Reliability Engineering and System Safety 142 (2015) 399–432

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Variable importance analysis: A comprehensive review

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 25 January 2014
Received in revised form
30 March 2015
Accepted 25 May 2015
Available online 9 June 2015

Keywords:
Variable importance analysis
Difference-based
Regression technique
Random forest
Variance-based
Moment-independent
Graphic variable importance measures

ABSTRACT

Measuring variable importance for computational models or measured data is an important task in many applications. It has drawn our attention that the variable importance analysis (VIA) techniques were developed independently in many disciplines. We are strongly aware of the necessity to aggregate all the good practices in each discipline, and compare the relative merits of each method, so as to instruct the practitioners to choose the optimal methods to meet different analysis purposes, and to guide current research on VIA. To this end, all the good practices, including seven groups of methods, i.e., the difference-based variable importance measures (VIMs), parametric regression and related VIMs, nonparametric regression techniques, hypothesis test techniques, variance-based VIMs, moment-independent VIMs and graphic VIMs, are reviewed and compared with a numerical test example set in two situations (independent and dependent cases). For ease of use, the recommendations are provided for different types of applications, and packages as well as software for implementing these VIA techniques are collected. Prospects for future study of VIA techniques are also proposed.

Contents

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

http://dx.doi.org/10.1016/j.ress.2015.05.018
0951-8320/

1. Introduction

Along with the rapid development of computer science and technique, a variety of computational models and numerical simulations have been developed for simulating and predicting the behavior of systems in nearly all fields of engineering and science such as aeronautical and astronautic engineering, chemistry and physics science, environmental science and technology, economics and education science. On the other hand, the last few decades have witnessed an explosive increase of the data volume in all kinds of large-scale scientific researches such as bioinformatics and related fields. To some degree, researchers from almost all the fields have reached an agreement on the necessity to perform variable importance analysis (VIA) based on these computational models and measured data. However, due to the wide dispersion of research fields and the lack of communication among different fields, the methodologies for VIA were independently developed in different research fields with different terminologies. These good practices in different disciplines, which will be reviewed in this article, are summarized in Fig. 1 with classification.

Researchers and practitioners working on computational models may face the problems of screening the relatively small group of important input variables from the tremendous candidate input variables (*variable prioritization setting*), fixing the large group of non-influential input variables at their nominal values without affecting the prediction accuracy or model output uncertainty (*variable fixing setting*), and determining how a reduction of the uncertainty of each input variable will influence the uncertainty in the output variable (*uncertainty reduction setting*) [1]. One can refer to Ref. [2] for an example of this type of analysis. VIA in these settings is mostly termed as "sensitivity analysis (SA)" in literature, where the word "sensitivity" used here is a general concept more related to "contribution" or "impact", not just the partial derivative which is commonly thought to be. This group of variable importance measures (VIMs) developed for computational models includes the difference-based VIMs, variance-based VIMs, moment-independent VIMs and the graphic VIMs, as shown in Fig. 1. This group of VIA techniques can also be termed as mathematical techniques.

In many disciplines such as bioinformatics, the objects operated by the analysts are measured data instead of computational models, and the analysts want to find the input variables that have obvious effect on the output variable based purely on data. This type of analysis is often dealt by statistical techniques such as measures of dependence, regression techniques and hypothesis tests. The correlation coefficient (CC), partial correlation coefficient (PCC), rank correlation coefficient (RCC), partial rank correlation coefficient (PRCC) and the moment-independent VIMs are all measures of dependence between the input and output variables. The parametric and nonparametric regression techniques aim at developing meta-model to approximate the true model response function. These techniques measure the variable importance either by the regression coefficients or by attributing the model output variance explained by the regression model to each of the input variables. The random forest, belonging to the group of nonparametric regression techniques, can provide the analysts with various types of VIMs, as indicated in Fig. 1. The hypothesis test techniques aim at testing the strength of relationship between the input and output variables, and use the probability-values (p-values) as measures of variable importance.

The reviews for "SA" methods developed for computational models are available in Refs. [3–12]. However, all these articles do not include the best practice for correlated input variables and the recently developed graphic VIMs. The reviews for statistical techniques (also called sampling-based techniques) are available