Technological Forecasting & Social Change 111 (2016) 124–134

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Improving scenario discovery by bagging random boxes

J.H. Kwakkel *, S.C. Cunningham

Delft University of Technology, Faculty of Technology, Policy and Management, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 24 January 2015
Received in revised form 10 June 2016
Accepted 12 June 2016
Available online 30 June 2016

Keywords:
Scenario discovery
Robust decision making
Exploratory modeling
Deep uncertainty

ABSTRACT

Scenario discovery is a model-based approach to scenario development under deep uncertainty. Scenario discovery relies on the use of statistical machine learning algorithms. The most frequently used algorithm is the Patient Rule Induction Method (PRIM). This algorithm identifies regions in an uncertain model input space that are highly predictive of model outcomes that are of interest. To identify these regions, PRIM uses a hill-climbing optimization procedure. This suggests that PRIM can suffer from the usual defects of hill climbing optimization algorithms, including local optima, plateaus, and ridges and valleys. In case of PRIM, these problems are even more pronounced when dealing with heterogeneously typed data. Drawing inspiration from machine learning research on random forests, we present an improved version of PRIM. This improved version is based on the idea of performing multiple PRIM analyses based on randomly selected features and combining these results using a bagging technique. The efficacy of the approach is demonstrated using three cases. Each of the cases has been published before and used PRIM. We compare the results found using PRIM with the results found using the improved version of PRIM. We find that the improved version is more robust to new data, can better cope with heterogeneously typed data, and is less prone to overfitting.

1. Introduction

Scenario discovery (Bryant and Lempert, 2010) is an approach for addressing the challenges of characterizing and communicating deep uncertainty associated with simulation models (Dalal et al., 2013). Deep uncertainty is encountered when the different parties to a decision do not know or cannot agree on the system model that relates actions to consequences, the exogenous inputs to the system model (Lempert et al., 2003). Decision problems under deep uncertainty often involve decisions that are made over time in dynamic interaction with the system (Hallegatte et al., 2012). When confronted by deep uncertainty, it is possible to enumerate the possibilities (e.g. sets of model inputs, alternative relationships inside a model), without ranking these possibilities in terms of perceived likelihood or assigning probabilities to them (Kwakkel et al., 2010). Scenario discovery addresses the challenge posed by deep uncertainty by exploring the consequences of the various deep uncertainties associated with a simulation model through conducting a series of computational experiments (Bankes et al., 2013). The resulting data set is subsequently analyzed using statistical machine learning algorithms in order to identify regions in the uncertainty space that are of interest (Bryant and Lempert, 2010, Kwakkel et al., 2013). These identified regions, which are typically characterized by only a small subset of the deeply uncertain factors, can subsequently be communicated to the actors involved in the decision problem. Preliminary experiments with real world decision makers suggest that scenario discovery results are decision relevant and easier to interpret for decision makers than probabilistic ways of conveying the same information (Parker et al., 2015).

Currently, the main statistical rule induction algorithm used for scenario discovery is the Patient Rule Induction Method (PRIM) (Friedman and Fisher, 1999). PRIM can be used for data analytic questions, where the analyst tries to find combinations of values for input variables that result in similar characteristic values for the outcome variables. Specifically, PRIM identifies one or more of hyper rectangular subspaces of the model input space within which the values of a single output variable are considerably different from its average values over the entire model input space. These subspaces are described as hyper-rectangular boxes of the model input space. To identify these boxes, PRIM uses a non-greedy, or patient, and hill climbing optimization procedure.

There are two key concerns when using PRIM for scenario discovery. The first concern is the interpretability of the results. Ideally the subspaces identified through PRIM should be composed of only a small subset of the uncertainties considered. If the number of uncertainties that jointly define the subspace is too large, interpretation of the results becomes challenging for the analyst (Bryant and Lempert, 2010). But, perhaps even more importantly, communicating such results to the stakeholders involved in the process becomes substantially more challenging (Parker et al., 2015). The second concern is that the uncertainties in the subset should be significant. That is, PRIM should

* Corresponding author.
E-mail address: j.h.kwakkel@tudelft.nl (J.H. Kwakkel).

http://dx.doi.org/10.1016/j.techfore.2016.06.014

only include uncertain factors in the definition of a subspace that are truly predictive for the characteristic values of the outcome variable. This concern is particularly important given that PRIM uses a lenient hill climbing optimization procedure for finding the subspaces. As such, PRIM suffers from the usual defects associated with hill climbing. The main defect is that hill climbing can only find a local optimum. Moreover, PRIM can get stuck on a plateau where the performance does not change resulting in an early stop of the optimization. PRIM can also get stuck by ridges and valleys which prevent the hill climbing algorithm from further improving the performance. Together, these defects imply that there might exist boxes that offer a better description of the data, but which cannot be found by the hill climbing optimization algorithm.

In current scenario discovery practice, the interpretability concern is addressed primarily by performing PRIM in an interactive manner. By keeping track of the route followed by the lenient hill climbing optimization procedure used in PRIM, the so-called peeling trajectory, a manual inspection can reveal how the number of uncertainties that define the subspace varies as a function of density (precision) and coverage (recall). This allows for making a judgment call by the analyst balancing interpretability, coverage, and density. To avoid the inclusion of spurious uncertainties in the subset, Bryant and Lempert (2010) propose a resampling procedure and a quasi-p-values test. This resampling test assesses how often essentially the same subspace is found by running PRIM on randomly selected subsets of the data. The quasi-p-value test, essentially a one sided binomial test, is an estimate of the likelihood that a given uncertainty is included in the definition of the subspace purely by chance.

