Iranian Journal of Fuzzy Systems

Volume 19, Number 4, (2022), pp. 73-88

Original paper

University of Sistan and Baluchestan

Fuzzy decision tree algorithm based on feature value's class contribution level

X. Y. Bai[1] and Y. L. Yang[2]

[1,2] *School of Mathematics and Statistics, Xidian University, Xi'an, PR China*
[1] *School of Science, Northwest A&F University, Yangling, PR China*

xuyingbai@nwafu.edu.cn, ylyang@mail.xidian.edu.cn

Abstract

Both fuzzy set theory and probability theory could handle uncertainty, and researchers have always compared the two to try to figure out which is better. Meanwhile, fuzzy decision tree algorithms represent classification knowledge more naturally to the way of human thinking. However, most generalizations of fuzzy decision tree algorithms focus only on numerical fine-tuning and interpretability of the algorithms. In order to make full use of the information of data sets and improve the performance of fuzzy decision trees. In this paper, we apply prior probability knowledge to the construction of fuzzy decision tree. Therefore, first of all, based on prior probability knowledge, for a certain feature variable, we present a concept of feature value's class contribution level which shows the different roles the same feature values play in different classes. Then, on the basis of feature value's class contribution level, we put forward a new algorithm called Feature Contribution Fuzzy Decision Tree (FCFDT). FCFDT algorithm has good classification results especially on data set with outliers. And it also maintains the interpretability of the fuzzy decision tree algorithms. The proposed algorithm is implemented and validated with seven state-of-the-art decision tree algorithms on 15 real-life data sets coming from the UCI Repository of machine learning. The experimental results obtained clearly indicate the superiority of the proposed scheme over the baseline methods.

Keywords: Fuzzy decision tree, fuzzy rule, fuzzy membership degree, feature value's class contribution level, prior probability knowledge.

1 Introduction

Fuzzy set theory provides a basis for fuzzy representation [31]. Rule-based fuzzy classification algorithms have become an important research area for classification tasks, due to the nature of continuous attributes and the various uncertainties present in problems. Because of the comprehensibility of fuzzy partitions, the complexity of fuzzy rules, and the simplicity of fuzzy reasoning, rule-based fuzzy classifiers are interpretable and transparent [12, 28]. Moreover, rule-based fuzzy reasoning methods conform to human epistemic habits. Therefore, the research of fuzzy classification algorithms has attracted more and more attention. Furthermore, rule-based fuzzy classification algorithms are fundamental to many application domains like computer vision, decision making, information retrieval, natural language processing, bioinformatics, pattern recognition, etc [5, 15, 23, 33, 34]. For example, based on fuzzy-rule, Pozna C. et al. proposed an approach that applies the signatures to expert systems modelling [20]. R. A. Borzooei et al. developed a product of fuzzy adjacency matrix based on t-norm fuzzy graphs [2]. In order to deal with big data sets, treating fuzzy C-means (FCM) and K-means (KM) algorithms as single configurable algorithm, Borlea I. D. et al. presented Unified Form (UF) clustering algorithm [1].

Decision trees are one of the most well-known classification methods, and there are many decision tree induction algorithms encountered in literature [11, 13, 19, 29]. Fuzzy decision trees more naturally represent the classification

Corresponding Author: X. Y. Bai
Received: October 2021; Revised: February 2022; Accepted: May 2022.
https://doi.org/10.22111/IJFS.2022.7088

knowledge of the human thinking and are more robust in tolerating imprecise, conflicting information. To deal with imprecise and uncertain data sets, rule-based fuzzy decision tree classification algorithms have been an active research topic. Researchers mainly focus on two aspects of fuzzy decision trees: feature selection and fuzzy rule extraction [14, 17, 24, 25, 27, 30, 32]. Quinlan proposed a fuzzy ID3 algorithm, which uses fuzzy sets to describe features and then constructs a decision tree using the ID3 method. Feature selection is then accomplished by a measure of fuzzy entropy, information gain or gain ratio [21]. Yuan and Shaw introduced a fuzzy decision tree induction method, which selects features based on the reduction of classification ambiguity with fuzzy evidence [30]. An induction of fuzzy decision trees is proposed by Boyen et al., in which the fuzzy sets are constructed automatically during the growth of trees, and a normalized Kolmogorov-Smirnov discrimination quality measure is applied to select features used in the node splitting [3, 4]. Liu et al. presented a fuzzy decision tree algorithm based on Axiomatic Fuzzy Set theory to extract fuzzy classification rules, which could be applied to data sets with mixed type attributes [17]. Most fuzzy decision trees take into account only one feature at each node. However, based on association rules extraction algorithm, Wang et al. put forward fuzzy rule based decision tree algorithm in which multiple features are considered at one node [25]. This algorithm is not only highly interpretable but also small in size. In general, the research of fuzzy decision tree algorithms has been flourishing and has made great achievements.

Uncertainty is a pervasive problem in real-world data. It is well known that both fuzzy set theory and probability theory could deal with uncertainty problems. However, since the introduction of the concept of fuzzy sets, the debate about what is better between fuzzy sets and probability theory has not stopped. It is often said that the mathematical modeling of fuzzy concepts lacks foundations. Drakopoulos has shown that for any given finite space, probabilities have strictly greater representational power than fuzzy sets [8]. Laviolette et al. believed that probability methods are sufficient for handling fuzziness [16]. Nguyen argued that if probabilistic approach is used as absolute meaning representations of fuzzy concept, one might make the mistake of considering the logical equivalence in [0, 1]-value logic to be equal to the logical equivalence in {0, 1}-value logic [18]. Actually, probabilistic approach and fuzzy set approach have their own advantages in solving uncertainty problems. Fuzzy set methods could effectively deal with the fuzziness of uncertainty, while probabilistic methods could well address the randomness of uncertainty. Most often, fuzziness and randomness of concepts are tightly related and inseparable.

Therefore, the approach of combining probability with fuzzy set theory is encountered in many literatures. Durairaj and Asha utilized fuzzy probability methods to identify people based on the probability values of real and imposter scores [10]. Cloud model describing qualitative concepts provides a new approach to linguistic decision making [26]. Based on probability theory and fuzzy set theory, it depicts the fuzziness and the randomness of qualitative concepts in detail. This makes it possible to convert between qualitative concepts and quantitative values. In literature [12], Gu et al. proposed Bayesian TSK fuzzy classifier, which casted the Takagi-Sugeno-Kang fuzzy classifier into the Bayesian inference framework. In this way, fuzziness and probability in the new classifier are collaboratively modeled to enhance the performance of TSK fuzzy classifier, in terms of classification performance and interpretability.

In order to make full use of the information of data sets to improve the classification performance of the fuzzy classifier, in this article, we apply prior probability knowledge to construct a fuzzy decision tree. For the convenience of calculation, without loss of generality, we assume that variables of instances are independent of each other. Since generalized fuzzy decision tree algorithms only focus on numerical fine-tuning and interpretability of algorithms, the contribution of specific feature values are neglected in the classification process. Due to the complexity and randomness of data sets, it is often the case that for a specific feature variable, different instances with the same feature value may belong to different classes. Obviously, this same feature value plays different roles in different classes. However, Wang et al. ignored this fact in the construction of the fuzzy rule based decision tree (FRDT) model [25]. Therefore, the FRDT algorithm is more sensitive to data sets containing outliers. Consequently, in this article, based on prior probability knowledge, for a given feature variable, we propose a concept of feature value's class contribution level, which is used to describe the role of a specific feature value for a certain class. Then, based on the FRDT algorithm, we apply the class contribution levels of feature values to the process of constructing fuzzy decision trees, and then present a new fuzzy classification algorithm called feature contribution fuzzy decision tree (FCFDT) algorithm. FCFDT algorithm refines the calculation of the membership degree of an instance belonging to each class and fully applies the information of data sets. It is a supplement to fuzzy algorithms based on fuzzy set theory, and improves the classification performance of classifiers. Meanwhile, the FCFDT algorithm has a good classification effect on data set with outliers, and it is generally robust. As shown in Figure 1, we present a flow chart of the FCFDT algorithm.

