Information Sciences 278 (2014) 715–735

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Fuzzy partitioning of continuous attributes through discretization methods to construct fuzzy decision tree classifiers

Mohsen Zeinalkhani *, Mahdi Eftekhari

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Article history:
Received 21 April 2012
Received in revised form 7 February 2014
Accepted 15 March 2014
Available online 3 April 2014

Keywords:
Discretization
Fuzzy decision tree
Fuzzy partitioning
Membership function

ABSTRACT

Generating Membership Functions (MFs) from data is one of the fundamental challenges associated with the applications of fuzzy set theory. This paper proposes a new two-step algorithm, which uses discretization methods for initial partitioning, to generate MFs from data. In the first step, discretization algorithm divides domain of attributes to several partitions, and then, in the second step, an MF is defined on each partition. Four different methods are proposed to define MFs in the second step: the first method is based on partition width, the second is based on standard deviation of examples, the third is based on Coverage Rate of Neighbor Partitions (NPCR) and the last one is based on Coverage Rate of Partition (PCR). Coverage rate of partition and coverage rate of neighbor partition are two new introduced parameters, which can be used for MF generation. In addition, this paper proposes a new MF generation algorithm, called Fuzzy Entropy Based Fuzzy Partitioning (FEBFP), which is a specific version of the proposed two-step algorithm with some modifications. FEBFP uses fuzzy entropy of partition to generate MFs and involves the parameters of MFs in the process of MF generation to combine two steps of the algorithm. Non-parametric statistical tests are used to compare Fuzzy Decision Trees (FDTs) induced using the MFs generated by the proposed methods (employing different discretization algorithms as well as four MF generation methods). Experimental results show that eight methods outperform the others in terms of both accuracy and number of nodes. Among them, trapezoidal MFs that are defined by PCR on partitions generated by Zeta discretization algorithm, outperform the others when the accuracy and complexity of FDT have the same degree of importance. Moreover, the results show that the PCR and NPCR MF definition methods perform better than the other ones.

1. Introduction

Performance of fuzzy rule-based classifiers is significantly influenced by employed Membership Functions (MFs). There are two general methods to determine MFs: Manual and Automatic [41]. In manual methods, experts manually determine MFs based on their experiments while in automatic methods, an algorithm automatically generates MFs based on available data. Automatic MF generation from data is one of the fundamental challenges in the application of fuzzy set theory. There are no guidelines or rules that can be used to choose the appropriate membership generation technique [37].

* Corresponding author. Tel./fax: +98 341 3235901.
E-mail address: zeinalkhani@gmail.com (M. Zeinalkhani).

http://dx.doi.org/10.1016/j.ins.2014.03.087
0020-0255/

Because evaluating the correctness of generated MFs is difficult, MFs reflecting subjective perceptions about imprecise concepts cannot be easily generated for many practical problems [37]. This problem becomes more complex when MFs are generated automatically from data and without the human interference.

There are many methods in the literature for MF generation from data, which their comprehensive survey can be found in Ref. [37]. Heuristic methods [26] make use of predefined shapes for MFs and have been applied successfully in rule-based pattern recognition applications. Histogram-based methods [37] use histograms of attributes that provide information regarding the distribution of input attribute values. Probability-based methods define MFs by converting probability distributions to possibility distributions [15]. Methods based on fuzzy nearest neighbor techniques use fuzzy nearest neighbor introduced by Keller et al. [28] to define MFs. Neural network based methods utilize feed forward multi-layer neural networks to generate MFs [37]. Clustering based methods [37] make use of clustering techniques like fuzzy c-mean to determine MFs. Recently, Malchiodi and Pedrycz [36] used a modified support vector clustering method to learn MFs, and Hong et al. [25] proposed a method based on ant colony to find MFs.

Fuzzy Decision Tree (FDT) induction algorithm [27,49,55] searches the domain of all possible fuzzy rules and selects a subset of them that leads to construction of the most accurate and simplest FDT. MFs defined on the domain of attributes determine the size and quality of fuzzy rules available in search space. Designing good MFs for all continuous attributes is the most important challenge of developing FDT classifier.

This paper proposes a new two-step algorithm based on discretization algorithms to generate MFs from data and uses the accuracy of FDT as a quality measure of generated MFs. In the first step, discretization algorithm divides domain of attributes to several partitions, and then, in the second step, an MF is defined on each partition. This paper also introduces four new methods for MF definition in the second step: 1 – MF definition based on width of partition; 2 – MF definition based on standard deviation of examples in the partition; 3 – MF definition based on coverage rate of neighbor partition; 4 – MF definition based on coverage rate of partition.

The proposed two-step algorithm uses a discretization method to partition the domain of attributes independent of the shape and parameters of used MFs. In addition, a new MF generation algorithm, called Fuzzy Entropy Based Fuzzy Partitioning (FEBFP), is proposed. This method considers the shape and parameters of MFs in the MF generation process. FEBFP is a modified version of Fayyad and Irani discretization method [16] that partitions the domain of attribute in a binary recursive manner such that minimizes the weighted entropy of generated partitions.

Non-parametric statistical tests are used to compare Fuzzy Decision Trees (FDTs) induced using the MFs generated by the proposed methods (employing different discretization algorithms as well as four MF generation methods). Experimental results show that eight MF generation methods (i.e. PCR-CAIM, PCR-Zeta, PCR-FEBFP, PCR-MantarasDist, PCR-Fayyad, NPCR-FEBFP, NPCR-Fayyad and NPCR-MantarasDist) outperform the others in terms of both accuracy and number of nodes. Among them, trapezoidal MFs that are defined by PCR on partitions generated by Zeta discretization algorithm, outperform the others when the accuracy and complexity of FDT have the same degree of importance. Moreover, the results reveal that the PCR and NPCR MF definition methods perform better than the other ones. In addition, the FUSINTER discretization method is the only one in which tuning its parameters can lead to performance improvement of FDTs in terms of both accuracy and complexity.

The organization of the paper is as follows. Section 2 briefly reviews the fuzzy decision trees. Section 3 provides a brief overview of discretization methods. Section 4 considers different kinds of membership functions. Section 5 describes proposed methods including fuzzy partitioning based on discretization methods, some other methods to define MFs over partitions, and FEBFP method. The experimental results are presented in Section 6. Finally, the last section concludes the paper.

2. Fuzzy decision tree

Fuzzy decision tree classifiers [1,27,49,50,55] combine decision trees with approximate reasoning offered by fuzzy representation to deal with language and measurement uncertainties. Fuzzy decision trees use fuzzy linguistic terms to specify branching condition of nodes and allow examples to simultaneously follow down multiple branches with different satisfaction degrees ranged on [0, 1]. Fig. 1 shows a typical expanded node of FDT. Each node of FDT has a dataset that provides

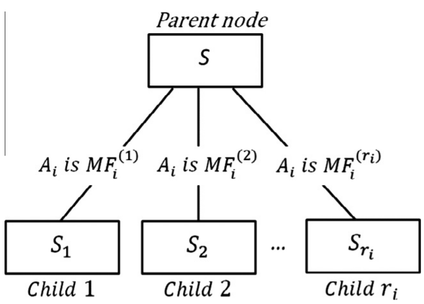

Fig. 1. A typical expanded node of fuzzy decision tree.