ELSEVIER

Reliability Engineering and System Safety 79 (2003) 123–129

RELIABILITY ENGINEERING & SYSTEM SAFETY

www.elsevier.com/locate/ress

Application of classification trees in the sensitivity analysis of probabilistic model results

Srikanta Mishra[a,*], Neil E. Deeds[a], Banda S. RamaRao[b]

[a]*INTERA Inc., 9111 Research Blvd. Austin, TX 78758, USA*
[b]*Fromatome ANP DE&S, 9111 Research Blvd. Austin, TX 78758, USA*

Abstract

The complexity of some integrated-system models necessitates using a probabilistic approach to quantify uncertainty in model projections. In this work, we demonstrate how classification trees can be used to perform sensitivity analyses on probabilistic results. The classification tree technique is applied to results from the probabilistic total system performance assessment model used in the Yucca Mountain project. The technique proves effective in delineating the variables that most influence low and high outcomes.

Keywords: Probabilistic models; Classification trees; Sensitivity analyses

1. Introduction and scope

Computer models are increasingly being used to predict the future behavior and associated uncertainties of complex systems such as nuclear waste repositories [7], oil production facilities [17] and global climate change dynamics [9]. Such integrated-system models, when executed in a probabilistic mode to enable quantification of uncertainties in model projections, often include hundreds of parameters that are uncertain and/or variable and whose interaction with one another can also be complex and/or highly nonlinear. It is difficult to obtain an understanding of exactly how the model works, and what the critical uncertainties and sensitivities are, from a simple evaluation of model results. To this end, sensitivity analysis provides a structured framework for unraveling the results of probabilistic model runs by examining the sensitivity of model results to the uncertainties and assumptions in model inputs.

Sensitivity analysis, in its simplest sense, involves quantification of the change in model output corresponding to a change in one or more of the model inputs. In the context of probabilistic models, however, sensitivity analysis takes on a more specific definition, viz. identification of those input parameters that have the greatest influence on the spread or variance of the model results. This is sometimes referred to as global sensitivity analysis [19] to distinguish it from the classical (local) sensitivity analysis measures typically obtained as partial derivatives of the output with respect to inputs of interest.

The contribution to output uncertainty (variance) by an input is a function of both the uncertainty of the input variable and the sensitivity of the output to that particular input. In general, input variables identified as important in global sensitivity analysis have both characteristics; they demonstrate significant variance and large sensitivity coefficients. Conversely, variables which do not show up as important per these metrics are either restricted to a small range in the probabilistic analysis, and/or are variables to which the model outcome does not have a high sensitivity.

Commonly used global sensitivity analysis techniques for probabilistic models include regression analysis [6], variance decomposition methods [20], screening methods [16] and partitioning techniques [5]. Several applications of these techniques have recently been described in proceedings of the SAMO conferences [2,4,18] as well as in the open literature [21].

The objective of this paper is to present a new global sensitivity analysis methodology which enables the analyst to determine what variables or interactions of variables drive model output into particular categories. The proposed methodology utilizes the classification tree analysis technique that is widely used in the field of medical decision-making and other scientific disciplines [1]. Tree-based modeling is an exploratory technique for uncovering structure in categorical and continuous data with such practical applications as rapid determination of prediction

* Corresponding author.
E-mail address: smishra@intera.com (S. Mishra).

0951-8320/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0951-8320(02)00222-3

rules, summary of large multivariate data sets and variable screening. Although tree-based models are useful for both classification and regression problems, we focus here on the former because standard global sensitivity analysis techniques are generally restricted to continuous rather than categorical outcomes. In particular, tree-based models are likely to be helpful in determining factors responsible for the separation between high- and low-dose outcomes, zero- and non-zero release outcomes, etc.

In what follows, we first review the principles of classification tree analysis along with some implementation details specific to the sensitivity analysis problem. Next, we describe illustrative applications of the methodology in a recently concluded probabilistic performance assessment study for the proposed nuclear waste repository at Yucca Mountain, Nevada, USA. Finally, we present some comments regarding the general applicability of the classification tree technique and how it compares with other common sensitivity analysis methods.

2. Classification tree analysis principles

Decision trees were first demonstrated and utilized in an automated fashion by social scientists. In the early 1960s, Morgan and Sonquist [15] developed automatic interaction detection, which was later extended into the THAID approach by Morgan and Messenger [14]. The fundamental reference in statistics regarding classification and regression trees is by Breiman et al. [1]. The computer algorithms used in the current study were implemented first by Clark and Pregibon [3] and then improved and extended by Venables and Ripley [22]. The tree-building theory in this section is based primarily on that reference.

The fundamental goal of a binary classification tree is to provide a set of rules or binary classifiers that enable the partitioning of an output variable y into two or more categories based on the set of input variables x. This partitioning occurs over a series of binary splits, with each split determined by the appropriate classifier.

Fig. 1 shows a simple example of binary splitting. In Fig. 1(a) we see the unordered data set, with two input variables $x1$ and $x2$, and a categorical (Low/High) output variable y. Fig. 1(b) shows the first binary split, where the data have been sorted in ascending order by variable $x2$, and split occurs between the values 6 and 7. This split results in a set of only High values ($x2 > 6.5$) and a mixed set of High and Low values ($x2 < 6.5$). Fig. 1(c) shows the second binary split. The data in the mixed set has been sorted in ascending order by variable $x1$, and split occurs between the values 4 and 5. This second split divides the mixed set into two sets of purely High and Low values, resulting in a total of three pure sets. Here, the purity of a set refers to the predominance of one category in that set. The bottom half of Fig. 1 shows a graphical representation of the classification tree. The points where the splits occur are called 'nodes'

x1	x2	y
4	5	High
3	1	High
1	3	High
5	4	Low
9	2	Low
11	6	Low
2	7	High
6	8	High
9	9	High
8	10	High
6	11	High
7	12	High

(a)

x1	x2	y
3	1	High
9	2	Low
1	3	High
5	4	Low
4	5	High
11	6	Low
2	7	High
6	8	High
9	9	High
8	10	High
6	11	High
7	12	High

(b)

x1	x2	y
1	3	High
3	1	High
4	5	High
5	4	Low
9	2	Low
11	6	Low
2	7	High
6	8	High
9	9	High
8	10	High
6	11	High
7	12	High

(c)

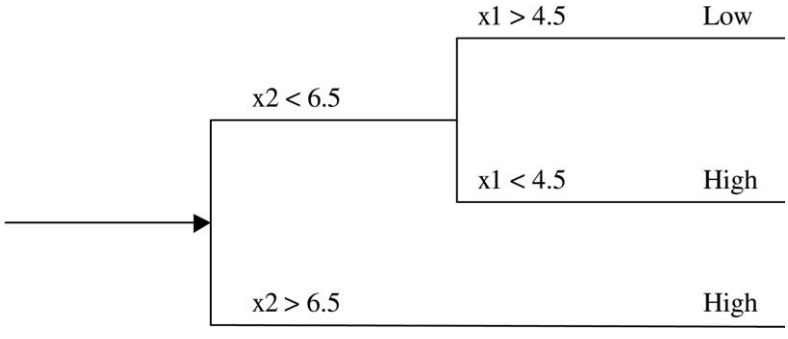

Fig. 1. Example of a simple classification tree, showing (a) the original dataset; (b) the first split using variable $x2$; (c) the second split using variable $x1$; and (d) the final categorical outcomes.