An Algorithm Combining Random Forest Classification and Fuzzy Comprehensive Evaluation

Jing Zhu[1,2], Song Huang[1*], Yaqing Shi[1], Kaishun Wu[1], Yanqiu Wang[3]

[1] Command and Control Engineering College, Army Engineering University of PLA, China
[2] Training Management Department, Navy Command College, China
[3] Baopo technology co. LTD, China
391740172@qq.com, huangsong@aeu.edu.cn, 77745227@qq.com, ksw.macs@gmail.com, 317177028@qq.com

Abstract

Random forest algorithm is a common classification method. However, if the weights of many attributes in a data set are not same or close to each other, the direct use of this algorithm for data training will lead to the neglect of the interrelationships between these attributes, and it is difficult to reflect the differences brought by different weights of different attributes. Worse, if the number of attributes in the data set is relatively large, many attributes will be given very little weight when normalization is satisfied, which will also lead to information loss. All of these will have a negative impact on the final result. To solve these problems, this paper proposes an algorithm combining random forest classification and fuzzy comprehensive evaluation, which not only take into account the correlation between attributes in data training, but also retain the information in the original data set to the maximun. At the same time, this algorithm significantly improves the accuracy of random forest training results.

Keywords: Random forest, Fuzzy, Comprehensive evaluation

1 Introduction

When we deal with a data set with a large number of attributes, if the values of these attributes are not specific or clearly defined, the amount of calculation in data training will be large and the classification effect will be poor. At the same time, even though the values of each attribute are accurately defined, due to the large number of attributes, it is inevitable that the weight assigned to some attributes will be very small, and the loss of information will be inevitable [1-2]. In that case, the accuracy of the results would be negatively affected when the random forest method is directly used to train such data sets [3].

In addition, during sample training, the inherent defects of random forest will also have an uncertain impact on the training results, due to the large number of attributes and the large number of attribute value division. In order to solve this problem, we adopt the method of fuzzy comprehensive evaluation in fuzzy theory. The method divided all the attributes into two levels, which not only significantly reduced the number of splits, but also improved the accuracy of the final results through fuzzy comprehensive evaluation followed by random forest training [4].

At present, there are also some algorithms that try to combine random forest theory and fuzzy theory, however, to date there is almost no use of the theory of fuzzy comprehensive evaluation algorithm; moreover, these algorithms are using fuzzy theory to deal with the results of the random forest algorithm instead of using the fuzzy comprehensive evaluation method to improve the accuracy of the random forest training results, which is exactly what this paper studies.

For data sets with a large number of attributes, when the values of these attributes are not clearly defined or the values are not absolute but in an ambiguous state, this paper proposes an algorithm combining random forest classification and fuzzy comprehensive evaluation. This algorithm can not only reduce the number of attributes involved in the calculation, but also further improve the accuracy of random forest training results while preserving the information of the original data set to the greatest extent.

A preliminary version of the partial context of this paper was presented at the 8th International Conference on Dependable Systems and Their Applications (DSA) [5].

In the first section, the paper introduces the algorithm combining random forest classification and fuzzy comprehensive evaluation.

In the second section, the paper gives a brief description of the related work and research in quo on random forest and fuzzy comprehensive evaluation.

In the third section, the theory of the algorithm combining random forest classification and fuzzy comprehensive evaluation is expounded.

In the fourth section, K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF), the four of the commonly used single classification methods, and four kinds of improved RF methods based on fuzzy comprehensive evaluation, are all experimented on 21 data sets. After the experiment, it is verified that among the 4 improved RF methods, 3 of them have higher average accuracy than the RF method itself, whose average accuracy is obviously higher than the KNN, NB, and DT methods.

In the fifth section, this paper concludes what the optimal fuzzy operator is, and shows that the average accuracy of the RF method based on fuzzy comprehensive evaluation with this operator is 74.74%, which is 4.94% higher than the result of the RF method itself. Therefore, the new methods proposed in this paper are worthy of further research and expansion.

*Corresponding Author: Song Huang; E-mail: huangsong@aeu.edu.cn
DOI: 10.53106/160792642022072304009

2 Related Work

RF classification was proposed by Leo Breiman and Adele Cutler in 1995 and belongs to the category of machine learning [6]. The fuzzy theory was developed for many years, which originated from the concept of fuzzy sets proposed by an American automatic control expert Professor L. A. Zadeh in 1965, which is mainly used to express the uncertainty of transactions [7].

The algorithm combining random forest classification and fuzzy comprehensive evaluation proposed in this paper not only solves the uncertainty caused by the fuzziness of data, but also makes use of the advantages of comprehensive evaluation and combines with the random forest training effect to further improve the accuracy.

2.1 The Method of Fuzzy Comprehensive Evaluation

The condition of a thing is often related to a variety of factors, and the so-called comprehensive evaluation is to make a general evaluation of the thing or phenomenon determined by a variety of factors. It allows an object to have a hierarchy of membership between full membership and non-membership, which reflects the degree of an element or a factor belong to the set.

The fuzzy factor set $U = \{u_1, u_2, \cdots, u_{\mathrm{n}}\}$ refers to a factor-set that influences the result of evaluation. When it comes to a specific problem, each factor itself is determined by many child-factors $\{u_1, u_2, \cdots, u_n\}$. At this point, the set U can also be regarded as a parent-factor. Fuzzy comprehensive evaluation is to evaluate each parent-factor individually and then make comprehensive evaluation.

In 2003, Tzung-Pei Hong et al. proposed a new learning algorithm based on rough sets to find cross-level certain and possible rules from training data with hierarchical attribute values, which is more complex than learning rules from training examples with single-level values, but may derive more general knowledge from data [8]. In 2009, Tzung-Pei Hong et al. extended their previous approach to deal with the problem of producing a set of cross-level maximally general fuzzy certain and possible rules from examples with hierarchical and quantitative attributes, which combines the rough-set theory and the fuzzy-set theory to learn [9]. In 2017, W. Ma and Y. Wang et al. combined AHP and fuzzy theory in order to evaluate interuniversity collaborative learning, which provides a new thought for interuniversity collaborative learning evaluation based on network [10]. In 2019, Zhu et al. developed a fuzzy comprehensive evaluation method based on cloud model, whose outcomes display the consistency, representativeness, robustness, and superiority of this evaluation method, which make the evaluation results more scientific and objective [11]. In 2020, Xueling Wu and Fang Hu also used a fuzzy comprehensive method and an analytic hierarchy process to provide reasonable weights for ECC evaluation modelling by combining subjective and objective weights [12].

The biggest advantage of the fuzzy comprehensive evaluation method is that it can make a scientific, reasonable and realistic quantitative evaluation of the fuzzy evaluation object by using accurate numerical means. That is, the method does not characterize the object studied as either/or, but gives each object a membership degree to describe the object more accurately. In a word, fuzzy evaluation can not only describe the object more accurately, but also process the obtained information.

2.2 The Method of Random Forest Classification

In the field of machine learning, random forest is a classifier containing multiple decision trees. It contains many decision trees whose final outputs are determined by vote of these trees [13-14].

In 2010, Piero Bonissone et al. proposed a multiple classifier system based on a forest of fuzzy decision trees, that is, a fuzzy random forest that combines the robustness of multiple classifier systems, the power of the randomness to increase the diversity of the trees, and the flexibility of fuzzy logic and fuzzy sets for imperfect data management [15]. In 2013, Jose M. Cadenas et al. proposed a new method of feature selection that can handle both crisp and low quality data, and this approach is based on a fuzzy random forest and it integrates filter and wrapper methods into a sequential search procedure with improved classification accuracy of the features selected [16]. In 2019, Mohammed Ozigis et al. compared the Fuzzy Forest (FF) and Random Forest (RF) methods in detecting and mapping oil-impacted vegetation from a post spill multispectral sentinel 2 image and multi-frequency C and X Band Sentinel-1, COSMO Skymed and TanDEM-X images, and employed FF and RF classifiers to discriminate oil-spill impacted and oil-free vegetation in a study area [17].

Compared with the traditional classification method, the RF method has many advantages, such as fewer parameters to be adjusted, efficient processing of large sample data, no need to worry about overfitting, and a strong signal tolerance, which can effectively prevent the problem of sparse data in the decision tree.

