doi:10.11887/j.cn.202102004 http://journal.nudt.edu.cn

基于随机森林的重要性测度指标体系*

宋述芳,何入洋
(西北工业大学 航空学院,陕西 西安 710072)

摘 要:重要性测度分析可以找出重要特征变量,从而降低输入空间的维数,节约运算成本。基于随机森林重要性测度的分析原理,探寻随机森林的重要性测度指标与基于方差的全局灵敏度指标之间的联系,得到求解方差灵敏度主指标 S_i 及其总指标 S_i^T 的新途径。建立基于随机森林的单变量、组变量重要性测度指标,并明确具体的求解过程,完善基于随机森林的重要性测度指标体系。通过算例验证了所提基于随机森林的重要性测度指标体系的有效性及其与方差灵敏度指标之间关系的正确性。

关键词:随机森林;重要性测度;全局灵敏度;组变量;降维
中图分类号:V221;TB114.3 **文献标志码**:A **文章编号**:1001-2486(2021)02-025-08

Importance measure index system based on random forest

SONG Shufang, *HE Ruyang*
(School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The importance measure analysis can find out the important feature variables of model, which can effectively reduce the variable dimension and decrease the computation time. The relationship between the important measure of random forest and the variance-based global sensitivity measure was explored, which can give a novel way to solve variance-based main sensitivity index S_i and total sensitivity index S_i^T. The importance measure of single and group variables based on random forest were established to improve the corresponding measure index system. Several examples are given to verify the validity of the proposed important measures and the correctness relation derivation about variance-based sensitivity indices.

Keywords: random forest; importance measure; global sensitivity; group variables; dimension-reduction

灵敏度分析可以反映输入变量对输出响应的影响程度,主要分为局部灵敏度分析与全局灵敏度分析[1]。局部灵敏度反映的是名义值处输入变量对输出响应的影响,它受制于名义值的选取,缺乏全局性与计算的稳定性。全局灵敏度(又称重要性测度)能够从平均的角度衡量输入变量在整个分布区域内对输出响应的影响[2-5]。目前,重要性测度模型主要分为三类:非参数化模型、基于方差的模型和矩独立模型。基于方差的重要性测度分析应用最为广泛,可以用单层/双层 Monte Carlo 模拟法求解方差灵敏度主指标 S_i 及其总指标 S_i^T[6]。

随机森林(Random Forest, RF)是 Breiman 于2001年提出的一种统计学习理论方法[7]。首先,通过 Bootstrap 重采样技术从原始样本集中抽取多个训练样本集,然后再利用抽取的样本集建立相应的决策树,并组建随机森林。随机森林应用广泛,不仅可以处理分类、回归问题,对于降维也有很好的适用性。随机森林对异常值与噪音也有很好的容忍度,稳健性强,不容易出现过拟合,被 Iverson 誉为当前最好的算法之一[8]。现有的基于随机森林的重要测度指标主要有两种:基于 Gini 指数的平均不纯度减少指标(Mean Decrease Impurity, MDI)和基于袋外(Out-Of-Bag, OOB)数据置换的平均精确率减少指标(Mean Decrease Accuracy, MDA)[9-11]。基于 Gini 指数的 MDI 指标对离散特征存在偏向性,且重要性分析结果与特征变量的选择顺序有关[12-13]。基于 OOB 数据置换的 MDA 指标则可以直接度量每个特征变量对模型精确率的影响程度,不存在偏向问题,应用广泛。此外,基于 OOB 数据置换的 MDA 指标求解过程与基于方差的全局灵敏度分析的单层 Monte Carlo 模拟法相似,可由此作为切入点寻找两者之间的关系。

* **收稿日期**:2019-09-16
基金项目:国家数值风洞工程资助项目(NNW 2019ZT2-A05);国家自然科学基金资助项目(11902254)
作者简介:宋述芳(1982—),女,副教授,博士,硕士生导师,E-mail:shufangsong@nwpu.edu.cn

本文通过比较基于方差的全局灵敏度指标和基于 OOB 数据置换的 MDA 指标的求解过程，寻找两者之间的关系，并进一步建立基于随机森林的重要测度指标体系，包括单变量测度指标、组变量测度指标等，可为后期复杂环境、高维小样本数据的重要性测度分析奠定基础。

1 基于方差的全局灵敏度指标

由 Sobol′提出的基于方差的全局灵敏度指标能够反映输入变量在整个变化范围内对输出响应方差的影响程度。Sobol′指标不仅具有很强的模型通用性，而且还可对输入变量进行分组讨论以及量化输入变量之间的交互影响，因此在工程领域得到了广泛应用。ANOVA（analysis of variance）分解是方差灵敏度指标分析的基础[6]。

1.1 ANOVA 分解

响应函数 $Y=g(\boldsymbol{X})$ 存在唯一的 ANOVA 分解式为

$$g(\boldsymbol{X}) = g_0 + \sum_{i=1}^{n} g_i(X_i) + \sum_{1\leqslant i<j\leqslant n} g_{ij}(X_i,X_j) + \cdots + g_{1,\cdots,n}(X_1,X_2,\cdots,X_n) \tag{1}$$