The remainder of this paper is structured as follows. In section 2, we introduce the concept of fuzzy numbers and the extraction of fuzzy rules. Then, we mainly present the concept of features' class contribution level and its solution method. The proposed FCFDT algorithm is elaborated in section 3. In section 4, we evaluate FCFDT algorithm by running it on benchmark data sets, and analyze the performance of the proposed algorithm. Section 5 concludes this article.

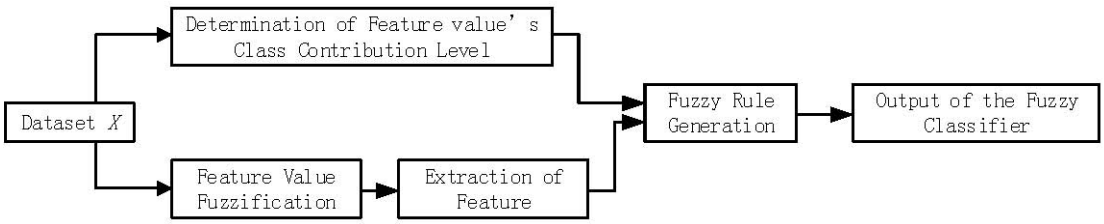


Figure 1: A flow chart of FCFDT algorithm.

2 Definitions and preliminaries

Fuzzy rule is an important part of fuzzy decision tree research, and fuzzy numbers are the basis for discussing the fuzzy rule. Therefore, in this section, we first introduce the concept of fuzzy numbers and the method of solving them [25]. And an example is given to illustrate the use of fuzzy membership degrees to represent linguistic interpretation. Then, the method of fuzzy rule extraction is presented [25]. At last, we elaborate on features' class contribution level. In general, the content of this section provides theoretical support for the presentation of FCFDT algorithm.

2.1 Fuzzy number

A data set X can be written as a $n(m+1)$ matrix. Each row of X represents an instance, and each column represents a feature variable except the last column which indicates the class labels of instances. Let's suppose f_j to denote the $j-$th column (feature variable) of X, $j = 1, 2, \cdots, m$, and $L = (y_1, y_2, \cdots, y_n)^T$ to denote the last column (class labels) of X. $X_i = (x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$ $(i = 1, 2, \cdots, n)$ represents the $i-$th instance, where x_{ij} is the value of the $j-$th feature of instance X_i, $(j = 1, 2, \cdots, m)$, and $y_i \in \{1, 2, \cdots, c\}$ is a class label of X_i, c is the number of classes. In data set X, Y_l contains all of instances which belong to the $l-$th class, $l \in \{1, 2, \cdots, c\}$.

Fuzzy numbers are the most basic concepts to describe fuzzy sets. There are many ways to construct fuzzy numbers, such as Gaussian membership functions, Sigmoid membership functions, rectangular membership functions, and so on. Different fuzzy numbers are applicable for different problems. Therefore, the choice of fuzzy numbers has a great influence on the classification effect of the model. Especially for complicated problems, the discussion of fuzzy numbers becomes particularly important. However, our research focuses on classification algorithms rather than fuzzy numbers. For the convenience of discussion, we utilize right-angle trapezoidal and triangular membership functions to form fuzzy numbers in this article. Each fuzzy number could be used to calculate the membership degree of a feature value belonging to the corresponding linguistic interpretation. We assume that the number of the fuzzy numbers defined for the $j-$th feature f_j is equal to c which is the number of classes. And F denotes the set of all fuzzy numbers. Let $f_{k,j} \in F$ represent the $k-$th $(k = 1, 2, \cdots, c)$ fuzzy number formed for the $j-$th feature f_j. Each triangular fuzzy number $f_{k,j}$ is characterized by three parameters $m_{k-1,j}$, $m_{k,j}$ and $m_{k+1,j}$. Each trapezoidal fuzzy number $f_{k,j}$ is characterized by two parameters. If $k = 1$, trapezoidal fuzzy number $f_{1,j}$ has two parameters $m_{1,j}$ and $m_{2,j}$. If $k = c$, trapezoidal fuzzy number $f_{c,j}$ has two parameters $m_{c-1,j}$ and $m_{c,j}$. The total number of the parameters of the fuzzy numbers $f_{k,j}$ $(k = 1, 2, \cdots, c)$ defined for the $j-$th feature f_j is equal to c which is the least number required to describe c classes. The value of the parameter $m_{k,j}$ $(k = 1, 2, \cdots, c)$ is taken as the mean value of all training instances falling within the $k-$th class Y_k when considering feature f_j. Let $\{m'_{k,j}, k = 1, 2, \cdots, c\}$ be calculated by formula (1) and $\{m_{k,j}, k = 1, 2, \cdots, c\}$ be the same values after sorting them in an ascending order.

$$m_{k,j} = \frac{\sum_{X_i \in Y_k} x_{ij}}{|Y_k|}. \tag{1}$$

It is easy to see that the boundaries formed of triangular fuzzy number $f_{k,j}$ is $(m_{k-1,j}+m_{k,j})/2$ and $(m_{k,j}+m_{k+1,j})/2$, the boundaries of trapezoidal fuzzy number $f_{1,j}$ is $(m_{1,j} + m_{2,j})/2$, and the boundaries of trapezoidal fuzzy number $f_{c,j}$ is $(m_{c-1,j} + m_{c,j})/2$. Fuzzy number $f_{k,j}$ defined in this way can explicitly describes the $k-$th class Y_k in terms of feature f_j, so that the linguistic interpretation of fuzzy numbers becomes very evident. Figure 2 shows the function graph of five fuzzy numbers for the $j-$th feature f_j, when $c = 5$.

In order to understand fuzzy number deeply, we provide a simple example here.

Example 2.1. *A synthetic data set is supplied to illustrate fuzzy number. As shown in Figure 3, the data set has* 620 *instances,* 310 *data points come from class 1, and* 310 *points are from class 2. There are two features to describe them. Considering the first feature* f_1*, two right-angled trapezoidal fuzzy numbers* $f_{1,1}$ *and* $f_{2,1}$ *are formed. According*

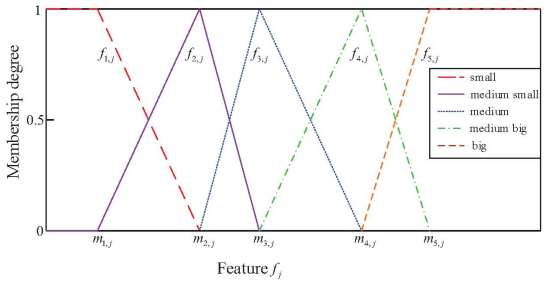


Figure 2: Right-angled trapezoidal and triangular forms of fuzzy numbers.

to Figure 4, $f_{1,1}(x_{i1})$ indicates the membership degree of feature value x_{i1} belonging to 'small', and $f_{2,1}(x_{i1})$ indicates the degree of membership of x_{i1} belonging to 'big'. Combined with Figure 3, it is easy to know that a data point would be in class 1 if the 1st feature value of it is small. A data point would belong to class 2, if the 1st feature value is big. Here, we utilize fuzzy number $f_{1,1}$ and $f_{2,1}$ to quantify linguistic description 'small' and 'big'.