In conclusion, this paper proposes an algorithm combining random forest classification and fuzzy comprehensive evaluation, which retains the information in the original data to the maximum extent and reduces the problem of information loss caused by more attributes and smaller weights. The results of fuzzy comprehensive evaluation are taken as the data that need to use random forest for training, which further improves the accuracy of the results by random forest training.

3 Proposed Method

In practice, however, data sets tend to be large, and the number of attributes that need to be trained by RF will also become huge, which will not only lead to a huge increase in the calculation, but also negatively affect the classification effect because the weight of each attribute and its correlation are not taken into account. Worse still, due to more attributes or ‘factors’, that as a term used in fuzzy theory, even if the weight distribution of each attribute or factor is clearly defined, in order to satisfy the normalization, the weight assigned must be small which is likely to cause information loss.

To solve these above problems, this paper proposes the method of RF classification based on fuzzy comprehensive evaluation, whose theoretical idea mainly is divided into three parts, as Figure 1 shows.

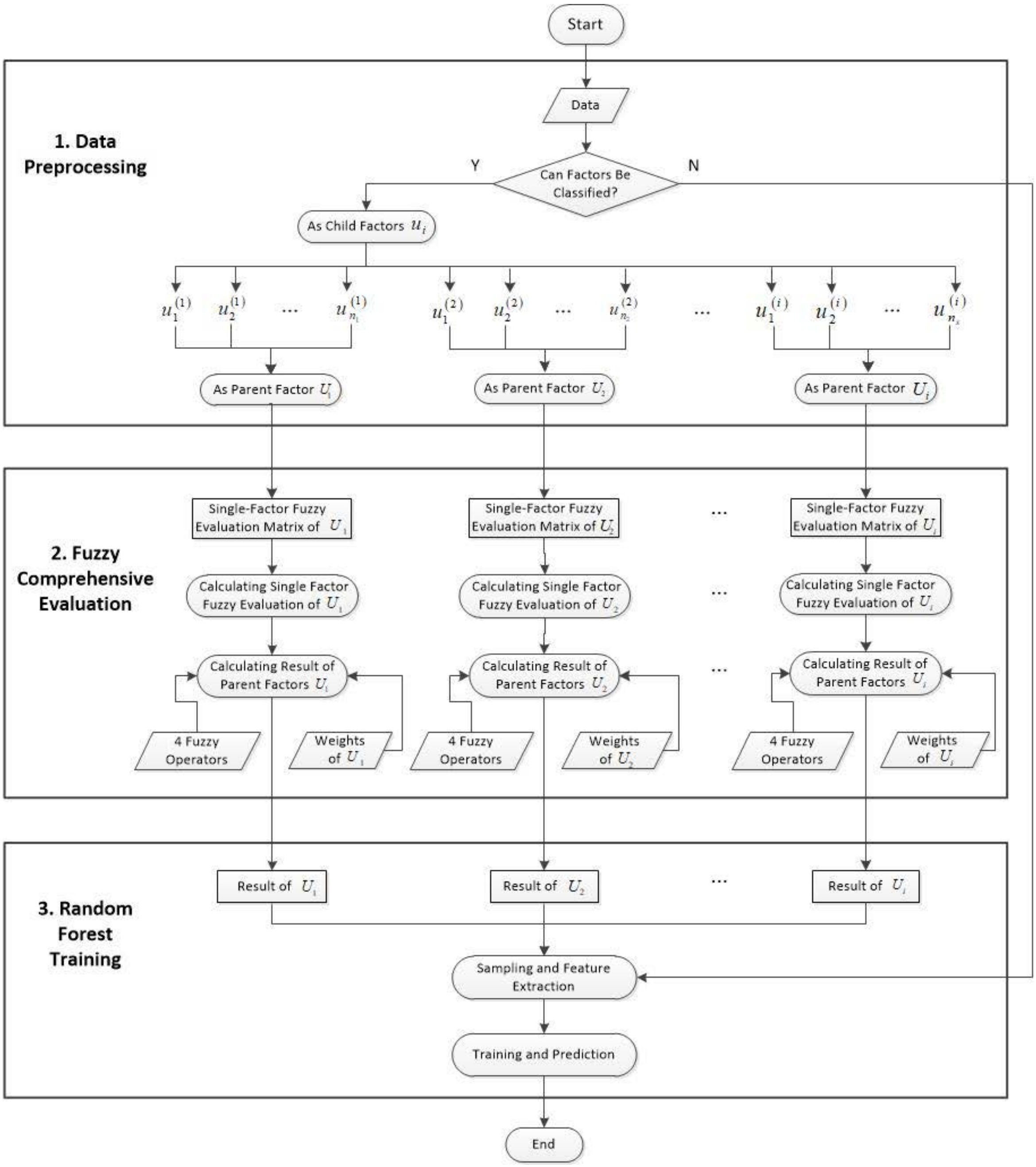


Figure 1. Flowchart of the improved method

3.1 Data Preprocessing

The first part is data preprocessing. The method of RF classification based on fuzzy comprehensive evaluation is for the data sets with large amount of data. These data sets have two characteristics: one is there are more attributes that need to be trained; the other is the value range of the attributes is not the Boolean data that can only take two values to represent yes or no, but can take multiple different values to represent multiple levels between the maximum and minimum values; that is, can take multiple values to represent different membership degrees.

It can be seen from the data preprocessing in Figure 1 that we first decide whether all the factors or attributes can be classified. Then, for those factors or attributes that can be classified, we carry out artificial discrimination, regarding the attributes with strong commonness as child-factors of a parent factor. In other word, we classify all the u_{i} that can be classified according to their specific meanings.

For example, Wikipedia's introduction to a country, such as Japan, includes the following attributes: 'Climate', 'Biodiversity', 'Environment', 'Agriculture and Fishery', 'Industry', 'Science and Technology', 'Art and Architecture', 'Etiquette', 'Cuisine' and so on. However, Wiki has already grouped these attributes: 'Climate', 'Biodiversity', 'Environment'; these factors are equivalent to$u_1^{(1)}$, $u_2^{(1)}$, and, $u_3^{(1)}$ in Figure 1, which describe the attribute 'Geography'; thus, we consider $u_1^{(1)}$, $u_2^{(1)}$, and $u_3^{(1)}$to be three child factors of U_1, i.e., the parent factor "Geography"; in the same way, 'Agriculture and Fishery', 'Industry', 'Science and Technology' , these factors are equivalent to $u_1^{(2)}$, $u_2^{(2)}$, and,

$u_3^{(2)}$ in Figure 1, which describe the attribute 'Economy'. So, we consider $u_1^{(2)}, u_2^{(2)}$, and $u_3^{(2)}$ to be three child factors of U_2, i.e., the parent factor "Economy"; finally, 'Art and Architecture', 'Etiquette', 'Cuisine', these factors are equivalent to $u_1^{(3)}$, $u_2^{(3)}$, and, $u_3^{(3)}$ in Figure 1, which describe the attribute 'Culture'; therefore, we consider $u_1^{(3)}, u_2^{(3)}$, and $u_3^{(3)}$ to be three child factors of U_3, i.e., the parent factor "Culture", as shown in the Figure 2.