其中，常量 g_0 为函数 $g(\boldsymbol{X})$ 的期望值，$g_i(X_i)$ 为单变量 X_i 的主效应分量。

$$g_i(X_i) = \int g(\boldsymbol{X}) \prod_{j\neq i}^{n} [f_{X_j}(x_j)\mathrm{d}x_j] - g_0 \tag{2}$$

式中，$f_{X_i}(x_i)$ 为变量 X_i 的概率密度函数。

多个变量交互作用的分量可由下式求得

$$g_{i,\cdots,s}(X_i,\cdots,X_s) = \int g(\boldsymbol{X}) \prod_{k\neq i,\cdots,s}^{n} [f_{X_k}(x_k)\mathrm{d}x_k] - \sum_{l=i}^{s-1} \sum_{k_i,\cdots,k_l\in(i,\cdots,s)} g_{k_i,\cdots,k_l}(X_{k_i},\cdots,X_{k_l}) - g_0 \tag{3}$$

1.2 方差灵敏度指标

基于式(1)，分别对各分解项进行积分，由于各分解项正交，响应函数的方差 $V = VAR(Y)$ 可以表示为各分解项的方差之和，即

$$V = \sum_{i=1}^{n} V_i + \sum_{1\leqslant i<j\leqslant n}^{n} V_{ij} + \cdots + V_{1,2,\cdots,n} \tag{4}$$

其中，

$$V_{1,\cdots,s} = \int g_{1,\cdots,s}^2(X_1,\cdots,X_s) \prod_{k=1,\cdots,s} [f_{X_k}(x_k)\mathrm{d}x_k] \tag{5}$$

用分解项的方差与响应函数的方差之比来衡量分解项的方差贡献率，即

$$S_{i_1,\cdots,i_s} = V_{i_1,\cdots,i_s}/V \tag{6}$$

其中，S_i 表示单变量 X_i 的灵敏度主指标。

将所有与变量 X_i 相关的影响效应求和，得到变量 X_i 的灵敏度总指标 S_i^{T}，即

$$S_i^{\mathrm{T}} = S_i + \sum_{j>i} S_{ij} + \sum_{k>j>i} S_{ijk} + \cdots + S_{12\cdots n} \tag{7}$$

由概率论知识可知，基于方差的全局灵敏度指标可表示为[14-16]

$$S_i = \frac{V_i}{V} = \frac{VAR[E(Y|X_i)]}{VAR(Y)} \tag{8}$$

$$S_i^{\mathrm{T}} = \frac{V_i^{\mathrm{T}}}{V} = 1 - \frac{VAR[E(Y|\boldsymbol{X}_{\sim i})]}{VAR(Y)} \tag{9}$$

其中，$\boldsymbol{X}_{\sim i}$ 表示除 X_i 外的所有变量组成的向量。

1.3 求解方差灵敏度指标的 Monte Carlo 法

采用传统的数字模拟法求解基于方差的全局灵敏度指标需要进行双层抽样，计算量大，不适用于复杂的工程问题分析[17]。单层 Monte Carlo 模拟法应用广泛，其求解步骤如下。

Step 1：根据输入变量 $\boldsymbol{X}$ 的联合分布，抽取两组容量为 N 的样本，分别记为矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$：

$$\boldsymbol{A} = \begin{bmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{n1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1N} & \cdots & x_{iN} & \cdots & x_{nN} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} x_{1(N+1)} & \cdots & x_{i(N+1)} & \cdots & x_{n(N+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1(N+N)} & \cdots & x_{i(N+N)} & \cdots & x_{n(N+N)} \end{bmatrix}$$

Step 2：将矩阵 $\boldsymbol{B}$ 中的第 i 列用 $\boldsymbol{A}$ 中的第 i 列代替，构造矩阵 $\boldsymbol{C}_i$：

$$\boldsymbol{C}_i = \begin{bmatrix} x_{1(N+1)} & \cdots & x_{i1} & \cdots & x_{n(N+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1(N+N)} & \cdots & x_{iN} & \cdots & x_{n(N+N)} \end{bmatrix}$$

Step 3：计算输入变量 X_i 的方差灵敏度主指标和总指标，即

$$S_i = \frac{\frac{1}{N}\sum_{j=1}^{N} y_A^{(j)} y_{C_i}^{(j)} - g_0^2}{VAR(Y)} \tag{10}$$

$$S_i^{\mathrm{T}} = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} y_B^{(j)} y_{C_i}^{(j)} - g_0^2}{VAR(Y)} \tag{11}$$

式中，$y_A^{(j)}$、$y_B^{(j)}$、$y_{C_i}^{(j)}$ 分别是矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}_i$ 为输入时的输出响应向量 $\boldsymbol{y}_A = [y_A^{(1)},\cdots,y_A^{(N)}]^{\mathrm{T}}$、$\boldsymbol{y}_B = [y_B^{(1)},\cdots,y_B^{(N)}]^{\mathrm{T}}$、$\boldsymbol{y}_{C_i} = [y_{C_i}^{(1)},\cdots,y_{C_i}^{(N)}]^{\mathrm{T}}$ 的第 j 个响应值。

2 基于随机森林的重要性测度分析

随机森林是一种统计学习理论方法，利用 Bootstrap 重采样方法从数据库中抽取样本，并运用决策树对每组 Bootstrap 样本进行建模，组合多棵决策树，通过投票（分类）或取平均值（回归）得

