Chapter XXXI

Fuzzy Decision–Tree–Based Analysis of Databases

Malcolm J. Beynon
Cardiff University, UK

ABSTRACT

The general fuzzy decision tree approach encapsulates the benefits of being an inductive learning technique to classify objects, utilising the richness of the data being considered, as well as the readability and interpretability that accompanies its operation in a fuzzy environment. This chapter offers a description of fuzzy decision tree based research, including the exposition of small and large fuzzy decision trees to demonstrate their construction and practicality. The two large fuzzy decision trees described are associated with a real application, namely, the identification of workplace establishments in the United Kingdom that pay a noticeable proportion of their employees less than the legislated minimum wage. Two separate fuzzy decision tree analyses are undertaken on a low-pay database, which utilise different numbers of membership functions to fuzzify the continuous attributes describing the investigated establishments. The findings demonstrate the sensitivity of results when there are changes in the compactness of the fuzzy representation of the associated data.

INTRODUCTION

The 1960s included the earliest discussion on an inductive approach towards the construction of a decision tree for the classification of objects that are described by attributes (Hunt, Marin, & Stone, 1966). It is not until over a decade later that the popularization of the decision tree method was materialized with the introduction of what developed into the Interactive Dichotomizer 3 (Quinlan, 1979, 1986). ID3, as it generally became to be known, involves the repetitive partitioning of objects through the augmentation of attributes down a tree until each subset of objects is associated with the same decision class or no attribute is available for further decomposition. Alternative
crisp decision tree methods include classification and regression trees (CARTs) and C4.5 (Breiman, Friedman, Olshen, & Stone, 1984; Janssens, Wets, Brijs, & Vanhoof, 2005; Quinlan, 1993).

Since the introduction of fuzzy set theory in Zadeh (1965) and its association with uncertain reasoning, including the incorporation of vagueness and ambiguity, it is understandable that the decision tree approach was developed in a fuzzy environment. The notion of fuzzy decision trees was first loosely referenced in the late 1970s (Chang & Pavlidis, 1977); like its crisp counterpart, it is closely involved with inductive learning, whereby it is the process of reaching a general conclusion from specific examples (Abdel-Galil, Sharkawy, Salama, & Bartnikas, 2005). Mitra, Konwar, and Pal (2002) succinctly describe a most important feature of decision trees, crisp and fuzzy, which is their capability to break down a complex decision-making process into a collection of simpler decisions and thereby provide an easily interpretable solution (see also Safavian & Landgrebe, 1991).

These characteristics are particularly pertinent when analysing databases (Bouchon-Meunier & Marsala, 1997), especially when an important feature of a fuzzy decision tree approach is the concomitant readability of the resultant fuzzy if-then decision rules constructed. The appropriateness of fuzzy decision trees in analysing databases was recently stated by Li, Zhao, and Chow (2006, p. 655):

*Decision trees based on fuzzy set theory combines the advantages of good comprehensibility of decision trees and the ability of fuzzy representation to deal with inexact and uncertain information.*

Their findings also highlight that fuzzy decision trees may not need extensive data to operate or intensive computing powers.

The formulations of fuzzy decision trees include derivatives of the well-known ID3 approach utilising fuzzy entropy (Ichihashi, Shirai, Nagasaki, & Miyoshi, 1996; their approach is subsequently regarded in terms of a neural network), as well as those that specifically take into account more cognitive uncertainties (Yuan & Shaw, 1995). The studies developing fuzzy decision trees directly from their crisp counterparts often include in their argument for the fuzzified adaptations of ID3 and similar that crisp decision trees suffer a weakness from having sharp boundaries when an attribute has continuous values. That is, small changes in attribute values describing an object may result in dramatic changes in its path down a crisp decision tree and its subsequent classification. In certain circumstances this may result in misclassification, which could have been avoided if the decision boundaries were more gradual (Crockett, Bandar, Melean, & O’Shea, 2006).

In contrast, the utilisation of fuzzy decision trees is not without its own potential problems. Crockett et al. (2006) states that the fundamental problem in developing any type of fuzzy system is in determining the shape, size, and number of fuzzy sets (commonly defined with membership functions, MFs) that should be used to represent the inputs and outputs of the proposed system. Moreover, it is usual for an expert’s intuition to be used to estimate the number of membership functions describing each attribute in the domain in question. It is worthy to note that Hong and Chen (1999) developed fuzzy if-then decision rule systems that derive the necessary membership functions from the available data. In their study, they included an entropy-based approach to measure the fitness of objects’ attributes to their consequents.

The particular fuzzy decision tree technique employed in this chapter is one of the independently defined approaches, and was introduced and developed in Yuan and Shaw (1995) and Wang, Chen, Qian, and Ye (2000). Their stated objective was to explicitly represent, measure, and incorporate cognitive uncertainties into a knowledge induction process for classification problems. Its use in real applications include the exposition of the antecedents of Sedge Warbler's song flight, in particular, a linguistic interpretation on the relationship between the birds’ characteristics like repertoire size and territory size against their