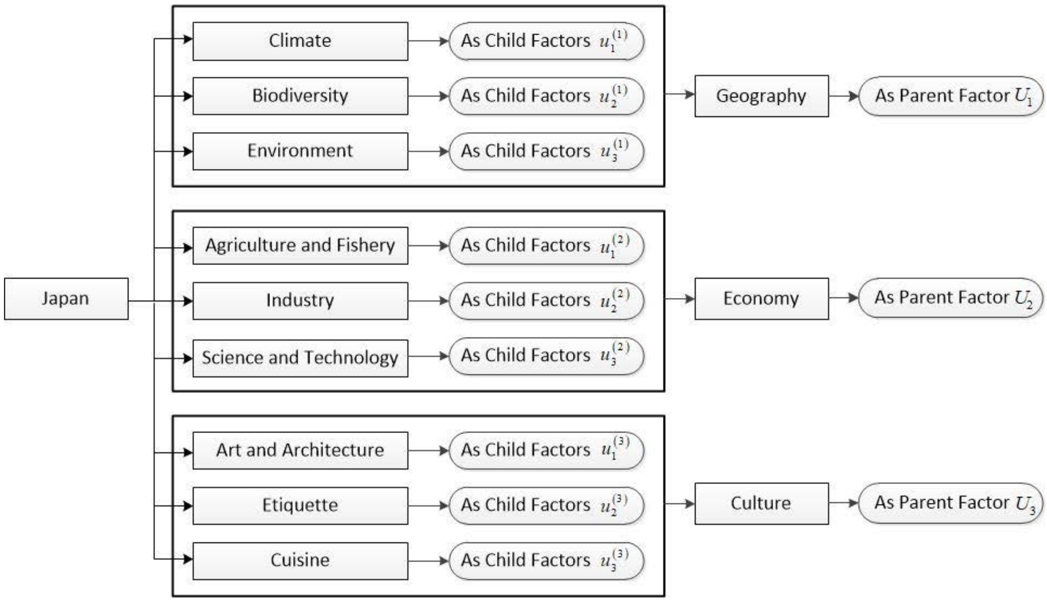


Figure 2. Classification of attributes of Japan

In a word, when there are many attributes and the value of each attribute is no longer limited to two limit values, the attributes with correlation are regarded as child factors under the same classification, which is usually based on the actual meaning. Then, we aggregate these common attributes, u_{n_i}, into a single parent-factor U_{i} and the final value of U_i is determined by these child factors. Therefore, after the first step of Data Preprocessing, we get the parent factor U_i.

As shown in Formula 1, $U_{\mathrm{i}}(i = 1, \cdots, n)$ is divided into s subsets according to the above rules, and satisfies the following conditions:

$$U_i = \left\{u_1^{(i)}, u_2^{(i)}, \cdots, u_t^{(i)}\right\}, t = 1,2,\cdots,s$$
$$\bigcup_{i=1}^{n} U_i = U \text{ ; } U_i \cap U_j = \emptyset, i \neq j. \quad \textbf{(1)}$$

3.2 Fuzzy Comprehensive Evaluation

The second part is fuzzy comprehensive evaluation. First of all, for all parent factors U_i obtained in data preprocessing, we regard each U_i as a single factor and calculate their single factor fuzzy evaluation matrixes. The reason is that only when we get the single factor fuzzy evaluation matrix of one parent factor, we can carry out the single factor fuzzy comprehensive evaluation based on its matrix.

Besides, in order to clarify the outcome of the fuzzy evaluation, in general, we will define the fuzzy evaluation set $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_\mathrm{n}\}$, which is a set of all kinds of overall evaluation results made by the evaluator to the objects that need to be evaluated. $\mathrm{v}_j(j = 1, 2, \cdots, n)$ represents the outcome of all possible judgments. .For example, v_1 means completely not to belong, while v_n means completely to belong, so, v_2, v_3 ,$\cdots$ and v_{n-1} , all represent the results of judgments between completely not to belong and completely to belong.

Second, we make a single-factor fuzzy evaluation, which refers to the evaluation of a factor to determine the membership of the object. Assuming that factor U_{i} is evaluated, and the membership degree of the element *j* in the evaluation set is r_{in}, then the result is denoted as $\mathrm{R_i} = \{\mathrm{r}_{i1}, \mathrm{r}_{i2}, \cdots, \mathrm{r}_{in}\}$.

If there are altogether *i* single factors, then each single factor set is calculated to form the single factor evaluation matrix R_{ij}, as shown in Formula 2. By this way, the fuzzy evaluation matrix R^{U_i} of all fuzzy factors $U_i(i = 1, 2, \cdots, s)$ can be obtained.

$$\mathrm{R}_{ij} = \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \cdots & \mathrm{r}_{1n} \\ \mathrm{r}_{21} & \mathrm{r}_{22} & \cdots & \mathrm{r}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{r}_{m1} & \mathrm{r}_{m2} & \cdots & r_{mn} \end{bmatrix} \quad \textbf{(2)}$$

Thirdly, we make fuzzy comprehensive evaluation of U_i. That is to say, to comprehensively consider the influence by calculating the results of each Parent Factor U_i and get the correct evaluation results. However, to this end, the weights of the each above object $\mathrm{a_k}$ and fuzzy operators should be determined in advance.

The most common method to determine the weight of $\mathrm{a_k}$ is with the Delphi method. The reason for choosing this method is that when the original amount of information is larger and there are more relevant factors involved in the decision making, it is not only expensive for a computer to process, but the cost-benefit ratio is not low either. As a result, from an efficiency point of view, the Delphi method is often a good choice [18].

In addition, there are 4 main types of fuzzy operators in common use: dominant factor types M(∧,∨) and M(•,∨), weighted average types M(∧,⊕) and M(•,⊕) [19]. The characteristics of these 4 fuzzy operators are shown in the Table 1.

Table 1. Characteristics of 4 fuzzy operators

Characteristics	Fuzzy operator			
	M(∧,∨)	M(●,∨)	M(∧,⊕)	M(●,⊕)
Role of weights	Unobvious	Obvious	Unobvious	Obvious
Comprehensive ability	Weak	Weak	Strong	Strong
Utilization of information	Insufficiency	Insufficiency	A little sufficient	Sufficient
Type	Dominant factor	Dominant factor	Weighted average	Weighted average

So, we calculate the result of these parent factor U_i with 4 fuzzy operators and their weights of each U_i.

Through these operations and calculations above, the big number of the original factors will be greatly reduced because the child factors with common meanings are aggregated together, which are regarded as the child factors of a parent factor. Now, we make a fuzzy comprehensive evaluation of these parent factors, and the result must contain the information about the original child factors in the original data sets. That is to say, although the amount used for random forest training has been greatly reduced, the information contained in the original dataset has not been lost. Therefore, the fuzzy evaluation results of these parent factors can replace the data in the original dataset for subsequent random forest training.

Finally, the fuzzy comprehensive evaluation is carried out to obtain the evaluation result B_U of the whole dataset, as shown in Formulas 3, 4, and 5.

$$B_i = (b_{i1}, b_{i2}, \cdots, b_{im}) = A_i \circ R_{ij} = (a_{i1}, a_{i2}, \cdots, a_{in_i}) \circ \begin{bmatrix} r_{i1}^1 & r_{i1}^2 & \cdots & r_{i1}^m \\ r_{i2}^1 & r_{i2}^2 & \cdots & r_{i2}^m \\ \vdots & \vdots & \ddots & \vdots \\ r_{in_i}^1 & r_{in_i}^2 & \cdots & r_{in_i}^m \end{bmatrix} \quad (3)$$

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_s \end{bmatrix} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{s1} & b_{s2} & \cdots & b_{sm} \end{pmatrix} \quad (4)$$

$$B_U = (b_1, b_2, \cdots b_m) = A \circ B = (a_1, a_2, \cdots, a_s) \circ \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{s1} & b_{s2} & \cdots & b_{sm} \end{bmatrix} \quad (5)$$

According to the principle of maximum membership, the evaluation grade v_j corresponding to the largest element b_j in the fuzzy evaluation set $B_U = (b_1, b_2, \cdots b_m)$ is taken to be the final comprehensive evaluation result.

3.3 Random Forest Training

The third part is Random Forest Training. Due to the evaluation result obtained in Fuzzy Comprehensive Evaluation, its representation is still consistent with a certain V_n value in the evaluation set V. Therefore, we can regard the fuzzy evaluation result of parent factor U_i obtained in the second step as the factors that need to carry out RF training. At the same time, all the factors that cannot be 'classified' in Data Preprocessing are also taken as the factors that need to carry out RF training. In this way, all the information in the original data set can be retained and can be used for sampling and feature extraction. And finally, we carry out training and prediction to get the final result.

Actually, when RF is used to process data with many instances and a large number of factors, if RF is directly used for training, although theoretically its prediction accuracy is higher than that of other methods, it is still difficult to solve the problem caused by the different weights of each factor as well as information loss. But, RF with fuzzy comprehensive evaluation method can solve this problem to the greatest extent. Therefore, the results of parent factors U_i, and factors that cannot be classified, are both jointly used as the training values for RF training.