出最终的预测结果[7]。随机森林具有很高的预测精度,鲁棒性好,防止过拟合,在分类、回归、降维等问题中得到了广泛应用。基于随机森林的重要性测度指标有:基于 Gini 指数的 MDI 指标和基于 OOB 数据置换的 MDA 指标。基于 OOB 数据置换的 MDA 指标可直接度量每个特征变量对模型精确度的影响程度,不存在 MDI 指标的偏向问题,使用范围广泛[9]。

基于 OOB 数据置换的 MDA 指标的主要思路:保证其他特征变量不变,只打乱 OOB 数据中的某个特征变量的顺序,破坏 OOB 数据的特征变量与输出之间的对应关系。利用决策树分别对打乱前与打乱后的 OOB 数据进行预测,将所有决策树前后两次预测的均方误差的平均值作为此特征变量的重要性测度结果[18]。基于 OOB 数据置换的 MDA 指标的求解过程如下。

Step 1:随机森林包含 M 棵决策树 $\boldsymbol{H}=\{h_1,h_2,\cdots,h_M\}$。分别利用每棵决策树 $h_m(m=1,\cdots,M)$ 对相应的 OOB 数据(OOB 数据的输入矩阵为 $\boldsymbol{x}_{\text{OOB}}$,输出响应向量为 $\boldsymbol{Y}$)的输入矩阵进行预测,预测结果为 $\boldsymbol{Y}_m$,则预测值 $\boldsymbol{Y}_m$ 与真实值 $\boldsymbol{Y}$ 的均方误差 $\varepsilon_m=\text{mean}(\boldsymbol{Y}_m-\boldsymbol{Y})^2$。

Step 2:保证 OOB 数据的其他特征变量不变,只打乱 X_i 的特征值顺序(即 $\boldsymbol{x}_{\text{OOB}}$的第 i 列),再利用决策树 h_m 对打乱顺序后的样本进行预测,则预测值 $\boldsymbol{Y}_m^i$ 与真实值 $\boldsymbol{Y}$ 的均方误差 $\varepsilon_m^i=\text{mean}(\boldsymbol{Y}_m^i-\boldsymbol{Y})^2$。

Step 3:特征变量 X_i 对决策树 h_m 预测精度的影响为 $mse_m^i=\varepsilon_m^i-\varepsilon_m$。

Step 4:重复 Step 1 ~ Step 3,遍历整个随机森林模型,得到特征变量 X_i 对所有决策树的影响 $(mse_1^i,mse_2^i,\cdots,mse_M^i)$,则特征变量 X_i 对随机森林准确率的总影响为

$$\eta_{\text{RF}}^i=\frac{1}{M}\sum_{m=1}^{M}mse_m^i \tag{12}$$

重要性测度 η_{RF}^i可能会出现三种情况:正值、负值和零[18]。当输入与输出之间有很强的关联性,在打乱 OOB 数据中特征变量的顺序后,关联性被破坏,则测度 η_{RF}^i为正;如果特征变量与输出响应无关,无论如何打乱顺序,其预测结果都不变,此时测度 η_{RF}^i应为零;如果打乱 OOB 数据中特征变量的顺序反而使得特征变量与输出的关联性加强,此时测度 η_{RF}^i应为负。

3　基于随机森林的重要性测度与方差全局灵敏度指标的关系

将高精度的 Kriging 模型作为随机森林中决策树的叶节点输出,替代原来的取平均值或线性拟合,以提高随机森林输出决策的精准度。如果忽略随机森林的预测误差,即第 2 节中 ε_m 为 0,则 $mse_m^i=\varepsilon_m^i$。仅考虑单棵决策树的 OOB 数据置换的均方误差,定义均方误差 ε_m^i 和 $\varepsilon_m^{\sim i}$ 用以表征基于随机森林的重要测度,并探寻 ε_m^i、$\varepsilon_m^{\sim i}$ 与 S_i、S_i^{T} 的关系。

3.1　均方误差 ε_m^i 与灵敏度总指标 S_i^{T} 的关系

对特征变量 X_i 的 MDA 指标中的均方误差 ε_m^i 进行分析,则

$$\begin{aligned}\varepsilon_m^i&=\text{mean}(\boldsymbol{Y}_m^i-\boldsymbol{Y})^2=\frac{1}{N}\sum_{j=1}^{N}(y_m^{(j)}-y^{(j)})^2\\&=\frac{1}{N}\sum_{j=1}^{N}[(y_m^{(j)})^2+(y^{(j)})^2-2y^{(j)}y_m^{(j)}]\\&=\frac{1}{N}\sum_{j=1}^{N}(y_m^{(j)})^2+\frac{1}{N}\sum_{j=1}^{N}(y^{(j)})^2-\frac{2}{N}\sum_{j=1}^{N}y^{(j)}y_m^{(j)}\end{aligned} \tag{13}$$

当 OOB 数据的样本量 N 充分大时,存在 $\frac{1}{N}\sum_{j=1}^{N}(y_m^{(j)})^2=\frac{1}{N}\sum_{j=1}^{N}(y^{(j)})^2$,表征决策树的叶节点 Kriging 函数的二阶原点矩。

