Environmental Modelling & Software 70 (2015) 178–190

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system

Pengfei Wei [a, *], Zhenzhou Lu [b], Jingwen Song [b]

[a] School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, 710072 Xi'an, People's Republic of China
[b] School of Aeronautics, Northwestern Polytechnical University, 710072 Xi'an, People's Republic of China

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form
25 April 2015
Accepted 27 April 2015
Available online 20 May 2015

Keywords:
High-dimensional model
Random forest
Permutation variables importance measure
Morris' screening design

ABSTRACT

Permutation variable importance measure (PVIM) based on random forest and Morris' screening design are two effective techniques for measuring the variable importance in high dimensions. The former technique is developed in the machine learning discipline and widely used in bioinformatics, while the latter technique is popular in scientific computing. We present three main contributions to variable importance analysis (VIA). First, through theoretical derivation, we show that the PVIM converges to double the non-standardized Sobol' total effect index. This observation indicates that the PVIM is especially useful for variable screening as it captures both the individual and interaction effects. Second, three numerical examples with different types of model behavior are presented for comparing the performances of these two techniques. The main conclusions are as follows. For high-dimensional additive or approximately additive models, the PVIM is much more efficient than Morris' screening design when used for both variable importance ranking and variable screening. For high-dimensional models mainly governed by interaction effects, the performance of PVIM degrades, but it is still a competitive technique. Finally, the two techniques are applied to an environmental multi-indicators system for improving the robustness of the partial order structure of this system.

1. Introduction

With the rapid development of computing power, more and more computational models are developed in many disciplines, such as environmental science and risk analysis, for understanding the behaviors of natural systems and supporting decisions. Meanwhile, in data sciences such as bioinformatics, the data volume is expanding dramatically. Variable importance analysis (VIA) based on computational models or training data has become a standard scheme in these disciplines (see for example Hall et al., 2009; Boulesteix et al., 2012; Butler et al., 2014; Della Peruta et al., 2014; Gan et al., 2014; Baroni and Tarantola, 2014; Saint-Geours et al., 2014).

Many techniques have been developed for measuring the importance of the input variables in computational models. The good practices include Sobol' indices (Sobol', 1993; Homma and Saltelli, 1996), moment-independent indices (Borgonovo, 2007; Pianosi and Wagener, 2015) and Morris' screening design (Morris, 1991; Campolongo et al., 2007). In the framework of Sobol' indices, the relative importance of each input variable is quantified by the contribution of this input variable to the model output variance. Considering that variance is not sufficient for characterizing uncertainty, the moment-independent indices were developed by Borgonovo (2007) for VIA. In Morris' screening design, a set of difference quotients (commonly called elementary effects, EEs) are firstly computed for each variable with well-designed trajectories, and then two importance measures (i.e., mu and sigma) are computed based on these EEs. Among these three techniques, the Sobol' indices have gained the most attentions as that they not only quantify the individual and total contributions of each input variable, but also reflect the structures of model response functions, and many rigorous numerical algorithms are available for computing these indices (see for example Tarantola et al., 2006; Ratto et al., 2009; Saltelli et al., 2010). However, for high-dimensional problems, the best practice is Morris' screening design (see subsection 6.5 of Saltelli et al., 2008).

* Corresponding author.
E-mail addresses: wpf0414@163.com, pengfeiwei@nwpu.edu.cn (P. Wei).

http://dx.doi.org/10.1016/j.envsoft.2015.04.015
1364-8152/

For high-dimensional data, good practices for VIA are linear regression based methods (Johnson and Lebreton, 2004; Bi, 2012) and random forest (RF) based methods (Breiman, 2001; Siroky, 2009). These two categories of methods have been compared with numerical simulations by Grömping (2009). The linear regression based methods are only suitable for linear or approximately linear models. The RF, developed by Breiman (2001), is a nonparametric machine learning algorithm. It can not only deal with the highly nonlinear models with extensive interactions, but can also be applied to problems with input dimension n much higher than the sample size N, thus it has been regarded as a standard method in many disciplines such as bioinformatics (see for example Nicodemus and Malley, 2009; Boulesteix et al., 2012). The RF (classification and regression) is developed for prediction, yet it can also be used for VIA. Along with the proposition of the RF, two variable importance measures (VIMs), called Gini VIM (GVIM) and permutation VIM (PVIM), were developed (Breiman, 2001). These two VIMs can be especially useful for extracting the small group of important variables from a large number (e.g., several thousands) of candidate inputs. Studies show that, when the model inputs are categorical variables with different number of candidate values, the GVIM is biased, and it tends to overestimate the importance of the categorical variables with more candidate values (Strobl et al., 2007), therefore it is rarely adopted in practical applications. The PVIM is biased only when the input variables are correlated (Strobl et al., 2008), and is more frequently used.

However, till now, no work has ever been presented for comparing the performance of the Morris' screening design and PVIM when applied to high-dimensional computational models. This motivates us to carry out this work.

Many studies (see for example Lunetta et al., 2004; Strobl et al., 2009; Winham et al., 2012) have empirically shown that PVIM captures both the individual and interaction effects of the input variables, but no theoretical proof has been presented before. The first aim of this work is to provide theoretical evidence for this observation by investigating the relation between the PVIM and Sobol' total effect indices.

For high-dimensional computational models, a common way to handle VIA is to firstly screen the large number of non-influential variables with screening techniques to substantially reduce the dimension of uncertain input variables, then further discriminate the individual effects of each single input and their interaction effects using other VIA techniques such as Sobol' indices (see for example Ge et al., 2015). Therefore, the second aim of this work is to compare the PVIM and Morris' screening design when applied to high-dimensional models with an emphasis on their abilities of screening non-influential variables. To address this, three numerical high-dimensional models with different types of behaviors are investigated, and several conclusions on the relative merits of both techniques are drawn.