4 Experiments and Result Analysis

In order to verify the accuracy of this method of RF based on fuzzy comprehensive evaluation, 21 data sets were carried out for experiments.

4.1 Preparation for the Experiment

In order to check the experimental effect more objectively and comprehensively. Three traditional methods, KNN, NB and DT, are used to compare with the RF method and the improved RF methods, which is based on fuzzy comprehensive evaluation.

KNN, in statistics, the K-nearest neighbors algorithm, is one of the most commonly used classification methods in data mining, which was first proposed by Cover and Hart in 1968, which is a relatively mature method in theory. The idea of this method is very simple and intuitive: if most of the K most similar samples in the feature space, that is, the closest neighbors in the feature space, belong to a certain category, then the sample also belongs to that category. This method only determines the category of the samples to be classified according to the category of the nearest one or several samples. KNN method is simple, easy to understand and implement, and does not need to estimate parameters. It is suitable for automatic classification of class fields with large sample size.

NB, that is the Naive Bayesian algorithm, is one of the most widely used classification algorithms, with a solid mathematical foundation and stable classification efficiency. This method is simplified on the basis of Bayesian algorithm, that is, when the target value is given, the conditions of the attributes are independent of each other, which greatly simplifies the complexity of Bayesian method in practical application. The advantage of NB method is that its logic is very simple and its algorithm is relatively stable. When the data presents different characteristics, the performance of naive Bayes classification will not be significantly different. Naive Bayes classification algorithm has better performance when the relationship between attributes of data set is relatively independent.

DT, that is a Decision Tree, is a kind of Decision analysis method which obtains the probability that, the expected value of net present value is greater than or equal to zero, by constructing Decision Tree on the basis of knowing the probability of occurrence of various situations. Decision Tree method, which can evaluate project risks and judge its feasibility, is a graphical method that intuitively applies probability analysis. The advantage of decision tree is that it is easy to understand and implement, can directly reflect the characteristics of data, and can understand the meaning expressed by decision tree as long as it is explained. At the same time, decision tree can make feasible and good results for large data sources in a relatively short time.

For the RF method and the improved RF methods，we collected 21 data sets.

The first data set is obtained after random interview and questionnaire conducted by our team on campus, which is about how much correlation there is between students' general evaluation in school and their off-campus family background and living environment. This data set has 382 valid instances and each instance has 9 factors. In these factors, 'Evaluation' is the target value, and the other ones can be regarded as the determining factors of 'Evaluation'. The number 0 to 4, or 1 to 4, or 1 to 5 indicate the values of all the determining factors.

The other 20 data sets are all from Kaggle data sets and UCI data sets. Each data set used in this paper is fully described on its web page, and their addresses are shown in the Table 2 below.

In the first data set, there are 382 students in a school, and each student has 13 different teachers, who evaluate various factors in the students' family background and living environment through the form of scores. That is to say, 13 different teachers conducted questionnaires with students to investigate the factors and make data statistics, whose results are recorded in the form of scores. The score ranged from 0 to 5, among which, 0 or 1 indicating lower or worse, and 4 or 5 indicating higher or better. Finally, a target factor in the data set is the student's general evaluation, with 0 representing unqualified and 1 representing qualified. Through the calculation of the data, we will examine whether the general evaluation of students is 0 or 1, and whether it can be predicted according to the students' family background and living environment.

Table 2. Source of the other 20 data sets

Data set	Address
2	https://www.kaggle.com/ronitf/heart-disease-uci
3	https://www.kaggle.com/unsdsn/world-happiness
4	https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
5	https://www.kaggle.com/START-UMD/gtd
6	https://www.kaggle.com/spscientist/students-performance-in-exams
7	https://www.kaggle.com/abcsds/pokemon
8	https://www.kaggle.com/mohansacharya/graduate-admissions
9	https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
10	https://www.kaggle.com/wsj/college-salaries
11	https://www.kaggle.com/shivam2503/diamonds
12	https://archive.ics.uci.edu/ml/datasets/Audiology+%28Standardized%29
13	https://archive.ics.uci.edu/ml/datasets/Australian+Sign+Language+signs+%28High+Quality%29
14	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
15	https://archive.ics.uci.edu/ml/datasets/Dresses_Attribute_Sales
16	https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
17	https://archive.ics.uci.edu/ml/datasets/Flags
18	https://archive.ics.uci.edu/ml/datasets/Horse+Colic
19	https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
20	https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29
21	https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

It should also be noted that, for sensitive reasons such as personal privacy, all the values of these factors are obtained through anonymous questionnaires on a voluntary basis. At the same time, the number of students is 382, which actually refers to the number of valid questionnaires finally obtained; that is, the number of valid questionnaires obtained by the students who voluntarily participated in the questionnaire is 382.

4.2 Acquisition of Child Factors and Parent Factor

The first data set has 9 factors for data preprocessing, namely "Medu", "Fedu", "Famrel", "Traveltime", "Studytime", "Freetime", "Goout", "Health" and "Evaluation". It should be noted that the values of these factors are not 0 or 1 representing "good" or "bad" respectively. Instead, the number 0 to 4, or 1 to 4, or 1 to 5 indicate that the value is between "good" and "bad"; in other words, it is a kind of fuzzy evaluation with different values. The specific description of each factor is shown in the Table 3.

In these factors, 'Evaluation' is the target value, and the other ones can be regarded as the determining factors of 'Evaluation'.

Table 3. Specific description of each factor

Name	Meaning	Rules of scoring
Medu	Mother's education	0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
Fedu	Father's education	
Famrel	Quality of family relationships	from 1 - very bad to 5 - excellent
Traveltime	Home to school travel time	1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour
Studytime	Weekly study time	1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - > 10 hours
Freetime	Free time after school	from 1 - very little to 5 - very much
Friendship	Going out with friends	from 1 - very few to 5 - very many
Health	Current health status	from 1 - very bad to 5 - very good
Evaluation	General evaluation of this course	0- unqualified, 1- qualified

By analyzing the actual meanings of these factors, it can be seen that the level of the mother's education, the level of the father's education, and the quality of his or her family relationship actually describe the student's 'Family Educational Environment' (FEE). So, we can consider the 'FEE' as a parent-factor U_1 and 'Medu', 'Fedu', 'Famrel' can be considered as the 3 child-factors that affect the value of U_1. According to the actual meaning of these factors, some of them are constructed into a two-layer factor set, as shown in the Table 4.

In the same way, we can find that 'Traveltime', 'Studytime' describe the students' 'Time Efficiency' (TE), and 'Freetime', 'Friendship', 'Health' describe the students' 'Physical and Mental State' (PMS). We can regard respectively 'TE' and 'PMS' as the two parent-factors U_2, U_3, and 'Traveltime', 'Studytime' can be considered as the two child factors u_1, u_2, which affect the value of U_2; and 'Freetime', 'Friendship', 'Health' can be considered as the three child factors u_1, u_2, u_3, which affect the value of U_3, as Table 5 shows.

Table 4. Two-Layer structure of U_1

The first layer		The second layer	
U_1	FEE	u_1	Medu
		u_2	Fedu
		u_3	Famrel

Table 5. Two-Layer structure of U_2, U_3

The first layer		The second layer	
U_2	TE	u_1	Traveltime
		u_2	Studytime
U_3	PMS	u_1	Freetime
		u_2	Goout
		u_3	Health

It is worth noting that in U_2, u_1 is rated in ascending order from less to more, but logically, the less time spent on commuting, the less negative the effect is on his or her total score. Hence, it is better to preprocess the data and rescore u_1 in descending order from less to more.

4.3 Acquisition of One Student's Fuzzy Relation Matrix R_i of U_i

Since the value range of u_1, u_2, u_3 in U_1 is $V = \{0, 1, 2, 3, 4\}$, the final R_1 should be a fuzzy relational matrix with 3 rows and 5 columns.