单层 Monte Carlo 模拟法求解变量的灵敏度总指标 S_i^{T} 为

$$\begin{aligned}S_i^{\text{T}}&=1-\frac{\frac{1}{N}\sum_{j=1}^{N}y_B^{(j)}y_{C_i}^{(j)}-g_0^2}{VAR(Y)}\\&=\frac{\frac{1}{N}\sum_{j=1}^{N}[y_B^{(j)}]^2\cdot\frac{1}{N}\sum_{j=1}^{N}y_B^{(j)}y_{C_i}^{(j)}}{VAR(Y)}\end{aligned} \tag{14}$$

对比式(13)和式(14)可以得出

$$\varepsilon_m^i=2S_i^{\text{T}}\times VAR(Y) \tag{15}$$

3.2　均方误差 $\varepsilon_m^{\sim i}$ 与灵敏度主指标 S_i 的关系

在对 OOB 数据进行打乱时,如果保持特征变量 X_i 的顺序不变,将其他特征变量的顺序打乱,决策树进行预测得出预测值为 $\boldsymbol{Y}_m^{\sim i}$,其余步骤不变,定义指标 $\varepsilon_m^{\sim i}=\text{mean}(\boldsymbol{Y}_m^{\sim i}-\boldsymbol{Y})^2$。

$$\begin{aligned}\varepsilon_m^{\sim i}&=\text{mean}(\boldsymbol{Y}_m^{\sim i}-\boldsymbol{Y})^2=\frac{1}{N}\sum_{j=1}^{N}(y_m^{\sim i(j)}-y^{(j)})^2\\&=\frac{1}{N}\sum_{j=1}^{N}[(y_m^{\sim i(j)})^2+(y^{(j)})^2-2y^{(j)}y_m^{\sim i(j)}]\\&=\frac{1}{N}\sum_{j=1}^{N}(y_m^{\sim i(j)})^2+\frac{1}{N}\sum_{j=1}^{N}(y^{(j)})^2-\frac{2}{N}\sum_{j=1}^{N}y^{(j)}y_m^{\sim i(j)}\end{aligned} \tag{16}$$

当 OOB 数据的样本量 N 充分大时,存在 $\frac{1}{N}\sum_{j=1}^{N}(y_m^{\sim i(j)})^2=\frac{1}{N}\sum_{j=1}^{N}(y^{(j)})^2$。

将单层 Monte Carlo 模拟法求解变量的灵敏度主指标 S_i 的公式进行拆分,先减 1 再加 1,整理得出

$$S_i = \frac{\frac{1}{N}\sum_{j=1}^{N} y_A^{(j)} y_{C_i}^{(j)} - g_0^2}{VAR(Y)} - 1 + 1 = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} (y_A^{(j)})^2 - \frac{1}{N}\sum_{j=1}^{N} y_A^{(j)} y_{C_i}^{(j)}}{VAR(Y)} \tag{17}$$

对比式(16)和式(17)可以得出

$$\varepsilon_m^{\sim i} = 2 \times (1 - S_i) \times VAR(Y) \tag{18}$$

由此,建立了基于随机森林的重要性测度与方差灵敏度指标之间的联系。ε_m^i 表示特征变量 X_i 对输出响应的总影响,由 $\varepsilon_m^{\sim i}$ 可推导出特征变量的灵敏度主指标 S_i,从而得到了基于随机森林求解 S_i 与 S_i^{T} 的新途径,它能够继承随机森林的高效性与稳定性。对于线性问题,即 $S_i = S_i^{\mathrm{T}}$ 时,由式(15)与式(18)可知,$\varepsilon_m^i + \varepsilon_m^{\sim i} = 2 \times VAR(Y)$。

4 基于随机森林的组变量重要性测度

随机森林在进行特征变量重要性分析时,仅给出了单变量 X_i 的均方误差 ε_m^i,它与灵敏度总指标 S_i^{T} 存在线性变换关系,能够反映特征变量本身以及与其他变量相互作用下对模型精确率的影响程度。此外,定义的 $\varepsilon_m^{\sim i}$ 可以反映特征变量自身对模型精确度的影响。

在此基础上,提出组变量测度指标 $\varepsilon_m^{\sim ij}$ 以表征变量 X_i 与 X_j 共同作用对模型精确度的影响,并推导与基于方差的二阶灵敏度指标 S_{ij} 的关系。在对 OOB 数据打乱过程中,保持特征变量 X_i 与 X_j 不变,将其他特征变量的顺序打乱,决策树进行预测得出预测值 $\boldsymbol{Y}_m^{\sim ij}$,其余步骤不变,定义指标 $\varepsilon_m^{\sim ij} = \mathrm{mean}(\boldsymbol{Y}_m^{\sim ij} - \boldsymbol{Y})^2$。