In many disciplines such as environmental science, researchers face the problem of ranking multiple objects, each of which is characterized by a number of indicators. These systems are commonly termed as multi-indicators systems. Two popular techniques for ranking the objects are Copeland Score (Al-Sharrah, 2010) and Hasse Diagram (Brüggemann et al., 1995). In real applications, due to the present of epistemic uncertainties presented in the performance values of indicators, the researchers often find it difficult to produce robust rankings. The third aim of this work is to apply the PVIM and Morris' screening design to an environmental multi-indicators system to improve the robustness of the object ranking.

2. Materials and methods

2.1. Description of the variable importance analysis methods

Our aim is to compare the performance of PVIM and Morris' screening design when applied to high-dimensional computational models, thus it is necessary to briefly review these two techniques. We also review the Sobol' indices since they will be used for investigating the property of PVIM and test the effectiveness of the two concerned techniques.

2.1.1. Sobol' indices

Only the computational model with deterministic response function is considered. Let $Y = g(\mathbf{X})$ denote the model response function, where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is the n-dimensional vector of input variables and Y is the scale output variable. Throughout this paper we assume that the input variables are all independently and uniformly distributed in $[0, 1]$ for the purpose of simplicity, and the input space is denoted as $\mathbf{H}^n = [0, 1]^n$.

The Sobol' indices are derived from the functional analysis-of-variance (ANOVA) decomposition (Sobol', 1993; Homma and Saltelli, 1996). When the input variables are independent with each other, the model output variance $V(Y)$ can be decomposed into 2^{n-1} partial variance terms of increasing order:

$$V(Y) = \sum_{i=1}^{n} V_i + \sum_{i \neq j} V_{ij} + \ldots + V_{1,2,\ldots,n}, \tag{1}$$

where the first order partial variance $V_i = V[E(Y|X_i)]$ indicates the model output variance explained by X_i individually, thus quantifies the individual contribution of X_i; the second order partial variance $V_{ij} = V[E(Y|X_i, X_j)] - V_i - V_j$ measures the interaction effect between X_i and X_j; higher order partial variances indicates higher order interaction effects.

Based on the variance decomposition in Eq. (1), the Sobol' main effect index for X_i is defined as follows (Sobol', 1993):

$$S_i = \frac{V_i}{V(Y)}, \tag{2}$$

which is a standardization of the first order partial variance. The Sobol' total effect index for X_i is defined by (Homma and Saltelli, 1996):

$$\begin{aligned} S_{Ti} &= \frac{V_{Ti}}{V(Y)} = \frac{E[V(Y|\mathbf{X}_{\sim i})]}{V(Y)} = \frac{1}{V(Y)}\left(V_i + \sum_{j=1, j\neq i}^{n} V_{ij} + \ldots + V_{1,2,\ldots,n}\right) \\ &= S_i + \sum_{j=1, j\neq i}^{n} S_{ij} + \ldots + S_{1,2,\ldots,n}, \end{aligned} \tag{3}$$

where $\mathbf{X}_{\sim i}$ indicates the vector including all the input variables but X_i, $V_{Ti} = E[V(Y|\mathbf{X}_{\sim i})]$ is the total partial variance, $S_{ij} = V_{ij}/V(Y)$ indicates the normalization of the second order partial variance, which is commonly termed as Sobol' second order effect index. The Sobol' higher order effect indices are defined similarly. By definition, the total effect index S_{Ti} includes not only the individual contribution of X_i, but all the interaction contributions of X_i with the other variables. Therefore, if the total effect index of one input variable is equal to zero, this variable must make no contribution to the model output variance, and can be regarded as a non-influential variable. This property makes the total effect index especially useful for variable screening.

The Monte Carlo procedure is a commonly used algorithm for computing the Sobol' indices. The derived estimators often involve extensive number of function evaluations especially in the case of high dimension (Saltelli et al., 2010). However, as long as the amount of sample is sufficiently enough, these estimators will always converge to the true values. For more information on these Monte Carlo estimators, see Saltelli et al. (2010) and Appendix A.

2.1.2. Random forest and permutation variable importance measure

Let $\mathbf{D} = \{(\mathbf{x}^{(j)}, y^{(j)})\}_{j=1,2,\ldots,N}$ denote a set of N sample points. The classical RF consists of a set of classification or regression trees grown by the classification and regression tree (CART) algorithm (Breiman et al., 1984; Breiman, 2001), thus it is commonly denoted as CART-RF. There are also several other improved versions of RF such as the RF based on conditional inference tree (Hothorn et al., 2006), the RF based on reinforcement learning tree (Zhu et al., 2012) and the dynamic RF (Bernard et al., 2012). These improved RFs aim at improving the performance of RF in specific applications. In this presentation, only the classical regression RF is considered. Before discussing the RF, an introduction to CART algorithm is necessary.

Briefly, the CART algorithm grows the binary regression tree by recursively partitioning the training data space into more and more homogeneous rectangular parts with the principle of maximizing the decrease of node impurity at each splitting node. The root node contains all the training data, and for regression problem, its impurity is measured by the variance of model output samples contained in it. The splitting variable (say X_i) as well as the splitting criterion (say x_i^*) are specified with the principle of maximizing the reduction of node impurity. The root node as well as the sample space is then divided into two parts by attributing the training data with the value of X_i less than x_i^* to the left daughter node, and the remaining data contained in the root node into the right daughter node. The left and right daughter nodes as well as the subspaces continue to split in the same manner until some stopping criteria, for example, when the node impurity is less than a prespecified threshold value, are reached. The training data dropping into the same