For U_1 of one student, 13 different teachers scored the student's three child-factors respectively, and the membership degree of each child-factor u_i belonging to a certain V_j is expressed as r_{ij}, as shown in Formula 6.

$$r_{ij} = \frac{\text{Teachers evaluated } u_i \text{ as the score of the number } j \text{ in the set V}}{n=13} \quad (6)$$

As defined above, R_1 is calculated as shown in Formula 7.

$$R_1 = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} \end{bmatrix} \quad (7)$$

Similarly, we can obtain the fuzzy relation matrices R_2 and R_3 of the second parent-factor U_2 and the third parent-factor U_3 respectively, as shown in Formulas 8 and 9.

$$R_2 = \begin{bmatrix} r'_{11} & r'_{12} & r'_{13} & r'_{14} \\ r'_{21} & r'_{22} & r'_{23} & r'_{24} \end{bmatrix} \quad (8)$$

$$R_3 = \begin{bmatrix} r''_{11} & r''_{12} & r''_{13} & r''_{14} & r''_{15} \\ r''_{21} & r''_{22} & r''_{23} & r''_{24} & r''_{25} \\ r''_{31} & r''_{32} & r''_{33} & r''_{34} & r''_{35} \end{bmatrix} \quad (9)$$

Also in the similar way, we can get the fuzzy relation matrices $R_1^t, R_2^t, R_3^t, t \in [1, 2, \cdots, 382]$ for all students.

4.4 Acquisition of the Weights a_k and Fuzzy Operators

It is necessary to assign different weights to each u_i because of its different influence.

Through Delphi methods, 30 teachers in the school were asked to assign weights to each factor respectively in the form of questionnaire for calculation and statistics. After statistics, the approximate values of acquired weights are shown in the Table 6.

Table 6. Values of weights

Values of weights	u_1	u_2	u_3
A_{U_1}	0.3	0.3	0.4
A_{U_2}	0.4	0.6	NULL
A_{U_3}	0.3	0.2	0.5

As to fuzzy operators, the 4 kinds of fuzzy operators are all used in the experiment for comparison.

4.5 Results of Experiment

The experimental results on the 21 data sets are shown in Table 7.

In Table 7, 'All-Attribute RF' refers to the original RF method itself, which treats all factors in the data set for training; the 'Accuracy' refers to the correct ratio after the comparing the predicted results from the above different methods, with the real evaluation values in the test set; in the column 'Accuracy Difference' of the table, it is the differences between the results of RF and those of all other methods.

At the same time, the methods KNN, NB, DT and the improve RF method, which is based on fuzzy comprehensive evaluation with 4 fuzzy operators, are adopted to carry out experiments on these 21 data sets respectively for comparison. The number of random forests is $N = \{100, 200, 300, 500\}$, and the depth of the forest can be $n = \{3, 5, 7, 9\}$. These two sets of parameters were cross-tested separately to find the best result in each case. We take the accuracy of RF as a benchmark to compare the experimental results of other methods.

After the above 8 methods were experimented on 21 data sets, their respective accuracy rates were compared, as shown in Figure 3. The average accuracy rate of each method was shown in Figure 4.

In order to make the accuracy of RF method compared more intuitively with that of the 4 improved RF methods, the results of these 5 methods are shown in Figure 5 below.

The accuracy difference between the 3 methods of RF with the operators M(●., ⊕), M(∧, ⊕), M(●, ∨) which have higher average accuracy than that of RF, and the average accuracy of RF itself, are shown in Figure 6.

4.6 Analysis of Results

It can be seen from the Table 7 and Figure 3 that for 21 data sets, the average accuracy of the methods of KNN, NB and DT is lower than that of the RF method. So, the accuracy difference of these methods are negative, among which, method KNN has the lowest accuracy. On the other hand, it can also be seen that there are 3 improved RF methods with fuzzy operators whose values are higher than that of RF method.

In Figure 4, it can be seen that the average accuracy of RF method without any fuzzy operator is 69.80%, and the average accuracy of KNN, NB and DT, is 55.10%, 59.60% and 64.46% respectively. That is to say, the experiment shows that the RF method is more accurate, so it has a wider application value in the actual process. Therefore, it is of more practical significance to improve the RF method.

However, from the values in Figure 5 and Figure 6, the RF method based on fuzzy comprehensive evaluation with 4 fuzzy operators has different accuracy. Among them, there are 3 that have higher accuracy than that of RF, which are the RF with M(●, ∨)、RF with M(∧, ⊕)、RF with M(●, ⊕), whose accuracy respectively are 79.43%, 71.03%, and 74.74%, while the accuracy of the RF with M(∧, ∨) is even slightly lower than that of RF.

When using fuzzy operator M(∧, ∨), the accuracy is 67.33%, which is lower than that of RF itself. The reason is that the accuracy of this fuzzy operator is much lower in the case of more factors and smaller weights. However, some literature proves that the low accuracy of this operator is due to the defects of its algorithm itself [20].

The accuracy of the result of using fuzzy operator M(●, ∨) is the highest, reaching 79.43%, which is 9.63% higher than that of RF. The result is good, but a lot of information is lost during the calculation [21]. Meanwhile, the information utilization ratio is lower, which can give rise to fake higher numerical accuracy if without sufficient data. However, it should be noted that an inflated accuracy is also a risk and does not have widespread portability [22]. Hence, this operator is more suitable for the datasets with more redundant data. The accuracy of the result of using fuzzy operator M(∧, ⊕) is 71.03%, which is slightly higher than that of RF. The reason is that the advantage of this operator is that it can make full use of all information, which means it can synthesize the values of all attributes or factors as far as possible. Therefore, after training the results calculated by this operator, the predicted value differs little from the original actual value, which only increases 1.23%. The disadvantage of this operator in calculation is that it fails to make full use of the different weights of attributes or factors, that is, the advantages of assigning different weights to different u_i are not reflected. Hence, this operator is more suitable for datasets with the same or similar weights of all attributes or factors.

The accuracy of the result of using the fuzzy operator M(●, ⊕) is 74.74%, which is 4.94% higher than that of the RF. Its advantages lie in the loss of information is less, and the integrated degree is higher; meanwhile, this operator can make full use of existing information to carry out fuzzy evaluation, and the influence of attributes or factors with different weights is also fully considered on the final result of fuzzy comprehensive evaluation.

Therefore, after using this operator to make fuzzy comprehensive decision and then using RF with this operator to make prediction, the results have higher accuracy, objectivity, good portability and universality, and this operator is more suitable for datasets with a large number of attributes or factors [23].