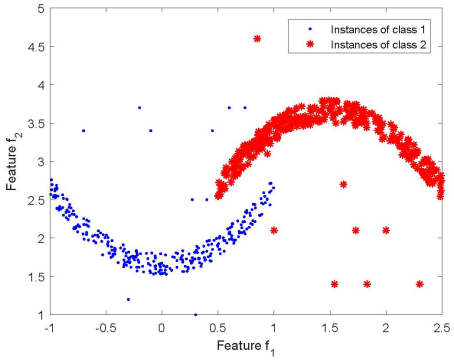


Figure 3: A synthetic data set.

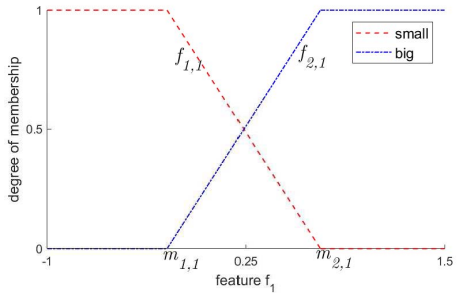


Figure 4: Two trapezoidal fuzzy number defined for the first feature f_1 of the synthetic data.

2.2 Extraction of fuzzy rule

Now, we will show the method of extracting fuzzy rules. In the process of forming fuzzy rules, fuzzy confidence level is utilized as a criterion for selecting fuzzy numbers. l is the class label of Y_l. For class Y_l, fuzzy confidence is defined as following.

$$Fconf(FR^{(w)} \Rightarrow l) = \frac{\sum_{X_i \in Y_l} FR^{(w)}(X_i)}{\sum_{X_t \in X} FR^{(w)}(X_t)}, \tag{2}$$

where $FR^{(w)}$ consists of w fuzzy numbers, i.e., $FR^{(w)} \subseteq F$. And $FR^{(w)}(X_i)$ would be acquired from formula (3).

$$FR^{(w)}(X_i) = \sum_{f_{k,j} \in FR^{(w)}} \frac{f_{k,j}(x_{ij})}{w}, \tag{3}$$

In formula (3), x_{ij} is the value taken for instance X_i at feature f_j. Fuzzy number $f_{k,j}(x_{ij})$ represents the membership degree of x_{ij} belonging to one linguistic interpretation. $FR^{(w)}$ ($FR^{(w)} \subseteq F$) consists of w fuzzy numbers with high values of Fuzzy Confidence, $(1 \leq w \leq min(MaxL, c * m))$, where $MaxL$ is the maximum length of fuzzy rules. Then, $FR^{(w)}$ could be used as the antecedent of fuzzy "if-then" rule of class Y_l. An optimal value of w could be acquired by the following formula.

$$w = \min\{v \mid Fconf(FR^{(v)} \Rightarrow l) - Fconf(FR^{(v+1)} \Rightarrow l) > \beta\}, \tag{4}$$

where β is a certain threshold to establish a sound balance between the length of the fuzzy rule and its fuzzy confidence level. Therefore, the value of parameter β is set to 0.02 to keep the length of fuzzy rules low, and the maximum length of fuzzy rules MaxL is 5, because five features are enough to describe the region of a class [17, 25].

2.3 Feature value's class contribution level

In classification learning, a classification model may be a great one when the probability distribution of data for each class is well described. The contribution level is defined as a probability distribution function $P(\cdot)$, that belongs to a definite group of distributions, denoted as $\{P(\cdot|\theta), \theta \in \Theta\}$, where, θ is element of the model parameters and Θ denotes the entire set of observations [7].

The objective of $P(\cdot)$ is to determine an estimator that would predict the true value of the parameter, θ. Assuming that instances $X_1, X_2, \cdots, X_t$ are independent of each other ($\{X_1, X_2, \cdots, X_t\} \subseteq Y_l$), $P(\cdot)$ can be interpreted as a joint probability distribution of all observations belonging to a specific class, which can be defined as

$$P(X_1, X_2, \cdots, X_t|l) = P(X_1|l)P(X_2|l)\cdots P(X_t|l), \tag{5}$$

where l is the class label of instances, and $l = 1, 2, \cdots, c$.

$P(X_i|l)$ represents the conditional probability of event X_i given that its class label is l. That is to say, in the distribution of the $l-$th class instances, the probability of X_i is $P(X_i|l)$. It is well known that different features play different roles in the process of classification. Consequently, it is necessary to study the contribution level of each feature to every class. In this article, under the assumption of independence of the features (namely, $P(x_{i1}, x_{i2}, \cdots, x_{im}) = P(x_{i1})P(x_{i2})\cdots P(x_{im})$), we study the distribution of each feature. As described in Section 2.1, we utilize f_j to represent the $j-$th feature column of a data set, where $f_j = \{x_{1j}, x_{2j}, \cdots, x_{nj}\}$. Then, we present the definition of feature's class contribution level.

Definition 2.2. *Given a data set X, $P(Y_l|f_j)$ means the conditional probability of Y_l given that we know the instances' $j-$th feature column is f_j. Then, we call $P(Y_l|f_j)$ as the $j-$th feature f_j to Y_l's class contribution level, denoted as $Contr(f_j \Rightarrow Y_l)$.*

Obviously, $Y_1, Y_2, \cdots, Y_c$ are mutually exclusive and exhaustive, and prior probability $P(Y_1), P(Y_2), \cdots, P(Y_c)$ could be easily obtained from data set X. According to Bayes' rule, feature's class contribution level could be acquired by the following conclusion.

Proposition 2.3. *In data set X, for the $l-$th class Y_l, the $j-$th feature f_j's class contribution level is given by*

$$Contr(f_j \Rightarrow Y_l) = P(Y_l|f_j) = \frac{P(f_j|Y_l)P(Y_l)}{\sum_{i=1}^{c} P(f_j|Y_i)P(Y_i)}. \tag{6}$$

In feature column f_j, feature value u often appears many times. The class labels of the instances to which u belongs are not always the same. Obviously, u may play a great role in several classes while minor in others. Therefore, it is appropriate and important to employ feature value's class contribution level in the learning process of classification model. Now, in feature column f_j, we put forward the definition of the class contribution level of feature value $u(u \in f_j)$.

Definition 2.4. *Given a data set X, considering the $j-$th feature f_j, $P(Y_l|u)$ means the conditional probability of instances belonging to the $l-$th class Y_l given that we know the instances' $j-$th feature value is u. Then, we call $P(Y_l|u)$ as the feature value u to Y_l's class contribution level in the $j-$th feature column, denoted as $Contr(u_{f_j} \Rightarrow Y_l)$.*

Just like Proposition 2.3, applying the knowledge of conditional probability, we draw a conclusion about feature value's class contribution level.

Proposition 2.5. *Given a data set X, in the $j-th$ feature column f_j, feature value $u(u \in f_j)$ to Y_l's class contribution level $Contr(u_{f_j} \Rightarrow Y_l)$ is given by*

$$Contr(u_{f_j} \Rightarrow Y_l) = \frac{P(u|Y_l)P(Y_l)}{P(u)}. \tag{7}$$

Here, we would show how to calculate feature value's class contribution level in learning classification model. According to Glivenko-Cantelli theorem, we acquire the following conclusion.

Theorem 2.6. *In data set X, let's take the $j-th$ feature column f_j into account. u is a feature value which belongs to f_j. S_i $(i = 1, 2, \cdots, c)$ denotes the set of instances whose class label is i and of which the $j-th$ feature value is u. Then, in the feature column f_j, feature value u to Y_l's class contribution level, $Contr(u_{f_j} \Rightarrow Y_l)$, is given by*

$$Contr(u_{f_j} \Rightarrow Y_l) = \frac{|S_l|}{\sum_{i=1}^{c} |S_i|}, \tag{8}$$

where $|S_i|$ denotes the number of instances in set S_i, c is the number of classes of which data set X consists.