由 $\varepsilon_m^{\sim ij}$ 可得组变量的灵敏度主指标 $S_{[i,j]}$ 为

$$S_{[i,j]} = 1 - \frac{\varepsilon_m^{\sim ij}}{2 \times VAR(Y)} \tag{19}$$

在单层 Monte Carlo 模拟法中,矩阵 $\boldsymbol{B}$ 中的第 i、j 列被矩阵 $\boldsymbol{A}$ 中的第 i、j 列代替后可求得组变量的主指标 $S_{[i,j]}$,$S_{[i,j]}$ 与单一变量的主指标 S_i 与 S_j 以及两变量交互指标 S_{ij} 的关系为[1]

$$S_{[i,j]} = S_i + S_j + S_{ij} \tag{20}$$

综合式(18)~(20),可由 $\varepsilon_m^{\sim ij}$ 推得两变量交互作用的灵敏度指标 S_{ij} 为

$$S_{ij} = \frac{\varepsilon_m^{\sim i} + \varepsilon_m^{\sim j} - \varepsilon_m^{\sim ij}}{2 \times VAR(Y)} - 1 \tag{21}$$

其中,上标“$\sim i$”“$\sim j$”“$\sim ij$”分别表示带外数据中除第 i 列、第 j 列以及第 i 和第 j 列以外的数据打乱顺序带来的预测精度的影响。

对随机森林中每棵决策树的均方误差取平均,可得 $\eta_i^{\mathrm{T}} = \mathrm{mean}(\varepsilon_m^i)$,$\eta_i = \mathrm{mean}(\varepsilon_m^{\sim i})$,$\eta_{ij} = \mathrm{mean}(\varepsilon_m^{\sim ij})$。至此,随机森林单变量 X_i 影响的重要性测度指标 η_i,单变量 X_i 自身以及与其他变量交互作用的重要性测度指标 η_i^{T},以及组变量的重要性测度指标 η_{ij},共同构成了基于随机森林的重要性测度体系。

5 算例与分析

算例 1: 线性函数

$$Y = X_1 + X_2 + X_3$$

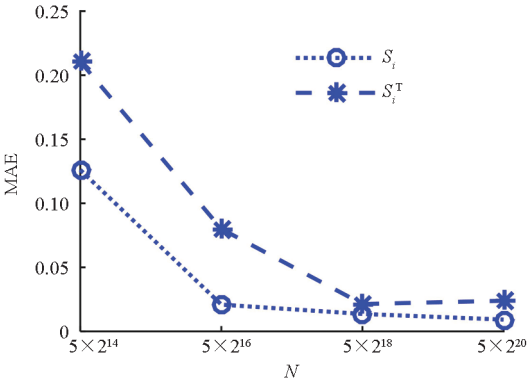
其中,$X_i(i = 1,2,3)$ 相互独立,均服从[0, 1]区间的均匀分布,解析可求得线性函数的方差 $VAR(Y) = 1/4$,基于方差的全局灵敏度指标 $S_i = S_i^{\mathrm{T}} = 1/3$,$S_{ij} = 0$。采用随机森林对线性函数进行重要性测度分析的结果见表 1。

表 1 线性函数的变量重要性测度分析结果

Tab. 1 The variable importance measures for linear function

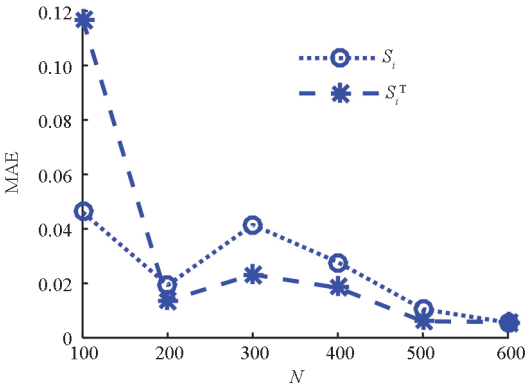
变量	η_i	$\eta_i \Rightarrow S_i$	误差/%	变量	η_i^{T}	$\eta_i^{\mathrm{T}} \Rightarrow S_i^{\mathrm{T}}$	误差/%	变量	η_{ij}	$\eta_{ij} \Rightarrow S_{ij}$
X_1	0.334 0	0.333 5	0.06	X_1	0.165 7	0.330 7	0.78	X_1X_2	0.165 5	0.000 2
X_2	0.332 5	0.336 4	0.93	X_2	0.166 2	0.331 7	0.48	X_1X_3	0.167 0	0.002 3
X_3	0.332 9	0.335 6	0.69	X_3	0.166 2	0.331 6	0.51	X_2X_3	0.166 4	0.004 0

线性函数 $S_i = S_i^{\mathrm{T}}$ 且 $S_{ij} = 0$,故有 $\eta_i + \eta_i^{\mathrm{T}} = 2 \times VAR(\boldsymbol{Y}) = 1/2$。图 1 给出了单层准 Monte Carlo 模拟法(Quasi-Monte Carlo, QMC)计算方差灵敏度指标、随机森林进行重要性测度推得方差灵敏度的平均绝对误差(Mean Absolute Error, MAE)随样本量的变化曲线。当样本总量为 600 时(400 个训练样本,200 个 OOB 数据),随机森林便可获得误差小于 1% 的测度指标,而单层 QMC 模拟的方差灵敏度分析则需要 5×2^{20} 个样本才能保证 S_i 与 S_i^{T} 的精度。