Table 7. Experimental results of 21 data sets

Results / Methods		Data Sets	Accuracy	N value	n value	Accuracy Difference	Data Sets	Accuracy	N value	n value	Accuracy Difference	Data Sets	Accuracy	N value	n value	Accuracy Difference
KNN		1	57.58%	/	/	-17.17%	8	56.15%	/	/	-16.22%	15	57.46%	/	/	-17.35%
Naive Bayes			62.63%	/	/	-12.12%		60.28%	/	/	-12.09%		68.54%	/	/	-6.27%
Decision Tree			63.67%	/	/	-11.08%		65.46%	/	/	-6.91%		71.90%	/	/	-2.91%
All-Attribute Random Forest			74.75%	100	3	0		72.37%	100	3	0		74.52%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		69.66%	100	5	-5.09%		70.05%	100	3	-2.32%		72.99%	100	3	-1.82%
	$M(\bullet,\vee)$		82.01%	100	3	7.26%		83.84%	200	3	11.47%		85.32%	100	7	10.51%
	$M(\wedge,\oplus)$		75.17%	100	3	0.42%		73.09%	100	3	0.72%		75.05%	100	3	0.24%
	$M(\bullet,\oplus)$		78.62%	100	7	3.87%		78.53%	100	3	6.16%		78.33%	100	3	3.52%
KNN		2	50.17%	/	/	-6.60%	9	52.77%	/	/	-16.47%	16	59.96%	/	/	-14.85%
Naive Bayes			53.06%	/	/	-3.71%		58.36%	/	/	-10.88%		63.84%	/	/	-10.97%
Decision Tree			54.51%	/	/	-2.26%		61.57%	/	/	-7.67%		70.33%	/	/	-4.48%
All-Attribute Random Forest			56.77%	100	5	0		69.24%	100	3	0		74.81%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		56.25%	100	3	-0.52%		68.11%	100	3	-1.13%		73.04%	100	3	-1.77%
	$M(\bullet,\vee)$		68.33%	500	9	11.56%		80.12%	100	7	10.88%		80.32%	100	7	5.51%
	$M(\wedge,\oplus)$		59.08%	200	3	2.31%		70.30%	100	3	1.06%		76.39%	100	3	1.58%
	$M(\bullet,\oplus)$		63.33%	100	3	6.56%		75.01%	100	3	5.77%		78.71%	100	3	3.90%
KNN		3	52.02%	/	/	-7.90%	10	58.24%	/	/	-14.91%	17	50.46%	/	/	-17.89%
Naive Bayes			54.33%	/	/	-5.59%		62.00%	/	/	-11.15%		57.58%	/	/	-10.77%
Decision Tree			55.89%	/	/	-4.03%		69.32%	/	/	-3.83%		62.67%	/	/	-5.68%
All-Attribute Random Forest			59.92%	100	3	0		73.15%	100	3	0		68.35%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		57.05%	100	3	-2.87%		70.56%	100	3	-2.59%		66.09%	100	3	-2.26%
	$M(\bullet,\vee)$		69.60%	200	5	9.68%		82.28%	300	5	9.13%		75.46%	300	7	7.11%
	$M(\wedge,\oplus)$		61.58%	100	3	1.66%		74.28%	100	3	1.13%		69.33%	100	3	0.98%
	$M(\bullet,\oplus)$		64.37%	100	3	4.45%		78.05%	100	3	4.90%		73.92%	100	3	5.57%
KNN		4	51.92%	/	/	-11.45%	11	53.66%	/	/	-13.17%	18	54.49%	/	/	-22.29%
Naive Bayes			56.83%	/	/	-6.54%		57.98%	/	/	-8.85%		59.02%	/	/	-17.76%
Decision Tree			59.21%	/	/	-4.16%		60.49%	/	/	-6.34%		73.54%	/	/	-3.24%
All-Attribute Random Forest			63.37%	100	3	0		66.83%	100	5	0		76.78%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		63.05%	100	3	-0.32%		65.11%	100	3	-1.72%		73.97%	100	3	-2.81%
	$M(\bullet,\vee)$		71.42%	100	7	8.05%		77.54%	100	7	10.71%		84.01%	300	7	7.23%
	$M(\wedge,\oplus)$		65.01%	100	3	1.64%		66.98%	100	3	0.15%		78.44%	100	3	1.66%
	$M(\bullet,\oplus)$		67.36%	100	3	3.99%		71.05%	100	3	4.22%		81.61%	100	3	4.83%
KNN		5	48.72%	/	/	-10.03%	12	50.93%	/	/	-12.31%	19	59.00%	/	/	-14.20%
Naive Bayes			50.33%	/	/	-8.42%		54.04%	/	/	-9.20%		64.89%	/	/	-8.31%
Decision Tree			54.91%	/	/	-3.84%		58.63%	/	/	-4.61%		66.02%	/	/	-7.18%
All-Attribute Random Forest			58.75%	100	3	0		63.24%	100	5	0		73.20%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		55.65%	100	3	-3.10%		60.82%	100	3	-2.42%		71.58%	100	3	-1.62%
	$M(\bullet,\vee)$		72.68%	300	5	13.93%		76.95%	200	7	13.71%		80.43%	500	5	7.23%
	$M(\wedge,\oplus)$		60.47%	200	3	1.72%		64.08%	200	3	0.84%		74.07%	200	3	0.87%
	$M(\bullet,\oplus)$		63.06%	100	3	4.31%		68.07%	100	3	4.83%		78.99%	100	3	5.79%
KNN		6	57.80%	/	/	-12.26%	13	55.48%	/	/	-15.29%	20	61.37%	/	/	-13.92%
Naive Bayes			62.63%	/	/	-7.43%		59.21%	/	/	-11.56%		66.59%	/	/	-8.70%
Decision Tree			67.48%	/	/	-2.58%		63.66%	/	/	-7.11%		70.18%	/	/	-5.11%
All-Attribute Random Forest			70.06%	100	5	0		70.77%	100	3	0		75.29%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		68.22%	100	3	-1.84%		67.56%	100	3	-3.21%		73.36%	100	3	-1.93%
	$M(\bullet,\vee)$		81.99%	100	3	11.93%		80.99%	500	7	10.22%		84.22%	500	5	8.93%
	$M(\wedge,\oplus)$		72.31%	100	3	2.25%		71.25%	300	3	0.48%		76.46%	200	3	1.17%
	$M(\bullet,\oplus)$		75.05%	100	3	4.99%		76.46%	100	3	5.69%		80.03%	100	3	4.74%
KNN		7	53.95%	/	/	-14.92%	14	59.22%	/	/	-17.39%	21	55.67%	/	/	-22.39%
Naive Bayes			57.13%	/	/	-11.74%		63.70%	/	/	-12.91%		59.64%	/	/	-18.42%
Decision Tree			61.15%	/	/	-7.72%		69.58%	/	/	-7.03%		73.48%	/	/	-4.58%
All-Attribute Random Forest			68.87%	100	5	0		76.61%	100	3	0		78.06%	100	3	0
Random Forest with Each Fuzzy Operator	$M(\wedge,\vee)$		64.26%	100	3	-4.61%		72.45%	100	3	-4.16%		74.20%	100	3	-3.86%
	$M(\bullet,\vee)$		76.46%	100	7	7.59%		89.09%	200	7	12.48%		84.98%	500	9	6.92%
	$M(\wedge,\oplus)$		70.57%	100	3	1.70%		78.43%	100	3	1.82%		79.39%	200	3	1.33%
	$M(\bullet,\oplus)$		74.35%	100	3	5.48%		82.21%	100	3	5.60%		82.35%	100	3	4.29%