Proof. According to Glivenko-Cantelli theorem, we could substitute the ratio of an event for its probability. $P(Y_l)$ represents the probability of instances of the $l-$th class Y_l appearing in the data set X. $P(u|Y_l)$ represents the probability of the $j-th$ feature value u occurring for all instances belonging to the $l-$th class Y_l. Consequently, probability $P(Y_l)$ and $P(u|Y_l)$ can be respectively described as

$$P(Y_l) = \frac{|Y_l|}{n}, \qquad P(u|Y_l) = \frac{|S_l|}{|Y_l|}, \tag{9}$$

where n is the number of instances of data set X.

Therefore, according to Proposition 2.5 and Bayes' Theorem, we have

$$Contr(u_{f_j} \Rightarrow Y_l) = \frac{P(u|Y_l)P(Y_l)}{P(u)} = \frac{P(Y_l)P(u|Y_l)}{\sum_{i=1}^{c} P(Y_i)P(u|Y_i)} = \frac{(|Y_l|/n)(|S_l|/|Y_l|)}{\sum_{i=1}^{c}(|Y_i|/n)(|S_i|/|Y_i|)} = \frac{|S_l|}{\sum_{i=1}^{c}|S_i|}. \tag{10}$$

□

In this paper, on the basis of fuzzy rule-based decision tree (FRDT) algorithm, applying the feature value's class contribution level, we propose a new fuzzy classification algorithm.

3 Over all architecture of FCFDT

Both fuzzy set and probability theory could deal with uncertain problems. Incorporating the probability knowledge into the learning process of the fuzzy classification algorithms not only makes full use of the information provided by data sets, but also improves the classification performance of classification algorithms. The feature value's class contribution level is proposed based on the knowledge of probability distribution. In this article, we apply the feature's class contribution level to the construction of fuzzy decision tree algorithm and propose a new classification algorithm called feature contribution fuzzy decision tree (FCFDT).

3.1 The algorithm of FCFDT

Different from the common axis-parallel decision tree that only considers one feature at each split node, FCFDT algorithm is an oblique decision tree that simultaneously takes multiple features into account at each split node. Each layer of the FCFDT decision tree contains c leaf nodes and one non-leaf node. Then, we will describe the main procedure of the FCFDT algorithm. As shown in Figure 5, the set of all training instances is recorded as the root node $UL^{0,\delta}$. According to the information of the data set, fuzzy rules of each class are extracted. The set composed of these fuzzy rules is denoted as FR_1. The unclassified instances in $UL^{0,\delta}$ constitute non-leaf node $UL^{1,\delta}$. We learn fuzzy rules in the second layer again. In this layer, FR_2 is the set of fuzzy rules obtained by learning instances in $UL^{1,\delta}$. The unclassified

instances of set $UL^{1,\delta}$ are recorded as node $UL^{2,\delta}$. Fuzzy rules are learned in the third layer once again. This is done until the end condition is satisfied. The entire procedure could be seen in Figure 5. In the fuzzy rule set FR_i obtained by learning the $i-$th data set $UL^{i-1,\delta}$, $FR_{i,l}$ represents the antecedent of a certain "IF-THEN" rule, and class label l denotes the consequent of it.

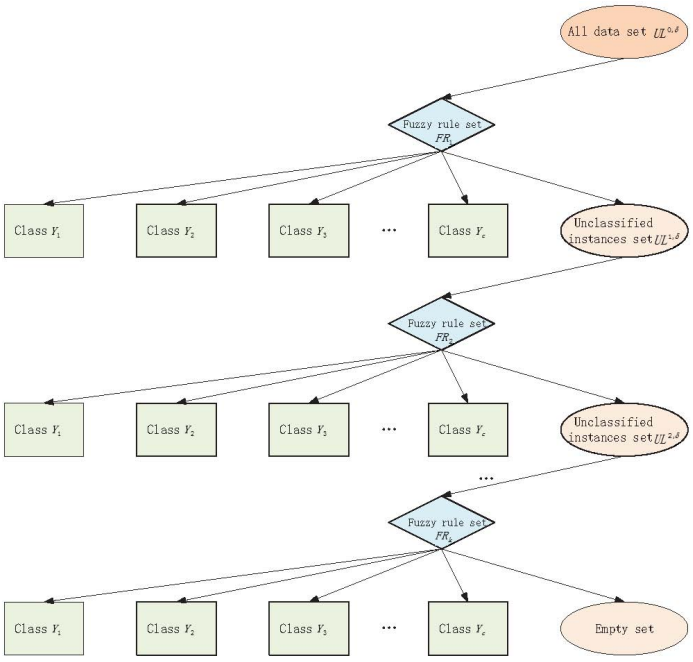


Figure 5: A structure of feature contribution fuzzy decision tree.

The FRDT algorithm determines the fuzzy membership degree of an instance belonging to each class based on the average value of the fuzzy numbers in the fuzzy rule $FR_{i,l}$ [25]. However, it ignores the probability distribution of the data set, and is particularly sensitive to data sets containing outliers. There may be such a case that a feature value of an instance belongs to a class with a large fuzzy number value, but in fact its contribution to the class is small. In this situation, it may lead to incorrect classification, if only the fuzzy number values of instances are considered. The application of the prior probability knowledge of data set provides more information and enhances the classification ability of the model. Therefore, it is necessary to integrate feature value's class contribution level into the fuzzy number values of feature value when we determine the fuzzy membership degree of an instance belonging to each class.

In training instances set, the class contribution level of each feature value should be calculated by formula (8). $h_l(x_{ij})$ denotes the l-th Y_l's class contribution level of feature value x_{ij}. $f_{k,j}$ represents the $k-$th fuzzy number of the $j-$th feature. $f_{k,j}(x_{ij})$ signifies the fuzzy number value. That is the membership degree of feature value x_{ij} belonging to fuzzy number $f_{k,j}$. In order to compute fuzzy number values, parameter $m_{k,j}$ should be determined according to formula (1).

Then, new fuzzy membership degree of an instance belonging to each class is generated by combining its feature value's class contribution level and fuzzy number values. For instance $x_t = (x_{t1}, x_{t2}, \cdots, x_{tm})$, in the $i-$th layer classification task, we utilize $PF_{i,l}(X_t)$ to represent the membership degree of X_t belonging to the $l-$th class Y_l. And $PF_{i,l}(X_t)$ is calculated according to the following formula.

$$PF_{i,l}(X_t) = \sum_{f_{k,j} \in FR_{i,l}} f_{k,j}(x_{tj}) h_l(x_{tj}) / |FR_{i,l}|, \tag{11}$$

where $|FR_{i,l}|$ represents the number of fuzzy numbers in set $FR_{i,l}$.

From formula (11), we could obtain that the fuzzy membership degree of an instance belonging to one class is determined by the fuzzy number values of its feature values and corresponding feature value's class contribution level. In the case of calculating fuzzy membership degree, the feature value's class contribution level is used as the weight to adjust the role of different fuzzy number values. This method overcomes the influence of outliers and noise on the classification algorithm to a certain extent.

Now, on the basis of the previous discussion, the pseudocode of FCFDT classification method will be shown in Algorithm 1. And CL_l denotes the set of instances which are labeled l ($l = 1, 2, \cdots, c$) by our new algorithm FCFDT.

Algorithm 1: Feature contribution fuzzy decision tree (FCFDT) algorithm.

