Decision Tree Applications for Data Modelling

Man Wai Lee
Brunel University, UK

Kyriacos Chrysostomou
Brunel University, UK

Sherry Y. Chen[1]
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

Many organisations, nowadays, have developed their own databases, in which a large amount of valuable information, e.g., customers' personal profiles, is stored. Such information plays an important role in organisations' development processes as it can help them gain a better understanding of customers' needs. To effectively extract such information and identify hidden relationships, there is a need to employ intelligent techniques, for example, data mining.

Data mining is a process of knowledge discovery (Roiger & Geatz, 2003). There are a wide range of data mining techniques, one of which is decision trees. Decision trees, which can be used for the purposes of classifications and predictions, are a tool to support decision making (Lee et al., 2007). As a decision tree can accurately classify data and make effective predictions, it has already been employed for data analyses in many application domains. In this paper, we attempt to provide an overview of the applications that decision trees can support. In particular, we focus on business management, engineering, and health-care management.

The structure of the paper is as follows. Firstly, Section 2 provides the theoretical background of decision trees. Section 3 then moves to discuss the applications that decision trees can support, with an emphasis on business management, engineering, and health-care management. For each application, how decision trees can help identify hidden relationships is described. Subsequently, Section 4 provides a critical discussion of limitations and identifies potential directions for future research. Finally, Section 5 presents the conclusions of the paper.

BACKGROUND

Decision trees are one of the most widely used classification and prediction tools. This is probably because the knowledge discovered by a decision tree is illustrated in a hierarchical structure, with which the discovered knowledge can easily be understood by individuals even though they are not experts in data mining (Chang et al., 2007). A decision tree model can be created in several ways using existing decision tree algorithms. In order to effectively adopt such algorithms, there is a need to have a solid understanding of the processes of creating a decision tree model and to identify suitability of the decision tree algorithms used. These issues are described in subsections below.

Processes of Model Development

A common way to create a decision tree model is to employ a top-down, recursive, and divide-and-conquer approach (Greene & Smith, 1993). Such a modelling approach enables the most significant attribute to be located at the top level as a root node and the least significant attributes to be located at the bottom level as leave nodes (Chien et al., 2007). Each path between the root node and the leave node can be interpreted as an 'if-then' rule, which can be used for making predications (Chien et al., 2007; Kumar & Ravi, 2007).

To create a decision tree model on the basis of the above-mentioned approach, the modelling processes can be divided into three stages, which are: (1) tree growing, (2) tree pruning, and (3) tree selection.

Tree Growing

The initial stage of creating a decision tree model is tree growing, which includes two steps: tree merging and tree splitting. At the beginning, the non-significant predictor categorises and the significant categories within a dataset are grouped together (tree merging). As the tree grows, impurities within the model will increase. Since the existence of impurities may result in reducing the accuracy of the model, there is a need to purify the tree. One possible way to do it is to remove the impurities into different leaves and ramifications (tree splitting) (Chang, 2007).

Tree Pruning

Tree pruning, which is the key elements of the second stage, is to remove irrelevant splitting nodes (Kirkos et al., 2007). The removal of irrelevant nodes can help reduce the chance of creating an over-fitting tree. Such a procedure is particularly useful because an over-fitting tree model may result in misclassifying data in real world applications (Breiman et al., 1984).

Tree Selection

The final stage of developing a decision tree model is tree selection. At this stage, the created decision tree model will be evaluated by either using cross-validation or a testing dataset (Breiman *et al.*, 1984). This stage is essential as it can reduce the chances of misclassifying data in real world applications, and consequently, minimise the cost of developing further applications.

Suitability of Decision Tree Algorithms

A review of existing literature shows that the most widely used decision tree algorithms include the Iterative Dichotomiser 3 (ID3) algorithm, the C4.5 algorithm, the Chi-squared Automatic Interactive Detector (CHAID) algorithm, and the Classification and Regression Tree (CART) algorithm. Amongst these algorithms, there are some differences, one of which is the capability of modelling different types of data. As a dataset may be constructed by different types of data, e.g., categorical data, numerical data, or the combination of both, there is a need to use a suitable decision tree algorithm which can support the particular type of data used in the dataset. All of the above-mentioned algorithms can support the modelling of categorical data whilst only the C4.5 algorithm and the CART algorithm can be used for the modelling of numerical data (see Table 1). This difference can also be used as a guideline for the selection of a suitable decision tree algorithm. The other difference amongst these algorithms is the process of model development, especially at the stages of tree growing and tree pruning. In terms of the former, the ID3 and C4.5 algorithms split a tree model into as many ramifications as necessary whereas the CART algorithm can only support binary splits. Regarding the latter, the pruning mechanisms located within the C4.5 and CART algorithms support the removal of insignificant nodes and ramifications but the CHAID algorithm hinders the tree growing process before the training data is being overused (see Table 1).

DECISION TREE APPLICATIONS

Business Management

In the past decades, many organizations had created their own databases to enhance their customer services. Decision trees are a possible way to extract useful information from databases and they have already been employed in many applications in the domain of business and management. In particular, decision tree modelling is widely used in customer relationship management and fraud detection, which are presented in subsections below.

Customer Relationship Management

A frequently used approach to manage customers' relationships is to investigate how individuals access online services. Such an investigation is mainly performed by collecting and analyzing individuals' usage data and then providing recommendations based on the extracted information. Lee et al. (2007) apply decision trees to investigate the relationships between the customers' needs and preferences and the success of online shopping. In their study, the frequency of us-