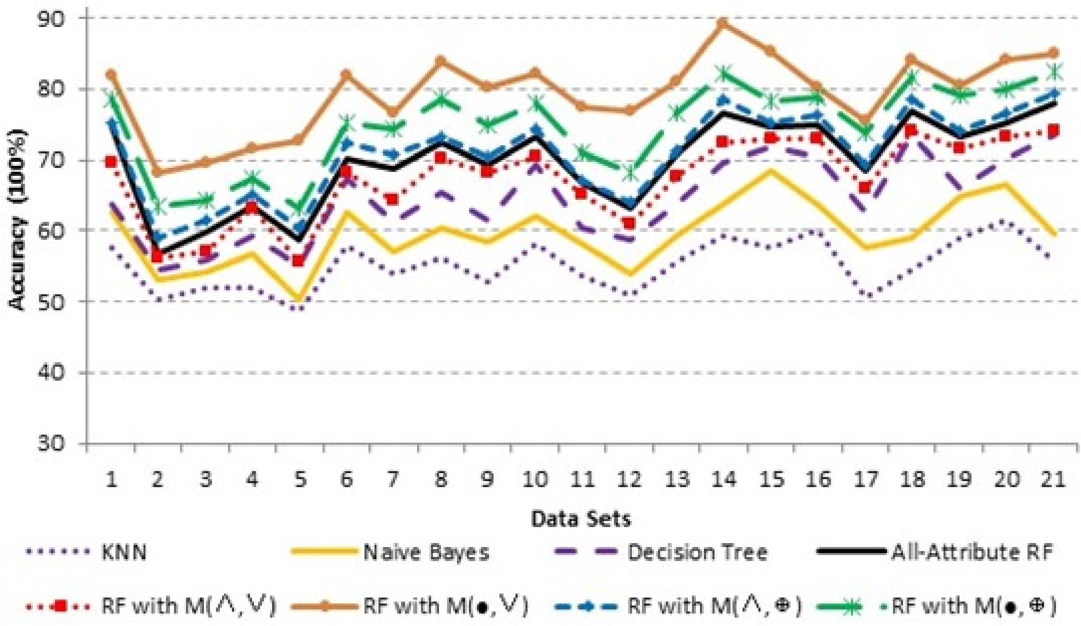


Figure 3. Respective accuracy rate of 8 methods

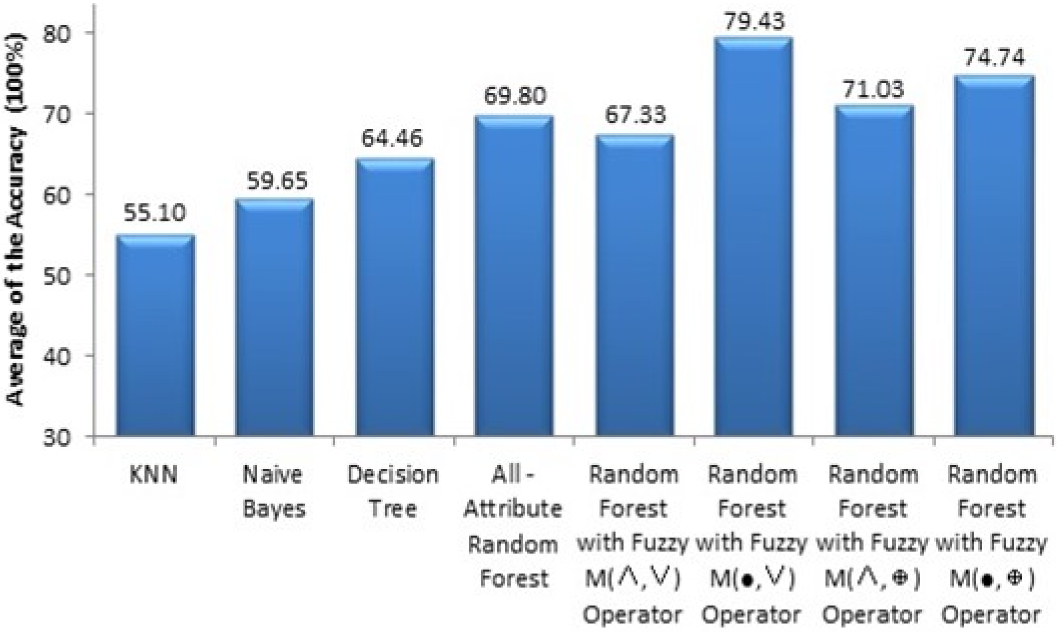


Figure 4. Average accuracy rate of 8 methods

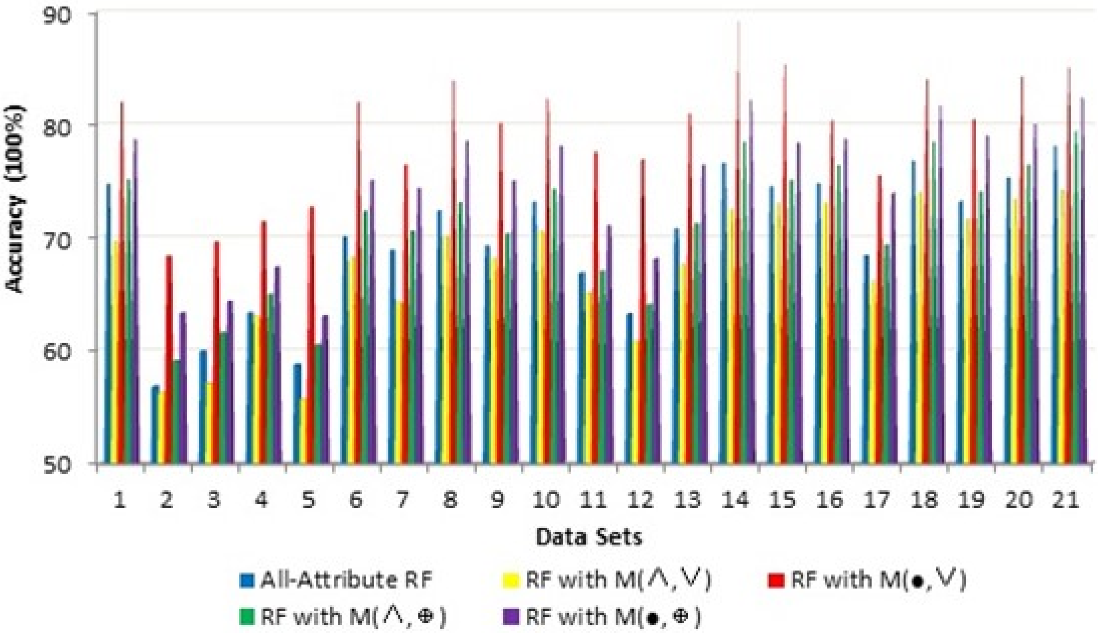


Figure 5. Intuitive comparison between RF method and approved RF methods

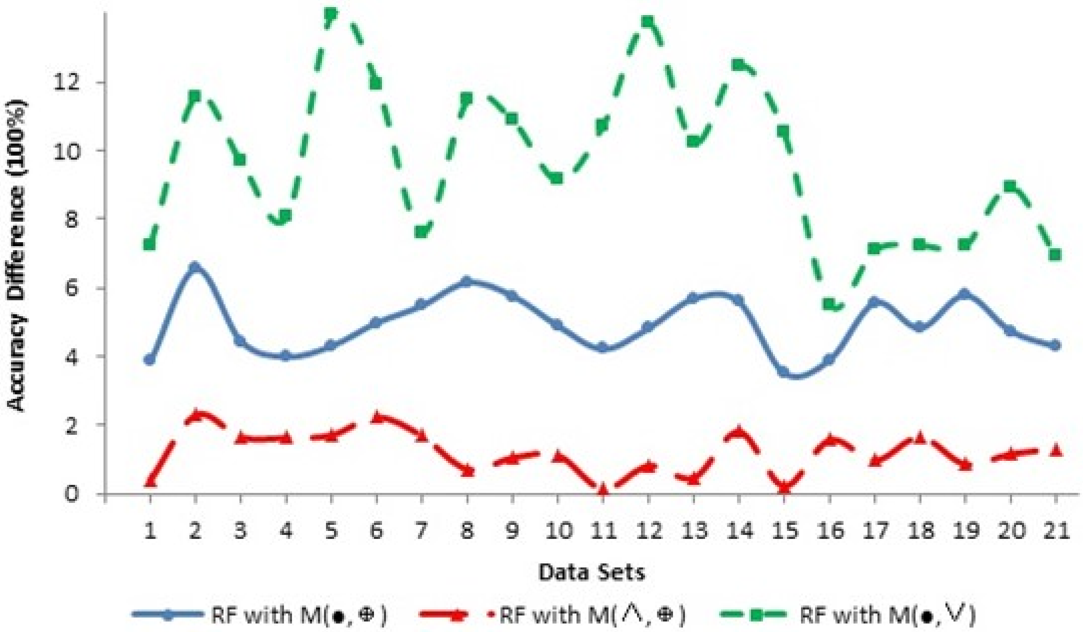


Figure 6. Accuracy difference between RF method and improved RF methods

5 Conclusion

This paper presents an algorithm combining random forest classification and fuzzy comprehensive evaluation, which can avoid the neglect of the correlation between attributes or factors, and also solves the problem of information loss caused by smaller weights due to more attributes in the original data set. In addition, the final random forest training results of this method has been further improved.

The experimental results show that compared to the training results of the all-attribute RF classification, 4 different fuzzy operators are used, and the accuracy of 3 of them are improved, as shown in the Table 8.

In Table 8, Growth in Accuracy refers to the comparison between the Accuracy of the three methods whose improved effect is better than the original RF algorithm and the original RF algorithm itself. Suitable Range refers to the Range that these three new and improved methods are more suitable for use than the original RF.

Table 8. Comparison of three improved RF methods with better results

Methods	Average accuracy rate	Growth in accuracy	Suitable range
RF	69.80%	0	Original range
RF with M(● ,∨)	79.43%	9.63%	Datasets with more redundant data
RF with M(∧ ,⊕)	71.03%	1.23%	Datasets with the same or similar weights of all attributes or factors
RF with M(● ,⊕)	74.74%	4.94%	Datasets with a large number of attribute or factors

From the Table 8, we can see that the accuracy of the result of using fuzzy operator M(●, ∨) is the highest, reaching 79.43%, which is 9.63% higher than that of RF. This operator is more suitable for the datasets with more redundant data. The accuracy of the result of using fuzzy operator M(∧ , ⊕) is 71.03%, which is slightly higher than that of RF. This operator is more suitable for datasets with the same or similar weights of all attributes or factors.

