2020, 22(3): bbaa189

[18] Tsai C F, Sung Y T. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. *Knowl Based Syst*, 2020, 203: 106097

[19] Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends. *Inf Fusion*, 2019, 52: 1

[20] Pes B. Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. *Neural Comput Appl*, 2020, 32(10): 5951

[21] Hartelman P A I, Maas H L J, Molenaar P C M. Detecting and modelling developmental transitions. *Br J Dev Psychol*, 1998, 16(1): 97

[22] Grasman R P P P, van der Maas H L J, Wagenmakers E J. Fitting the cusp catastrophe inR: AcuspPackage primer. *J Stat Soft*, 2009, 32(8): 1

[23] Aaron F, Cynthia R, Francesca D. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Machine Learning*

Research, 2019, 20(177): 1

[24] Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5

[25] Buscema M. Back propagation neural networks. *Subst Use Misuse*, 1998, 33(2): 233

[26] Karatzoglou A, Smola A, Hornik K, et al. Kernlab- AnS4Package for kernel methods inR. *J Stat Soft*, 2004, 11(9): 1

[27] Amar Aladžuz. Hotels accommodation prices dataset [DB/OL]. *Kaggle* (2020-12-24) [2021-07-16].https://www.kaggle.com/ala-dzuzamar/hotels-accommodation-prices-dataset

[28] Biecek P. DALEX: explainers for complex predictive models in R. *J Mach Learn Res*, 2018, 19(1): 3245

[29] Pei Z B, Cheng X Q, Yang X J, et al. Understanding environmental impacts on initial atmospheric corrosion based on corrosion monitoring sensors. *J Mater Sci Technol*, 2021, 64: 214