Input: X is a set of training data composed of c classes $Y_1, Y_2, \cdots, Y_c$, $0 < \delta < 1$, i represents the $i-$th layer of the tree.
Output: $FR_{i,l}, UL^{i,\delta}, CL_l$

For each $X_s = (x_{s1}, x_{s2}, \cdots, x_{sm}) \in X$, compute feature value' class contribution level $h_l(x_{sj})$.
For each feature f_j, compute the value of fuzzy number $f_{t,j}(x_{sj}), j = 1, 2, \cdots, m$.
$UL^{0,\delta} = X, i = 1, CL_l = \varnothing, l = 1, 2, \cdots, c$.
While $UL^{i-1,\delta} \neq \varnothing$
    $c_i \Leftarrow$the number of classes in $UL^{i-1,\delta}$
    If $c_i = 1$, break, end.
    Extract fuzzy rule set $FR_{i,l}$ in data set $UL^{i-1,\delta}, l = 1, 2, \cdots, c$.
    Compute fuzzy membership degree $PF_{i,l}(X_s) = \sum_{f_{k,j} \in FR_{i,l}} f_{k,j}(x_{sj}) h_l(x_{sj}) / |FR_{i,l}|$.
    $UE = UL^{i-1,\delta}, UL^{i,\delta} = \varnothing$.
    While $UE \neq \varnothing$
      If $\max_l\{PF_{i,l}(X_s)\} \geq \delta$
        $t_i = \arg\max_l\{PF_{i,l}(X_s)\}$.
        $CL_{t_i} = CL_{t_i} \cup \{X_s\}$.
      else $UL^{i,\delta} = UL^{i,\delta} \cup \{X_s\}$.
      end.
    $UE = UE \setminus \{X_s\}$.
    end.
    If $UL^{i,\delta} = UL^{i-1,\delta}$, or $UL^{i,\delta} = \varnothing$, break, end.
    $i = i + 1$
end.

• In the process of learning algorithm, there are three conditions for program termination. One is that node $UL^{i,\delta}$ is empty. Another one is that the instances in set $UL^{i-1,\delta}$ belong to the same class. The last one is that $UL^{i,\delta} = UL^{i-1,\delta}$. In this case, the instances in the set $UL^{i-1,\delta}$ are classified as the class with the largest number of instances.

• FCFDT decision tree is constructed under the guidance of the threshold δ. The parameter δ can control the size of the tree and has little influence on the classification effect of the FCFDT decision tree [25].

• Due to utilizing prior probability knowledge, FCFDT algorithm is robust, especially for data sets with outliers.

• The main advantage of the FCFDT algorithm is that the calculation method of the fuzzy membership degree of instances belonging to each class is improved by combining the prior probability knowledge. Therefore, the classification capability of the algorithm is restricted by the basic algorithm.

3.2 Analysis of time complexity

According to the FCFDT algorithm, for a feature column in a data set, the fewer the number of different feature values, the better the performance of the algorithm. In a data set with m feature variables, feature value's class contribution level is computed cm times corresponding to each class and each feature. In the worst case, for a feature column, all values are different. Thus, the frequency of computation of feature value's class contribution level for a particular data set is cmn. In each layer, it takes at most cmn calculations to obtain fuzzy rules FR, and at most n calculations to classify instances, respectively. The number of layers is n in the worst situation that is each layer of the resulting tree for the training data set can classify only one instance. In general, the number of the layer of the tree is no more than 10. Therefore, in this algorithm, the sum of frequencies of all sentences is $cmn + 10cmn + 10n$. Consequently, the total time complexity of the FCFDT algorithm is $O(cmn)$.

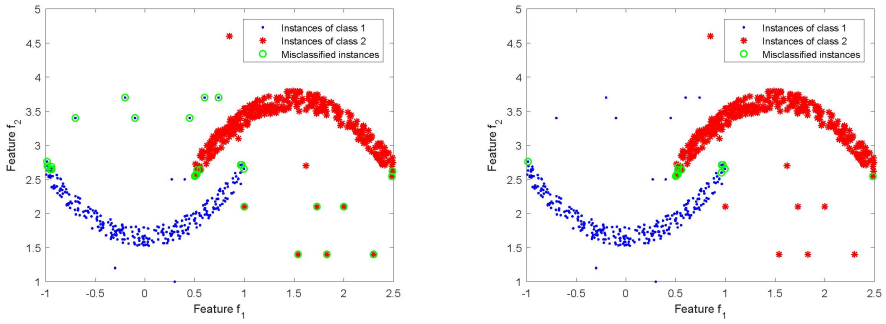
4 Numerical experiments and analysis

In this section, some numerical experiments are performed to demonstrate the performance of the FCFDT algorithm proposed in the previous section. The experiments mainly focus on the classification performance of the algorithm. We achieve this task in two ways. On the one hand, we compare the classification performance of FCFDT and FRDT algorithms on data sets containing outlier points. On the other hand, we compare the FCFDT algorithm with seven different decision tree algorithms on real data sets from the UCI Repository of machine learning databases. Finally, we analyze the effect of the parameter delta on the FCFDT model.

4.1 Experiments on data sets containing outliers

Since the basic algorithm of FCFDT is FRDT algorithm, we only compare FCFDT with FRDT algorithm when considering the classification performance of the proposed algorithm for data sets containing outliers.

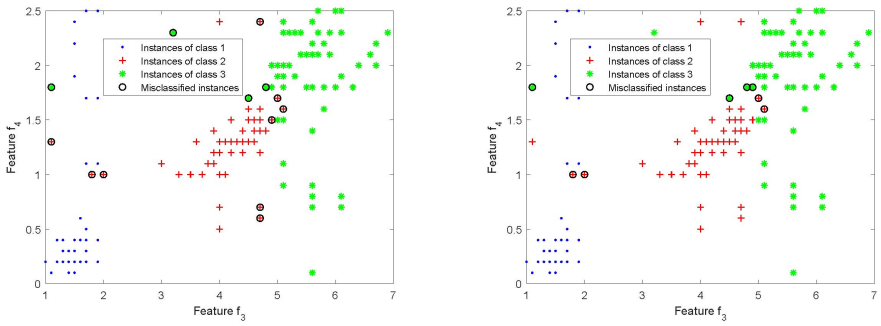
In the first place, let us consider the visualized two-dimensional synthetic data set with outliers given in Figure 3 above. The classification results are shown in Figure 6. The number of misclassified testing instances produced by FRDT is 29, while that produced by FCFDT is 13. In addition, it can be seen from Figure 6 that the number of outliers of misclassification by FCFDT is significantly smaller than that by FRDT algorithm.



(a) FRDT algorithm (b) FCFDT algorithm

Figure 6: Classification on the synthetic data set by FRDT and FCFDT, respectively.

Because the features of different classes of iris data set are obvious, the difference in the effect of classification by different algorithms is not very large. We add 30 instances to iris data set, and then classify it by FRDT and FCFDT algorithms. The number of misclassified testing instances by FRDT and FCFDT algorithms are 15 and 8, respectively. We project the four-dimensional iris data set onto the two-dimensional plane composed of the third and the fourth features. The classification results are shown in Figure 7. We note that the FCFDT algorithm outperforms the FRDT algorithm in classifying datasets with outliers.



(a) FRDT algorithm (b) FCFDT algorithm

Figure 7: Classification on iris data set by FRDT and FCFDT, respectively.

In order to more clearly compare the classification ability of the FCFDT and the FRDT algorithms for data sets containing outliers, we present the number of instances that are misclassified for outliers and boundary points in the form of tables.

Table 1: Nmuber of misclassified instances for outliers and boundary points on the synthetic data set.

Algorithms	Number of misclassified outliers	Number of misclassified boundary points
FRDT	12	28
FCFDT	0	13

Table 2: Number of misclassified instances outliers and boundary points on iris data set with outlier added.