(a) 基于单层 QMC 模拟的方差灵敏度分析

(a) Single-loop QMC simulation for variance-based sensitivity analysis

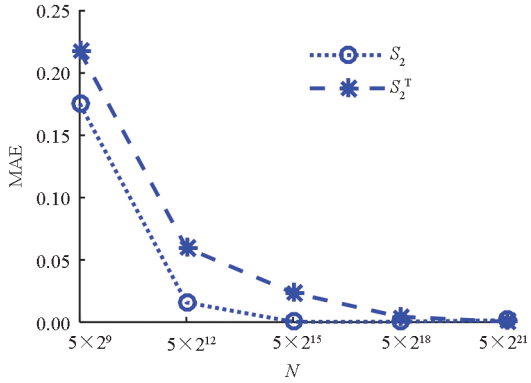


(b) 基于随机森林的重要测度分析

(b) Random forest for importance measure analysis

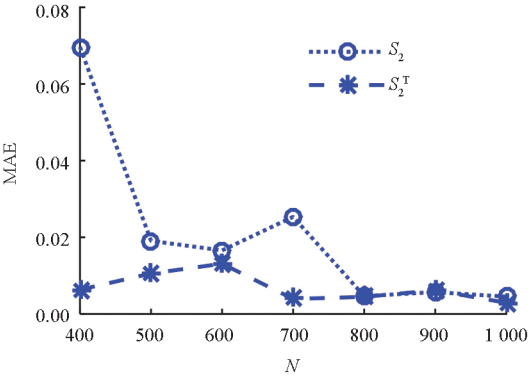
图 1　线性函数的方差灵敏度误差随样本量的变化曲线

Fig. 1　Error of variance-based sensitivity indices versus sampling number for linear function



(a) 基于单层 QMC 模拟的方差灵敏度分析

(a) Single-loop QMC simulation for variance-based sensitivity analysis



(b) 基于随机森林的重要测度分析

(b) Random forest for importance measure analysis

图 2　Ishigami 函数的方差灵敏度误差随样本量的变化曲线

Fig. 2　Error of variance-based sensitivity indices versus sampling number for Ishigami function

算例 2: Ishigami 函数[19]

$$Y=\sin(X_1)+7\sin^2(X_2)+0.1X_3^4\sin(X_1)$$

其中,$X_i(i=1,2,3)$相互独立,均服从$[-\pi,\pi]$区间的均匀分布。函数的方差 $VAR(Y)\approx13.846\ 0$。采用随机森林对 Ishigami 函数进行重要性测度分析,以变量 X_2 为例,基于单层 QMC 模拟的方差灵敏度指标、随机森林进行重要性测度推得方差灵敏度的误差随样本量的变化曲线如图 2 所示。随机森林用 300 个训练样本、700 个 OOB 样本进行重要性分析,可获得误差小于 2% 的测度指标,分析结果列于表 2。

Ishigami 函数中变量 X_3 的灵敏度主指标$S_3=0$,但是其灵敏度总指标 $S_3^T\approx0.243\ 7$,表 2 也给出了变量相互作用的方差灵敏度的二阶指标 S_{ij},可以得到变量 X_3 与变量 X_1 的交互重要性较大,因此导致了变量 X_3 的灵敏度总指标较大。

表 2　Ishigami 函数的变量重要性测度分析结果

Tab. 2　Variable importance measures for Ishigami function

变量	S_i	η_i	$\eta_i\Rightarrow S_i$	误差/%	变量	S_i^T	η_i^T	$\eta_i^T\Rightarrow S_i^T$	误差/%	变量	S_{ij}	η_{ij}	$\eta_{ij}\Rightarrow S_{ij}$	误差/%
X_1	0.313 9	18.996 5	0.314 0	0.03	X_1	0.557 6	15.359 1	0.554 6	0.53	X_1X_2	0.000	6.697 7	−0.002 8	
X_2	0.442 4	15.316 1	0.446 9	1.02	X_2	0.442 4	12.331 4	0.445 3	0.66	X_1X_3	0.243 7	12.412 8	0.241 1	1.06
X_3	0.000 0	27.784 1	−0.003 3		X_3	0.243 7	6.690 4	0.241 6	0.86	X_2X_3	0.000	15.363 7	0.001 6	

算例 3：系统失效树模型[20]

$$Y = X_1X_3X_5 + X_1X_3X_6 + X_1X_4X_5 + X_1X_4X_6 + X_2X_3X_4 + X_2X_3X_5 + X_2X_4X_5 + X_2X_5X_6 + X_2X_4X_7 + X_2X_6X_7 \quad (22)$$

式中，X_1、X_2 代表事件每年发生的次数，$X_3 \sim X_7$ 代表了基本事件的失效率，各变量相互独立，均服从对数正态分布，分布参数如表 3 所示。将大样本（$N = 9 \times 2^{21}$）下的单层 QMC 模拟的结果作为方差灵敏度的近似精确解，函数的方差 $VAR(Y) \approx 1.606\ 8 \times 10^{-8}$，与随机森林重要性测度分析结果对比见表 4。

算例 3 的变量维数 $n = 7$，需要较多的样本（3 000个训练样本，5 000 个 OOB 数据）来保证随机森林的精度。由表 4 的结果可以看出，基于随机森林的重要性测度推得的方差灵敏度与单层 QMC 模拟的近似精确解基本一致，变量的重要性排序相同，X_2、X_6、X_5 为重要变量。此外，对变量的交互作用也进行了重要性分析，得到最大的两个交互灵敏度指标为：$S_{25} \approx 0.021\ 9$，$S_{26} \approx 0.026\ 3$。