In this paper, we investigate an alternative approach that addresses the interpretability concern and the significance concern simultaneously. This alternative approach is inspired by the extensive work that has been done with Classification and Regression Trees (CART) (Breiman et al., 1984) and related classification tree algorithms. The basic idea behind this alternative is to perform multiple runs of the PRIM algorithm based on randomly selected features (Breiman, 2001) and combining these results using a bagging technique (Breiman, 1996). The resulting algorithm is known as random forest (Breiman, 2001). The idea of random feature selection is that all the data is used, but rather than including all uncertainties as candidate dimensions, only a randomly selected subset is used. So, instead of repeatedly running PRIM on randomly selected data as currently done in the resampling procedure suggested by Bryant and Lempert (2010), this random feature selection procedure randomly selects the uncertainties instead. Bagging is an established approach in machine learning for combining multiple versions of a predictor into an aggregate predictor (Breiman, 1996). The expectation is that this random boxes approach will outperform normal PRIM, analogous to how a random forest outperforms a single classification tree.

To demonstrate the proposed approach and assess its efficacy compared to the normal use of PRIM in the context of scenario discovery, we apply it to three cases. In particular, we apply it to the same data as used in the paper of Bryant and Lempert (2010) in which Scenario Discovery was first proposed, the case study of Rozenberg et al. (2013), and the case used by Hamarat et al. (2014). The first case covers continuous uncertain factors, the second case covers discrete uncertain factors, and the third case has continuous, discrete, and categorical uncertain factors. This allows for a comparison between the original algorithm and the proposed approach across cases with differently typed uncertain factors.

The remainder of this paper is structured accordingly. In Section 2, we preset a review of the scenario discovery literature. In Section 3, we outline the method in more detail. More specifically, we introduce PRIM in Section 3.1, random forests in Section 3.2, and the combined approach in Section 3.3. Section 4 contains the results. We discuss the results in Section 5. Section 6 contains the conclusions.

2. Prior research

Scenario discovery was first put forward by Bryant and Lempert (2010). Their work builds on earlier work on the use of PRIM and CART in the context of Robust Decision Making (Lempert et al., 2006; Groves and Lempert, 2007, Lempert et al., 2008). Scenario discovery forms the analytical core of Robust Decision Making (Walker et al., 2013). Many examples of the use of scenario discovery in the context of Robust Decision Making can be found in the literature (Lempert et al., 2006; Lempert and Collins, 2007, Dalal et al., 2013; Hamarat et al., 2013; Matrosov et al., 2013a, 2013b; Auping et al., 2015; Eker and van Daalen, 2015). Robust Decision Making aims at supporting the design of policies that perform satisfactory across a very large ensemble of future states of the world. In this context, scenario discovery is used to identify the combination of uncertainties under which a candidate policy performs poorly, allowing for their iterative improvement. The use of scenario discovery for Robust Decision Making suggests that it could also be used in other planning approaches that design plans based on an analysis of the conditions under which a plan fails to meet its goals (Walker et al., 2013). Specially, Kwakkel et al. (2015) and Kwakkel et al. (2016) suggest that the vulnerabilities identified through scenario discovery can be understood as a multi-dimensional generalization of adaptation tipping points (Kwadijk et al., 2010), which are a core concept in the literature on dynamic adaptive policy pathways (Haasnoot et al., 2013).

Increasingly, scenario discovery is used more general as a bottom up model-based approach to scenario development. (Gerst et al., 2013; Kwakkel et al., 2013; Rozenberg et al., 2013; Halim et al., 2015; Greeven et al., 2016). There exists a plethora of scenario definitions, typologies, and methodologies (Bradfield et al., 2005; Börjeson et al., 2006). Broadly, three schools can be distinguished: the *La Prospective* school developed in France; the Probabilistic Modified Trends school originating at RAND; and the intuitive logic school typically associated with the work of Shell (Bradfield et al., 2005; Amer et al., 2013). Scenario discovery can be understood as a model-based approach to scenario development belonging to the intuitive logic school (Bryant and Lempert, 2010).

Scenario discovery aims to address several shortcomings of other scenario approaches. First, the available literature on evaluating scenario studies has found that scenario development is difficult if the involved actors have diverging interests and worldviews (van 't Klooster and van Asselt, 2006, European Environmental Agency, 2009, Bryant and Lempert, 2010). Rather than trying to achieve consensus, facilitate a process of joint sense-making to resolve the differences between worldviews, or arbitrarily imposing one particular worldview, scenario discovery aims at making transparent which uncertain factors actually make a difference for the decision problem at hand. An illustration of this is offered by Kwakkel et al. (2013) who capture two distinct mental models of how copper demands emerges in two distinct System Dynamics models and apply scenario discovery to both models simultaneously. Similarly, Pruyt and Kwakkel (2014) apply scenario discovery to three models of radicalization processes, which encapsulates three distinct mental models of how home grown terrorists emerge.

Another shortcoming identified in the evaluative literature is that scenario development processes have a tendency to overlook surprising developments and discontinuities (van Notten et al., 2005; Goodwin and Wright, 2010, Derbyshire and Wright, 2014). This might be at least partly due to the fact that many intuitive logic approaches move from a large set of relevant uncertain factors to a smaller set of drivers or megatrends. The highly uncertain and high impact drivers form the scenario logic. In this dimensionality reduction, interesting plausible combinations of uncertain developments are lost. In contrast, scenario discovery first systematically explores the consequences of all the relevant factors, and only then performs a dimensionality reduction in light of the resulting outcomes — thus potentially identifying surprising