Algorithms	Number of misclassified outliers	Number of misclassified boundary points
FRDT	8	7
FCFDT	3	5

From the previous experiments and analysis, we obtain that the classification performance of the FCFDT algorithm is better than the FRDT algorithm. Especially for the data sets with outliers, the FCFDT algorithm is more robust.

4.2 Experiments on data sets coming from UCI Repository of machine learning

In order to further verification of the effectiveness of our new algorithm, we experimentally compared FCFDT with seven state-of-the-art algorithms on data sets coming from the UCI Repository of machine learning [9]. The 15 data sets are shown in Table 3.

Table 3: Statistics of data sets used in the experiment along with the number of features (m), the number of classes (c), and the number of instances (n) .

No.	Data set	m	c	n
1	iris	4	3	150
2	wine	13	3	178
3	wdbc	30	2	569
4	credit	14	2	690
5	heart	13	2	270
6	haberman	3	2	306
7	newthyroid	5	3	215
8	wobc	9	2	699
9	column_3C	6	3	310
10	wisconsin	9	2	683
11	balance_scale	4	3	625
12	blood	4	2	748
13	lymphography	18	4	148
14	cleveland	13	5	303
15	page-block	10	5	5472

The performance of the FCFDT algorithm is quantified by taking the classified accuracy of each data set. In order to make a clearer comparison, we present the number of misclassified instances of each algorithm. In these experiments, a ten-fold cross-validation is carried out.

The general procedure of ten-fold cross-validation is as follows. The data set is divided into ten equal parts, and nine of them are used as training data and one as test data in turn for the experiments. The corresponding misclassified instances are obtained for each experiment. The sum of the misclassified instances of the 10 results is used as an estimation of the classification effect of the algorithm. In this paper, we use the average of ten times 10-fold cross-validation results to measure the classification performance of each algorithm. In order to obtain the best results for the FCFDT model, we utilize two-step cross-validation to complete the tasks [22].

First, let $\delta = 0.1, 0.11, 0.12, \cdots, 0.60$, for each δ, and using the training set as the entire data set, we will derive the optimal result by running a five-fold cross-validation. The parameter corresponding to this optimal result is determined as the optimal parameter. Then, this parameter is used to construct the FCFDT tree.

In these experiments, we present a detailed comparative analysis of the proposed algorithm with other seven decision tree algorithms. They are C4.5 decision tree (C4.5), Best-first decision tree (BFT), A multi-class alternation decision tree (LAD), Simple cart (SC), Naive Bayes decision tree (NB), Fuzzy rule based decision tree (FRDT), and Evolutionary learning of globally optimal classification and regression tree (EVTREE).

The details can be found in the software Weka (http://www.cs.waikato.ac.nz/ml/weka/). For these seven algorithms we use their Weka implementation, and the values of the parameters for all the algorithms were set to their defaults being available in the toolkit. Moreover, the classification effect obtained by software Weka is often better than manual tuning.

Table 4: Number of misclassified testing instances and standard deviation for FCFDT and other decision tree classifiers, the best scores are indicated in boldface.

Data set	BFT	C4.5	LAD	SC	NBT	FRDT	EVTREE	FCFDT
iris	8.4±1.3	7.9±1.2	8.3±1.3	8.7±1.5	9.8±1.9	6.2± 10.6	9.6±2.0	**6.0 ±1.1**
wine	18.6±2.2	12.1±2.4	23±3.2	18.7±3.0	**7.0±2.4**	10.8±1.5	9.5±1.6	8.4±1.1
wdbc	39.6±3.7	35.5±2.8	53.2±6.9	38.9±3.5	34.4±4.6	28.2±1.7	36.7±4.7	**28.0±1.6**
credit	106.2±3.6	111.0±4.9	141.6±12.8	105.7±4.5	109.2±4.1	102.3±4.8	102.0±2.8	**96.3±2.5**
heart	61.5±4.6	59.0±6.1	74.6±5.2	59.2±4.4	51.5±3.3	44.6±3.0	51.1±4.3	**39.2±1.6**
haberman	84.4±4.5	85.2±3.4	90.2±7.4	81.9±3.7	87.0±4.0	81.2±2.3	81.1±3.4	**75.0±1.9**
newthyroid	15.2±1.7	15.9±2.1	23.9±2.2	17.5±2.0	16.4±3.0	14.4±3.1	15.9±1.9	**11.0±1.1**
wobc	38.8±3.8	34.9±3.1	42.7±5.3	36.8±2.6	25.4±3.0	33.0±2.6	30.3±4.1	**17.7±0.8**
column_3C	61.8±4.7	57.2±3.7	70.3±5.3	59.3±4.0	59.8±4.7	57.1±2.4	**54.0±2.7**	57.0±1.2
wisconsin	34.2±2.5	32.6±2.9	26.9±2.0	32.3±3.0	25.3±3.6	31.3±2.1	31.6±3.5	**15.1±0.6**
balance_scale	131.6±7.1	138.0±7.1	94.0±7.3	132.0±7.0	142.9±8.0	162.8±6.1	140.6±6.2	**71.4±1.9**
blood	164.3±5.3	162.9±2.9	**162.7±5.9**	165.7±3.7	180.9±5.2	190.1±1.8	177.7±2.8	166.8±1.2
lymphography	32±4.1	32.9±2.2	32.1±4.2	34.2±3.1	36.5±4.0	36.5±5.8	38.5±4.3	**19.1±1.7**
cleveland	138.8±3.0	142.2±6.8	143.3±3.4	135.7±3.4	139.5±5.0	133.6±8.4	135.8±7.2	**127.1±3.9**
page-block	186.1±6.3	170.1±8.8	210.1±6.8	177.5±5.4	174.2±11.0	491.0±13.6	**169.2±7.2**	313.0±7.6

The misclassified testing instances and standard deviation for the experiments on different data sets are summarized in Table 4.

The best scores are shown in boldface. FCFDT outperforms other algorithms on most data sets. Because of the application of prior probability knowledge, especially on the balance_scale data set, the FCFDT algorithm has great superiority. However, on the page-block data set, the classification effect of the FCFDT algorithm is not satisfactory.

This is because the FCFDT algorithm is developed on the basis of the FRDT algorithm, and the FRDT algorithm is sensitive to large scale class imbalance data. In order to observe the classification effect of eight algorithms on 15 different data sets more intuitively, we give some histograms shown in Figure 8.

In general, we would observe very clearly that the classification effect of the FCFDT algorithm is better than other methods.