表 3　失效树模型的变量分布信息

Tab. 3　Distribution information of input variables in fault tree model

变量	名义值	误差因子
X_1	2	2
X_2	3	2
X_3	1×10^{-3}	2
X_4	2×10^{-3}	2
X_5	4×10^{-3}	2
X_6	5×10^{-3}	2
X_7	3×10^{-3}	2

表 4　失效树模型的变量重要性测度分析结果对比

Tab. 4　Variable importance measures for fault tree model

变量	$S_i^{(\mathrm{QMC})}$	η_i	$\eta_i \Rightarrow S_i$	误差/%	变量	$S_i^{\mathrm{T}(\mathrm{QMC})}$	η_i^{T}	$\eta_i^{\mathrm{T}} \Rightarrow S_i^{\mathrm{T}}$	误差/%
X_1	0.035 5	3.11×10^{-8}	0.032 1	4.75	X_1	0.042 1	1.38×10^{-9}	0.043 2	2.03
X_2	0.326 2	2.16×10^{-8}	0.327 1	0.61	X_2	0.395 3	1.27×10^{-8}	0.394 4	0.11
X_3	0.015 5	3.17×10^{-8}	0.012 6	7.43	X_3	0.018 0	5.98×10^{-9}	0.018 7	3.36
X_4	0.085 3	2.95×10^{-8}	0.082 9	2.67	X_4	0.098 4	3.21×10^{-9}	0.100 4	1.62
X_5	0.174 3	2.66×10^{-8}	0.170 7	1.27	X_5	0.210 9	6.82×10^{-9}	0.213 2	0.62
X_6	0.221 4	2.51×10^{-8}	0.217 6	1.10	X_6	0.264 2	8.54×10^{-9}	0.265 2	0.54
X_7	0.048 2	3.07×10^{-8}	0.046 3	4.95	X_7	0.061 4	2.05×10^{-9}	0.064 0	3.92

算例 4：屋架结构

某屋架结构如图 3 所示，屋架的上弦杆和压杆采用钢筋混凝土杆，下弦杆和拉杆采用钢杆。设屋架结构承受垂直的均布载荷 q 的作用，将均布载荷 q 化成节点载荷 P，则 $P = ql/4$，通过力学知识可得 C 点的垂直位移为

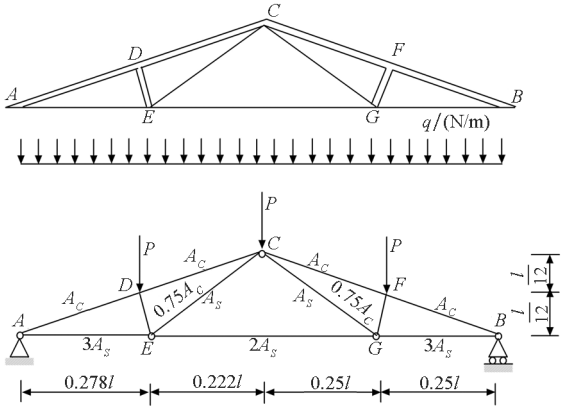


图 3　屋架结构的简单示意图

Fig. 3　Roof truss structure

$$\Delta_C = \frac{ql^2}{2}\left(\frac{3.81}{A_CE_C} + \frac{1.13}{A_SE_S}\right) \quad (23)$$

式中，A_C、A_S、E_C、E_S 分别为钢筋混凝土杆与钢杆的横截面积与弹性模量，l 为杆长，假设所有输入变量相互独立，且服从正态分布，分布参数如表 5 所示。

表 5　屋架结构的变量分布参数

Tab. 5　Distribution parameters of input variables in roof truss structure

变量	均值	标准差
q/(N/m)	20 000	1400
l/m	12	0.12
A_S/m^2	9.820×10^{-4}	5.982×10^{-5}
A_C/m^2	0.04	0.004 8
E_S/MPa	2×10^{11}	1.2×10^{10}
E_C/MPa	3×10^{10}	1.8×10^{9}

响应函数的方差 $VAR(\Delta_C) \approx 1.626\ 6 \times 10^{-6}$。

以大样本($N=8\times2^{20}$)下的单层 QMC 模拟结果作为近似精确解。随机森林用 1 000 个训练样本、5 000个 OOB 数据进行重要性测度分析。

表 6 列出了各特征变量的重要性测度结果,可以看出基于随机森林的重要性测度指标推得的 S_i、S_i^{T} 与单层 QMC 模拟的近似精确解吻合很好。S_i 与 S_i^{T} 的结果非常接近,说明该响应函数中特征变量的交互作用很小。从变量的重要性排序可以得到:均布载荷 q 的重要性最大,而钢筋混凝土杆的抗拉强度 E_C 与杆长 l 的重要性最小,因此在屋架结构的优化当中可以将 E_C 和 l 设置为常数,从而达到降低变量维数、简化模型的目的。

