Chapter 33
Text Classification:
New Fuzzy Decision Tree Model

Ben Elfadhl Mohamed Ahmed
Higher Institute of Management, Tunisia

Ben Abdessalem Wahiba
Taif University, Saudi Arabia

ABSTRACT

In this chapter, a supervised automatic text documents classification using the fuzzy decision trees technique is proposed. Whatever the algorithm used in the fuzzy decision trees, there must be a criterion for the choice of discriminating attribute at the nodes to partition. For fuzzy decision trees usually two heuristics were used to select the discriminating attribute at the node to partition. In the field of text documents classification there is a heuristic that has not yet been tested. This chapter tested this heuristic. The latter was presented in the works of Yuan and Shaw (1995) and was applied in a context different then the textual classification. This heuristic is analyzed and adapted to the author's approach for text documents classification.

INTRODUCTION

Day by day, the world faces a huge amount of information which continues to increase rapidly. This entire amount requires the availability of effective means for its good management. A preliminary classification of a great source of information facilitates access to its content and its later manipulation. This principle is used in various fields such as databases, presses mails, some websites (the hierarchical classification of Yahoo for example), etc. The classification is divided into two branches (Sebastiani, 2002; Raheel, 2010) supervised classification (also called categorization) and unsupervised classification (also known as segmentation or clustering). This chapter focuses on the first type.

Supervised classification is performed to assign automatically and independently one or more documents to one or more predefined categories (Sebastiani, 2002). There are various techniques for supervised classification, among the best known: the Bayesian networks, support vector machines, k-nearest neighbors, decision trees, etc. Among these techniques, only the decision trees easily generate a set of

DOI: 10.4018/978-1-5225-2229-4.ch033

rules justifying the generated classification decisions. Other techniques generate in a more difficult and complicated way such set of rules.

Despite the wide spread of decision trees, this technique suffers from a problem that may affect its effectiveness: the problem of continuous values attributes. Let's take the example of a tree that will classify two men according to their sizes. The first has a height of 181cm; the second has a height of 180. The tree classifies a man as tall, if he has a height strictly larger than 180. In this example the tree will classify the first man as tall, but not the second, despite the invisible difference between the two sizes of the two men in the real world.

One of the solutions used to solve the problem of the classification's results sudden changes following continuous values changes, is the integration of fuzzy set theory with decision trees. This theory takes into account the continuity of values describing the phenomena of the real world and describes them in a graduated way closer to the reality (Janikow and Kawa, 2005).

The fuzzy decision trees allow benefit from the advantages offered by the combination of classical decision trees and fuzzy set theory. This combination uses the fuzzy representation and approximate reasoning ability with the symbolic power and ease of the classic decision trees interpretation.

A fuzzy decision tree is a good choice to use in the field of text classification, to manage a big problem in this type of classification which is the uncertainty and ambiguity necessarily related to the use of human language terms in the documents to be classified. In addition to their ability to handle the noise, the problem of missing or erroneous attributes, classification with fuzzy decision trees still retains the advantage of being easily understandable and interpretable.

Different models have been developed in the literature to construct fuzzy decision trees. Most of these models are based on the fuzzy algorithm ID3 (Matiasko et al, 2006), which is a direct extension of the ID3 algorithm (Quinlan, 1986).

The difference between these models often lies in the selection criterion of discrimination attribute and the way used to find the membership degrees of the used variables.

For the selection of the discrimination attribute, two heuristics have been used in the literature: The first is based on the minimization of the fuzzy entropy; the second is based on minimizing the classification ambiguity (Wang and al, 2000). In the area of text classification with fuzzy decision tree, in the author's literature search, only the first heuristic has been implemented and tested (Wang and Wang, 2005). For the second heuristic based on the minimization of the classification ambiguity, it has not been yet implemented, nor tested; despite it seems to well fit the context of the text classification due to the existing ambiguity related to the use of human terms that always can't describe perfectly what we want to say.

Minimizing classification ambiguity has been used by (Yuan and Shaw, 1995) in their model with sample classification on sport to practice according to the state of the climate described by four attributes. In this chapter, the authors will study and apply this heuristic for text classification.

BACKGROUND

Some works has been developed in the literature to build different models for fuzzy decision trees. Most of these models are based on fuzzy ID3 algorithm (Matiasko et al, 2006), which is a direct extension of ID3 algorithm (Quinlan, 1986). In this section we present some of the most famous models and the most cited in the literature.