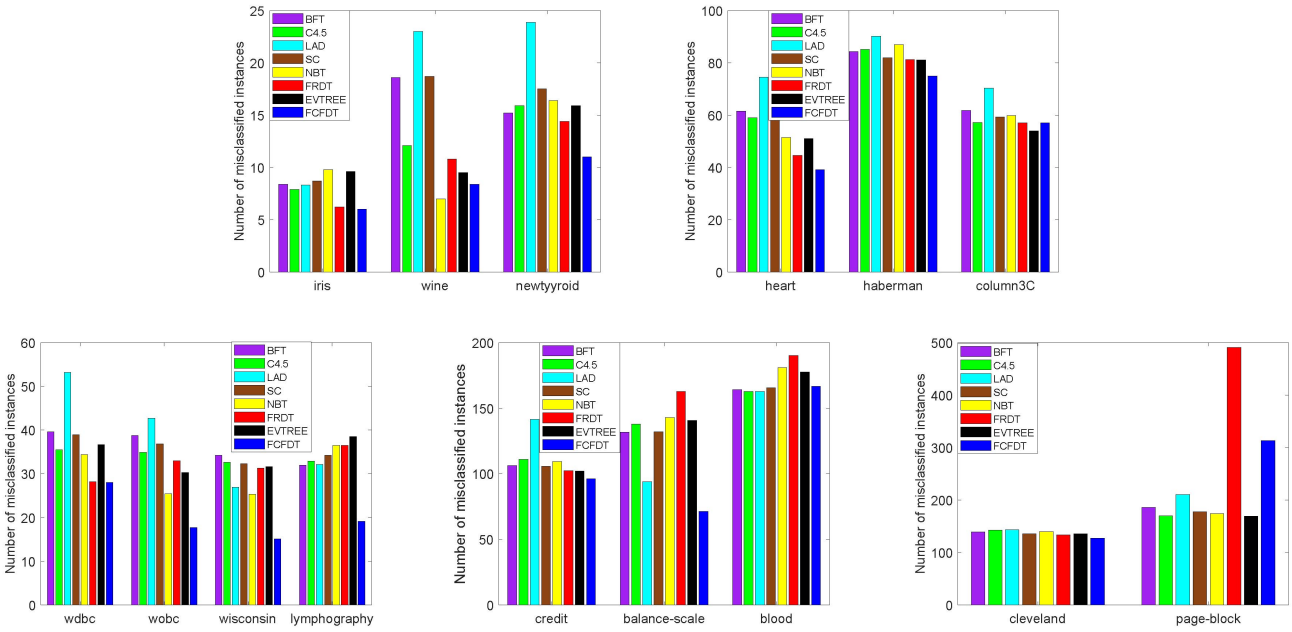


Figure 8: Number of misclassified instances of 8 algorithms in 15 data sets.

4.3 Analysis of parameter

Because unlike the traditional decision trees that only considers one feature variable at each split node, the FCFDT algorithm simultaneously considers multiple feature variables at each split node. Therefore, similar to the FDRT algorithm, the size of the FCFDT decision tree is much smaller than the traditional decision tree model.

FCFDT decision tree is constructed under the guidance of the threshold δ. From the classification procedure of the decision tree, the parameter δ can control the size of the tree. The smaller the δ, the smaller the purity of the leaf nodes and the smaller the tree.

Conversely, the larger the δ, the greater the purity of the leaf nodes, and the larger the size of the tree. In fact, the classification effect of the FCFDT algorithm is less affected by δ. Because multiple feature variables of the instance to be classified are considered at the same time, the possibility of the instance being classified into the correct class is very high.

In our numerical experiments, let $\delta = 0.1 : 0.01 : 0.6$, the classification accuracy of the FCFDT algorithm in each dataset is shown in Figures 9 and 10. We obtain that δ does not have a significant effect on the classification accuracy for most of the data sets, except for the balance_Scale and blood data sets.

Actually, judging from the characteristics of the feature variables of these two data sets, they are more suitable for being classified by a crisp classifier rather than a fuzzy classifier.

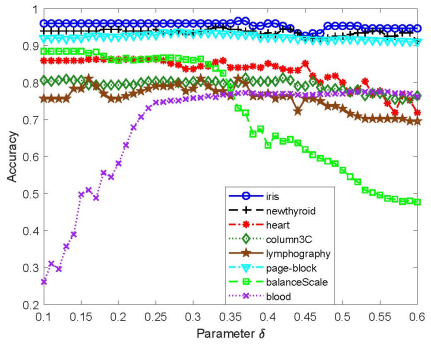


Figure 9: Accuracy of FCFDT for testing data for different δ for iris, newhyroid, heart, column_3C, lymphography, page-block, balance_Scale, blood.

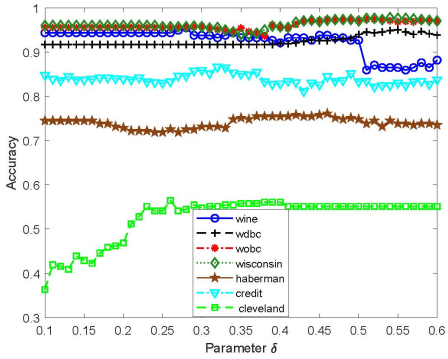


Figure 10: Accuracy of FCFDT for testing data for different δ for wine, wdbc, wobc, wisconsin, haberman, credit, cleveland.

4.4 Friedman test

However, due to the limitations of the testing data sets, using only classification accuracy is not enough to compare the classification performance of the algorithms. We should adopt other statistical and common-sense techniques to decide whether the difference between these algorithms are real or random. In this paper, we utilize Friedman test to examine whether the classification performance of the eight algorithms is the same. If the null-hypothesis that "all algorithms have the same performance" is rejected, the performance of these algorithms are significantly different. Then, we will proceed with a Holm post-hoc test [6].

Let N denote the number of the testing data sets, k be the number of the classifier algorithms. And R_i^j represents the rank of the $j-$th of k algorithms on the $i-$th of N data sets. The Friedman test compares the mean ranks of each algorithm, $Rank_j = \frac{1}{N}\sum_{i=1}^{N} R_i^j$. Under the null-hypothesis, which states that the ranks of all the algorithms are equivalent and so $Rank_j$ should be equal, the Friedman statistic

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\Big[\sum_j Rank_j^2 - \frac{k(k+1)^2}{4}\Big], \tag{12}$$

is distributed according to χ^2 with $k-1$ degree of freedom, when N and k are big enough (as a rule of a thumb, $N > 10$ and $k > 5$). Due to the "original Friedman's χ^2 test" is undesirably conservative, and a better statistic is derived

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}, \tag{13}$$

which is distributed according to the $F-$distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. The table of critical values could be found in any statistical book.

In our study, $Rank_{BFT} = 5.200, Rank_{C4.5} = 4.700, Rank_{LAD} = 6.067, Rank_{SC} = 5.133, Rank_{NBT} = 4.967, Rank_{FRDT} = 4.100, Rank_{EVTREE} = 4.033, Rank_{FCFDT} = 1.800$, $N = 15, k = 8$. Combining formulas (12) and (13), we get $\tau_F = 5.135$. With eight algorithms and 15 data sets, τ_F is distributed according to the F distribution with 7 and 98 degrees of freedom. The critical value of $F(7, 98)$ for significance level $\alpha = 0.05$ is 2.104, and $5.135 > 2.104$. Hence, we reject the null-hypothesis.

Then, we proceed with a Holm post-hoc test. The test statistics for comparing the $i-$th and $j-$th classifier using the following methods.

$$z = \frac{(Rank_i - Rank_j)}{SE}, \tag{14}$$

where $SE = \sqrt{\frac{k(k+1)}{6N}}$. The z value is used to find the corresponding probability value p from the table of normal distribution, which is utilized to compare with an appropriate α. We should sort the probability value p in ascending order, that is $p_1 \leq p_2 \leq \cdots \leq p_{k-1}$. The Holm's step-down procedure compares each p_i with $\alpha/(k-i)$, but differ in the order of the test. It starts with the most significant p value. If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on. Once a certain null-hypothesis cannot be rejected, all the remaining hypotheses are retained as well.

In this study, $SE = \sqrt{(8 \times 9)/(6 \times 15)} = 0.894$. The Holm procedure rejects the first, the second, the third, the fourth, the fifth, the sixth and the seventh hypothesis, since the corresponding p values are smaller than the adjusted α's (see Table 5). This shows that FCFDT performs significantly better than NBT, SC, BFT, LAD, C4.5 and EVTREE at the significance level $\alpha = 0.05$.