表 6　屋架结构的变量重要性测度分析结果对比

Tab. 6　Variable importance measures for the roof truss structure

变量	$S_i^{(\mathrm{QMC})}$	η_i	$\eta_i\Rightarrow S_i$	误差/%	变量	$S_i^{\mathrm{T(QMC)}}$	η_i^{T}	$\eta_i^{\mathrm{T}}\Rightarrow S_i^{\mathrm{T}}$	误差/%
q	0.448 5	1.78×10^{-6}	0.453 7	1.17	q	0.451 1	1.52×10^{-6}	0.467 8	3.71
l	0.036 6	3.13×10^{-6}	0.037 5	2.37	l	0.037 0	1.49×10^{-7}	0.045 8	2.37
A_S	0.143 1	2.80×10^{-6}	0.138 9	2.97	A_S	0.144 4	4.72×10^{-7}	0.145 0	0.38
A_C	0.185 8	2.68×10^{-6}	0.177 8	4.29	A_C	0.187 4	5.98×10^{-7}	0.183 6	2.01
E_S	0.138 8	2.83×10^{-6}	0.132 4	4.60	E_S	0.140 0	4.51×10^{-7}	0.138 4	1.13
E_C	0.043 2	3.12×10^{-6}	0.042 3	1.99	E_C	0.044 1	1.48×10^{-7}	0.045 6	3.36

6　结论

1)将决策树的叶节点由原始的取平均或线性拟合变为高精度的 Kriging 模型,使得改进后的决策树对原响应函数有更好的拟合精度。

2)在基于随机森林的 MDA 指标的分析基础上,提出了单变量和组变量重要性测度指标,完善了基于随机森林的重要性测度指标体系。

3)找到了基于随机森林的重要性测度指标与基于方差的全局灵敏度主指标、总指标之间的关系,可用随机森林的重要性测度指标推导出方差灵敏度指标,获得方差灵敏度指标求解的新途径。

4)本文只研究了独立变量对输出响应的影响,后续将开展基于随机森林的相关特征变量的重要性测度分析方面的研究。

参考文献(References)

[1] 吕震宙, 李璐祎, 宋述芳, 等. 不确定性结构系统的重要性分析理论与求解方法[M]. 北京: 科学出版社, 2015.
LYU Zhenzhou, LI Luyi, SONG Shufang, et al. Importance analysis theory and solution method with structural uncertainty[M]. Beijing: Science Press, 2015. (in Chinese)

[2] BORGONOVO E. A new uncertainty importance measure[J]. Reliability Engineering & System Safety, 2007, 92(6): 771 – 784.

[3] LIU Q, HOMMA T. A new computational method of a moment-independent uncertainty importance measure [J]. Reliability Engineering & System Safety, 2009, 94 (7): 1205 – 1211.

[4] CUI L J, LYU Z Z, ZHAO X P. Moment-independent importance measure of basic random variable and its probability density evolution solution [J]. Science China Technological Sciences, 2010, 53(4): 1138 – 1145.

[5] SALTELLI A. Sensitivity analysis for importance assessment[J]. Risk Analysis, 2002, 22(3): 579 – 590.

[6] SOBOL′ I M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates [J]. Mathematics and Computer in Simulation, 2001, 55 (1): 271 – 280.

[7] BREIMAN L. Random forests—random features [J]. Machine Learning, 2001, 45(1): 5 – 32.

[8] PRASAD A M, IVERSON L R, LIAW A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction [J]. Ecosystems, 2006, 9(2): 181 – 199.

[9] NICODEMUS K K. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures[J]. Briefings in Bioinformatics, 2011, 12 (4): 369 – 373.

[10] KELLIE A, KIMES R V. Empirical characterization of random forest variable importance measures [J]. Computational Statistics & Data Analysis, 2008, 52 (4): 2249 – 2260.

[11] MITCHELL M. Bias of the random forest out-of-bag (OOB) error for certain input parameters [J]. Open Journal of Statistics, 2011, 1(3): 205 – 211.

[12] STROBL C, BOULESTEIX A L, KNEIB T, et al. Conditional variable importance for random forests[J]. BMC Bioinformatics, 2008, 9(1): 307.

[13] NICODEMUS K K, MALLEY J D, STROBL C, et al. The behaviour of random forest permutation-based variable importance measures under predictor correlation [J]. BMC Bioinformatics, 2010, 11(1): 110.

[14] SALTELLI A, TARANTOLA S. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal[J]. Journal of the American Statistical Association, 2002, 97(459): 702 – 709.

[15] IMAN R L, HORA S C. A robust measure of uncertainty importance for use in fault tree system analysis[J]. Risk Analysis, 1990, 10(3): 401 – 406.

[16] HOMMA T, SALTELLI A. Importance measures in global sensitivity analysis on nonlinear models[J]. Reliability Engineering & System Safety, 1996, 52(1): 1 – 17.

[17] SALTELLI A. Sensitivity analysis for importance assessment[J]. Risk Analysis, 2002, 22(3): 579 – 590.

[18] CUTLER A, CUTLER D R, STEVENS J R. Ensemble machine learning[M]. US: Springer, 2011, 45(1): 157 – 176.

[19] SONG S F, WANG L. Modified GMDH-NN algorithm and its application for global sensitivity analysis[J]. Journal of Computational Physics, 2017, 348(1): 534 – 548.

[20] PARK C K, AHN K I. A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment[J]. Reliability Engineering & System Safety, 1994, 46(3): 253 – 261.