The RF method with the fuzzy operator M(●, ⊕), whose advantages lie in comprehensive data retention and strong ability to integrate information. The result accuracy of the improved RF method with the fuzzy operator M(●, ⊕) is 77.74%, 4.94% higher than that of RF method. By using RF with this operator to make prediction, the results have higher accuracy, objectivity, good portability and universality, and this operator is more suitable for datasets with a large number of attributes or factors.

Therefore, the algorithm combining random forest classification and fuzzy comprehensive evaluation proposed in this paper has a very broad application prospect. For example, by using this method, we could offer data projections for marketing purposes, such as the items that are expected to be purchased by customers based on their online behavioral history. In conclusion, this algorithm is worthy of further research and expansion for larger scale data sets with more attributes.

Acknowledge

The work reported here was supported by the 'National Key R&D Program of China' (No. 2018YFB1403400), Natural Science Foundation of China (No. 61702544), Natural Science Foundation of Jiangsu Province, China (No. BK20160769, BK20141072), and China Postdoctoral Science Foundation (No. 2016M603031).

References

[1] T. Wang, P. Chen, T. Bao, J. Li, X. Yu, Arrhythmia Classification Algorithm based on SMOTE and Feature Selection, *International Journal of Performability Engineering*, Vol. 17, No. 3, pp. 263-275, March, 2021.

[2] D. Li, W. E. Wong, W. Wang, Y. Yao, M. Chau, Detection and Mitigation of Label-flipping Attacks in Federated Learning Systems with KPCA and K-Means, *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, Yinchuan, Ningxia, China, September, 2021, pp. 551-559.

[3] V. Kulkani, P. Sinha, Pruning of Random Forest classifiers: A Survey and Future Directions, *2012 International Conference on Data Science & Engineering (ICDSE)*, Cochin, India, 2012, pp. 64-68.

[4] Advantages and disadvantages of random forest, https://blog.csdn.net/keepreder/article/details/47273297, 2015.

[5] J. Zhu, S. Huang, Y. Shi, K. Wu, Y. Wang, A Method of Random Forest Classification Based on Fuzzy Comprehensive Evaluation, *IEEE 8th International Conference on Dependable Systems and Their Applications (DSA)*, Yinchuan, Ningxia, China, 2021, pp. 178-183.

[6] Random Forest, https://en.wikipedia.org/wiki/Random_forest.

[7] Z.-G. Wu, S. Dong, P. Shi, D. Zhang, T. Huang, Reliable Filter Design of Takagi-Sugeno Fuzzy Switched Systems With Imprecise Modes, *IEEE Transactions on Cybernetics*, Vol. 50, No. 5, pp. 1941-1951, May, 2020.

[8] T. Hong, C. Lin, J. Lin, S. Wang, Learning from Hierarchical Attribute Values by Rough Sets, *Intelligent Systems Design and Applications, Advances in Soft Computing*, Tulsa, USA, 2003, pp. 559-567.

[9] T. Hong, Y. Liou, S. Wang, Fuzzy rough sets with hierarchical quantitative attributes, *Expert Systems with Applications*, Vol. 36, No. 3, pp. 6790-6799, April, 2009.

[10] W. Ma, Y. Wang, Fuzzy comprehensive evaluation model of interuniversity collaborative learning based on network, *Open Physics*, Vol. 15, No. 1, pp. 427-432, June, 2017.

[11] Y. Zhu, X. Wang, S. Deng, M. Zhao, X. Ao, Evaluation of Curtain Grouting Efficiency by Cloud Model – based Fuzzy Comprehensive Evaluation Method, *KSCE Journal of Civil Engineering*, Vol. 23, No. 7, pp. 2852-2866, July, 2019

[12] X. Wu, F. Hu, Analysis of Ecological Carrying Capacity using a Fuzzy Comprehensive Evaluation Method, *Ecological Indicators*, Vol. 113, pp.106243, June, 2020.

[13] A. Subudhi, M. Dash, S. Sabut, Automated Segmentation and Classification of Brain Stroke using Expectation-maximization and Random Forest Classifier, *Biocybernetics and Biomedical Engineering*, Vol. 40, No. 1, pp. 277-289, January-March, 2020.

[14] D. L. Manjushree, T. L. Varsha, W. K. Arvind, D. N. Laxman, Performance Analysis of the Impact of Technical Skills on Employability, *International Journal of Performability Engineering*, Vol. 17, No. 4, pp. 371-378, April, 2021.

[15] P. Bonissone, J. Cadenas, M. Garrido, R. Diaz-Valladares, A Fuzzy Random Forest, *International*

Journal of Approximate Reasoning, Vol. 51, No. 7, pp. 729-747, September, 2010.

[16] J. Cadenas, M. Garrido, R. Martinez, Feature Subset Selection Filter–wrapper based on Low Quality Data, *Expert Systems with Applications*, Vol. 40, No. 16, pp. 6241-6252, November, 2013.

[17] M. Ozigis, J. Kaduk, C. Jarvis, Mapping Terrestrial Oil Spill Impact using Machine Learning Random Forest and Landsat 8 Oli Imagery: a Case Site within the Niger Delta Region of Nigeria, *Environmental Science and Pollution Research*, Vol. 26, No. 4, pp. 3621-3635, February, 2019.

[18] M. Banno, Y.Tsujimoto, Y. Kataoka, The Majority of Reporting Guidelines are not Developed with the Delphi method: A Systematic Review of Reporting Guidelines, *Journal of Clinical Epidemiology*, Vol. 124, pp. 50-57, August, 2020.

[19] K. Wang, F. Shi, Fuzzifying Interval Operators, Fuzzifying Convex Structures and Fuzzy Pre-orders, *Fuzzy Sets and Systems*, Vol. 390, pp. 74-95, July, 2020.

[20] B. Pang, J. Mi, Z. Xiu, L-fuzzifying Approximation Operators in Fuzzy Rough Sets, *Information Sciences*, Vol. 480, pp. 14-33, April, 2019.

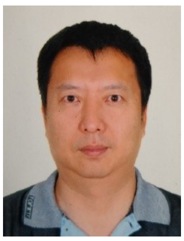
[21] L. R. O. Castillo, J. Soria, P. Melin, F. Valdez, C. Gonzalez, G. Martinez, J. Soto, A Fuzzy Hierarchical Operator in the Grey Wolf Optimizer Algorithm, *Applied Soft Computing*, Vol. 57, pp. 315-328, August, 2017.

[22] C. Wu, C. Lai, Data-driven Diversity Antenna Selection for MIMO Communication using Machine Learning, *Journal of Internet Technology*, Vol. 23, No. 1, pp. 1-9, January, 2022.

[23] N. Gao, Z. Zhu, Z. Weng, G. Chen, M. Zhang, A Supervised Named Entity Recognition Method Based on Pattern Matching and Semantic Verification, *Journal of Internet Technology*, Vol. 21, No. 7, pp. 1917-1928, December, 2020.

Biographies

Jing Zhu received the B.E. degree from Jiangsu University of Science and Technology in 2006, and the M.E. degree from Navy Command College in 2011. He is a Ph.D candidate in Software Engineering at Army Engineering University of PLA.



Song Huang received the B.S., M.S. and Ph.D degrees from the PLA University of Science and Technology. He is a professor of software engineering at Software Testing and Evaluation Center from Army Engineering University of PLA.

Yaqing Shi received the B.S. degree from Xi'an Technological University in 2004, and the M.S. and the Ph.D. degrees from Hehai University in 2007 and 2015, respectively. She is an associate professor in Army Engineering University of PLA.

Kaishun Wu received the B.S. degree from Zhejiang University in 2015, and the M.S. degree from the Army Engineering University of PLA in 2018. He has been a Ph.D. candidate in Software Engineering at Army Engineering University of PLA since 2018.

Yanqiu Wang received the B.E. degree in 1982. She had been a lecturer in the institute before retiring, teaching Information Fusion, C++, Machine Learning, and she also has been a private teacher who teaches relevant courses in another private institution.