Table 5: Holm test.

i	Classifier	z	p	$\alpha/(k-i)$
1	LAD	5.375	0.0000	0.0071
2	BFT	3.801	0.0001	0.0083
3	SC	3.727	0.0002	0.0100
4	NBT	3.541	0.0004	0.0125
5	C4.5	3.242	0.0012	0.0167
6	FRDT	2.572	0.0101	0.0250
7	EVTREE	2.497	0.0125	0.0500

5 Conclusions and future work

Fuzzy decision tree algorithms represent classification knowledge more naturally to the way of human thinking. Since their strong interpretability and good classification accuracy, fuzzy decision tree algorithms are increasingly favored by researchers. However, they focus on numerical fine-tuning and interpretability of algorithms, not considering the relevant prior probability knowledge provided by data sets. This will affect the classification performance of the algorithms, especially for data sets with outlier points.

In the process of learning model, for a given feature variable, one certain feature value may belong to different instances. And these instances may fall within different classes. As it is well known, the certain feature value may play different role in different class, but fuzzy rule-based decision tree algorithm treats them equally. Therefore, based on prior probability theory, we present the concept of feature value's class contribution level and apply it to fuzzy rule-based decision tree algorithm. A new algorithm called FCFDT is proposed.

As we all know, outliers in data sets have a very adverse impact on the learning of the algorithm, especially the fuzzy classification algorithms which use membership function to calculate the membership degree of instances belonging to the corresponding class. At the same time, we also know that outliers have interference effects in each class, and the positive effect is very small. Based on the feature value's class contribution level, FCFDT algorithm weakens the role of outliers in learning algorithms and achieves better classification results. Nevertheless, FCFDT algorithm does not have great advantages in dealing with data sets with redundant instances.

In direction of future work, the concept of feature values class contribution level could be incorporated into other fuzzy algorithms. Another research direction is the application of FCFDT to feature selection in fuzzy decision trees.

Conflict of interest

There are no conflicts of interest.

Acknowledgments

The author thanks the editors and the anonymous reviewers for helpful comments and suggestions. The research was supported by the Natural Science Basic Research Program of Shaanxi (Program No. 2021JM-133).

References

[1] I. D. Borlea, R. E. Precup, A. B. Borlea, D. Iercan, *A unified form of fuzzy C-means and K-means algorithms and its partitional implementation*, Knowledge-Based Systems, **214** (2021), 106731.

[2] R. A. Borzooei, B. Sheikh Hoseini, M. Mohseni Takallo, *Results on t-fuzzy graphs*, New Mathematics and Natural Computation, **16**(01) (2020), 143-161.

[3] X. Boyen, L. Wehenkel, *Fuzzy decision tree induction for power system security assessment*, IFAC Proceedings Volumes, **28**(26) (1995), 299-304.

[4] X. Boyen, L. Wehenkel, *Automatic induction of fuzzy decision trees and its application to power system security assessment*, Fuzzy Sets and Systems, **102** (1999), 3-19.

[5] W. Cao, X. Wang, Z. Ming, J. Gao, *A review on neural networks with random weights*, Neurocomputing, **275** (2018), 278-287.

[6] J. Demsar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, **7** (2006), 1-30.

[7] D. Devi, S. K. Biswas, B. Purkayastha, *Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance*, Pattern Recognition Letters, **93** (2017), 3-12.

[8] J. A. Drakopoulos, *Probabilities, possibilities, and fuzzy sets*, Fuzzy Sets and Systems, **75** (2017), 1-15.

[9] D. Dua, K. Taniskidou, *UCI machine learning repository*, Irvine, CA: University of California, School of Information and Computer Science, (2017), http://archive.ics.uci.edu/ml.

[10] M. Durairaj, J. H. M. Asha, *Fuzzy probability based person recognition in smart environments*, Journal of Fuzzy Systems, **40**(5) (2021), 9437-9452.

[11] N. EI, I. Karabadji, I. Khelf, H. Seridi, et al., *A data sampling and attribute selection strategy for improving decision tree construction*, Expert Systems with Applications, **129** (2019), 84-96.

[12] X. Gu, F. L. Chung, S. Wang, *Bayesian Takagi-Sugeno-Kang fuzzy classifier*, IEEE Transactions on Fuzzy Systems, **25**(9) (2017), 1655-1671.

[13] Q. Hu, X. Che, L. Zhang, et al., *Rank entropy-based decision trees for monotonic classification*, IEEE Transactions on Knowledge and Data Engineering, **24**(11) (2012), 2052-2064.

[14] C. Jin, F. Li, Y. Li, *A generalized fuzzy ID3 algorithm using generalized information entropy*, Knowledge-Based Systems, **64** (2014), 13-21.

[15] A. Kumar, M. Hanmandlu, H. M. Gupta, *Fuzzy binary decision tree for biometric based personal authentication*, Neurocomputing, **99**(1) (2013), 87-97.

[16] M. Lavioleue, J. W. Seaman, J. D. Barrett, W. H. Woodall, *A probabilistic and statistical view of fuzzy methods*, Technometrics, **37**(3) (1995), 249-261.

[17] X. Liu, X. Feng, W. Pedrycz, *Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (AFS) approach*, Data and Knowledge Engineering, **84** (2013), 1-25.

[18] H. T. Nguyen, *Fuzzy sets and probability*, Fuzzy Sets and Systems, **90** (1997), 129-132.

[19] A. Paez, F. Lopez, M. Ruiz, *Inducing non-orthogonal and non-linear decision boundaries in decision trees via interactive basis functions*, Expert Systems with Applications, **122** (2019), 183-206.

[20] C. Pozna, R. E. Precup, *Applications of signatures to expert systems modelling*, Acta Polytechnica Hungarica, **11**(2) (2014), 21-39.

[21] J. R. Quinlan, *Induction of decision trees*, Machine Learning, **1** (1986), 81-106.

[22] Y. Ren, X. Liu, J. Cao, *A parsimony fuzzy rule-based classifier using axiomatic fuzzy set theory and support vector machines*, Information Science, **181**(23) (2011), 5180-5193.

[23] L. Rokach, O. Z. Maimon, *Data mining with decision trees: Theory and applications*, World Scientific, (2nd edition), 2015.

[24] S. K. Savas, E. Nasibov, *A fuzzy ID3 induction for linguistic data sets*, International Journal of Intelligent Systems, **33** (2018), 858-878.

[25] X. Wang, X. Liu, W. Pedrycz, L. Zhang, *Fuzzy rule based decision trees*, Pattern Recognition, **48** (2015), 50-59.

[26] J. Wang, P. Lu, H. Zhang, X. Chen, *Method of multi-criteria group decision-making based on cloud aggregation operators with linguistic information*, Information Sciences, **274** (2014), 177-191.

[27] J. Wang, Y. Qian, F. Li, *Fusing fuzzy monotonic decision trees*, IEEE Transactions on Fuzzy Systems, **28**(5) (2020), 887-900.

[28] Y. Yang, X. Bai, *A research on classification performance of fuzzy classifiers based on fuzzy set theory*, Iranian Journal of Fuzzy Systems, **16** (2019), 15-27.

[29] W. Yi, M. Lu, Z. Liu, *Multi-valued attribute and multi-labels data decision tree algorithm*, International Journal of Machine Learning and Cybernetics, **2**(2) (2011), 67-74.

[30] Y. Yuan, M. J. Shaw, *Induction of fuzzy decision trees*, Fuzzy Sets and Systems, **69** (1995), 125-139.

[31] L. A. Zadeh, *Fuzzy sets*, Information and Control, **8** (1965), 338-353.

[32] J. Zhai, X. Z. Wang, S. Zhang, S. Hou, *Tolerance rough fuzzy decision tree*, Information Sciences, **465** (2018), 425-438.

[33] S. Zhao, H. Chen, C. Li, et al., *A novel approach to building a robust fuzzy rough classifier*, IEEE Transactions on Fuzzy Systems, **23**(4) (2015), 769-786.

[34] H. Zhu, X. Wang, *A cost-sensitive semi-supervised learning model based on uncertainty*, Neurocomputing, **251** (2017), 106-114.