Constructing a Fuzzy Decision Tree by Integrating Fuzzy Sets and Entropy

TIEN-CHIN WANG (王天津)1 HSIEN-DA LEE (李賢達)1,2
1Department of Information Management
I-Shou University
tcwang@isu.edu.tw
2 Fortune Institute of Technology
Kaohsiung, Taiwan
leesd@center.fjtc.edu.tw

Abstract: - Decision tree induction is one of common approaches for extracting knowledge from a sets of feature-based examples. In real world, many data occurred in a fuzzy and uncertain form. The decision tree must able to deal with such fuzzy data. This paper presents a tree construction procedure to build a fuzzy decision tree from a collection of fuzzy data by integrating fuzzy set theory and entropy. It proposes a fuzzy decision tree induction method for fuzzy data of which numeric attributes can be represented by fuzzy number, interval value as well as crisp value, of which nominal attributes are represented by crisp nominal value, and of which class has confidence factor. It also presents an experiment result to show the applicability of the proposed method.

Key-Words: Fuzzy Decision Tree, Fuzzy Sets, Entropy, Information Gain, Classification, Data Mining

1 Introduction

Decision trees have been widely and successfully used in machine learning. More recently, fuzzy representations have been combined with decision trees. Many methods have been proposed to construct decision trees from collection of data. Due to observation error, uncertainty, and so on, many data collecting in real world are obtained in fuzzy forms. Fuzzy decision trees treat features as fuzzy variables and also yield simple decision trees. Moreover, the use of fuzzy sets is expected to deal with uncertainty due to noise and imprecision. The researches on fuzzy decision tree induction for fuzzy data have not yet sufficiently performed. This paper is concerned with a fuzzy decision tree induction method for such fuzzy data. It proposes a tree-building procedure to construct fuzzy decision tree from a collection of fuzzy data.

Decision trees and decision rules are data-mining methodologies applied in many real-world applications as a powerful solution to classification problem [1]. Classification is a process of learning a function that maps a data item into one of several predefined classes. Every classification based on inductive-learning algorithms is given as input a sets of samples that consist of vectors of attribute values and a corresponding class. For example, a simple classification might group students into three groups based on their scores: (1) those students whose scores are above 90 (2) those students whose scores are between 90 and 70 and (3) those students whose scores are below 70.

1.1 Fuzzy set theory

Fuzzy set theory was first proposed by Zadeh to represent and manipulate data and information that posses non-statistical uncertainty. Fuzzy set theory is primarily concerned with quantifying and reasoning using natural language in which words can have ambiguous meanings. This can be thought of as an extension of traditional crisp sets, in which each element must either be in or not in a set. Fuzzy sets are defined on a non-fuzzy universe of discourse, which is an ordinary sets. A fuzzy sets F of a universe of discourse U is characterized by a membership function \( \mu_F(x) \) which assigns to every element \( x \in U \), a membership degree \( \mu_F(x) \in [0,1] \). An element \( x \in U \) is said to be in a fuzzy sets F if and only if \( \mu_F(x) > 0 \) and to be a full member if and only if \( \mu_F(x) = 1 \) [5]. Membership functions can either be chosen by the user arbitrarily, based on the user’s experience, or they can be designed by using optimization procedures[6][7]. Typically, a fuzzy subset \( A \) can be represented as,

\[
A = \{ \mu_A(x_1)/x_1, \mu_A(x_2)/x_2, \ldots, \mu_A(x_n)/x_n \}
\]

Where the separating symbol \( / \) is used to associate the membership value with its coordinate on the horizontal axis. For example, in Fig.1, let...
3.3 Extract classification rules

Data classification is an important data mining task[2] that tries to identify common characteristics in a set of N objects contained in a database and to categorize them into different groups. We extract classification IF-THEN rules from those equivalence classes. For equivalence class \{4,6,10\}, those samples all have the identical attribute values: ETS=high, Admission=yes

So, we use the condition attribute values (ETS=high) as the rule antecedent and use the class label attribute value (Admission= yes) as the rule consequent, we can get the following classification rule:

IF ETS=“high” THEN Admission=“yes”

Similarly, the other classification rules can be extracted at this manner. We can get those rules as follows:

1. IF ETS=“high” THEN Admission=“yes”
2. IF ETS=“low” THEN Admission=“no”
3. IF ETS=“middle” AND GPA=“high” THEN Admission=“yes”
4. IF ETS=“middle” AND GPA=“low” THEN Admission=“no”

4 Conclusion

The paper is concerned with fuzzy sets and decision tree. We present a fuzzy decision tree model based on fuzzy set theory and information theory. It proposes a fuzzy decision tree induction method for fuzzy data of which numeric attributes can be represented by fuzzy number, interval value as well as crisp value, of which nominal attributes are represented by crisp nominal value, and of which class has confidence factor. An example is used to prove the validity. First, we applied fuzzy set theory to transform real-world data into fuzzy linguistic forms. Secondly, we used information theory to construct a decision tree. Finding the best split point and performing the split are the main tasks in decision tree induction method. Through the integration of both fuzzy set theory and information theory, it can make classification tasks originally thought too difficult or complex to become possible. It provides an alternative for evaluating the best possible candidates.

References:
