An Online Integrated Classification Algorithm for Innovation and Entrepreneurship Teaching Data Based on Decision Tree

Juanjuan Zou(✉)

Chongqing Vocational Institute of Engineering, Jiangjin 402260, China
17726637816@163.com

Abstract. In order to improve the accuracy of data classification, reduce misclassification rates, and improve classification efficiency, a decision tree based online integrated classification algorithm for innovation and entrepreneurship teaching data is proposed. Establish a prototype system for online integration of innovation and entrepreneurship teaching data, and based on the results of data integration, preliminarily construct a decision tree model. Then, the fuzzy decision tree obtained by combining fuzzy theory with decision tree is used to solve the problems of data imbalance and missing data types. In order to further reduce the misclassification rate of data, the differential grey wolf optimization algorithm is used to optimize the decision tree, and the decision tree is improved through operations such as feature selection and decision tree pruning to obtain the optimal classification results of innovation and entrepreneurship teaching data. The experimental results show that the proposed method has high data classification accuracy, low misclassification rate, and high classification efficiency, which verifies the effectiveness of the method.

Keywords: Decision Tree · Innovation and Entrepreneurship · Data Classification · Differential Grey Wolf Optimization · Pruning Decision Tree

1 Introduction

With the continuous development of information technology and the Internet, people's demand for data is becoming increasingly urgent, especially in the field of innovation and entrepreneurship education. Data analysis and application have become important teaching content [1]. The online integration and classification of data are key technologies for achieving data sharing and utilization. Therefore, the background significance of studying the online integration and classification of innovation and entrepreneurship teaching data is to improve data utilization efficiency and accuracy, provide more comprehensive and accurate data support for students [2, 3]. However, due to the presence of a large number of redundant and irrelevant features in the dataset, this can affect the accuracy of classification algorithms and make the classification results more complex and inexplicable. Therefore, for each classification problem, it is necessary to correctly select features, eliminate redundant information, and improve data availability.

L. Yun et al. (Eds.): ADHIP 2023, LNICST 547, pp. 366–378, 2024.
https://doi.org/10.1007/978-3-031-50543-0_25

In the above context, relevant scholars have proposed some data classification methods, among which reference [4] proposes an imbalanced data classification algorithm based on undersampling and cost sensitivity. Firstly, before each iteration of the AdaBoost algorithm to train the base classifier, the majority class samples are sorted by weight from largest to smallest; Then normalize the weight of the sampled majority class samples and form a temporary training set with the minority class samples to train the base classifier; Secondly, in the weight update stage, a higher misclassification cost is assigned to minority classes, resulting in faster weight increase for minority class samples and slower weight increase for majority class samples, thus achieving data classification. The experimental results show that there is a problem of low data classification efficiency. Reference [5] proposed an unbalanced data classification method based on probability threshold Bagging algorithm, which combined threshold mobile technology technology with Bagging integrated algorithm, used the original distributed training set for training in the training phase, complete data classification. The experimental results show that this method has good classification advantages in handling imbalanced data, but there is a problem of inaccurate data classification when the data volume is large. Reference [6] proposes a data classification method based on the improved ID3 algorithm. This classification method determines the equilibrium coefficient through the modified information gain, optimizes the information gain obtained by the ID3 algorithm using the equilibrium coefficient, obtains the root node of the decision tree based on the optimized information gain, divides the nodes, classifies the attributes, constructs the decision tree, and achieves data classification through the decision tree. Through examples, it has been proven that this classification method can achieve control over multivalued bias and avoid selecting attributes with more values as branch nodes, but there is a problem of high misclassification rate. Reference [7] proposed an unbalanced data classification method based on the cost sensitive Activation function XGBoost, and introduced the cost sensitive Activation function to change the gradient change of the Loss function of samples under different prediction results, to solve the problem that the misclassified minority samples cannot be effectively classified in the XGBoost iteration process due to the small gradient change. The results show that the algorithm has a high detection rate for minority class samples, indicating that its classification effect for minority class samples is good. However, when the data volume is large, the accuracy of data classification is not high.

In order to solve the problems of low accuracy, low efficiency, and high misclassification rate in data classification of the existing methods mentioned above, a decision tree based online integrated classification algorithm for innovation and entrepreneurship teaching data is proposed. The main research content of this article's method is as follows:

(1) Establish a prototype system for online integration of innovation and entrepreneurship teaching data, improving data storage and access efficiency.
(2) Based on the online integration results of innovation and entrepreneurship teaching data, a decision tree model is constructed, and a decision tree based on differential grey wolf optimization is adopted. The data category weight modification strategy is used to reduce the weight of unimportant data and increase the weight of important data, in order to obtain the optimal result of data classification.

(3) Using data classification accuracy, efficiency, and misclassification rate as experimental indicators, compare the proposed method with traditional methods and draw relevant conclusions.

2 Online Integration of Innovation and Entrepreneurship Teaching Data

Before classifying innovation and entrepreneurship teaching data, the first step is to integrate the data, which aims to improve the coverage of the data. Integrating data from different sources can cover more comprehensive data and help improve the practical application value of data classification results. This article will use open-source software to study the online integration method of innovation and entrepreneurship teaching data, and design a prototype system to achieve online integration of innovation and entrepreneurship teaching data.

The prototype system adopts a browser/server (B/S) mode, with Windows as the operating system, and uses Ajax technology to develop web front-end interactive programs. The server side programs are developed using Apache network servers and PHP language. The framework structure of the system is shown in Fig. 1.

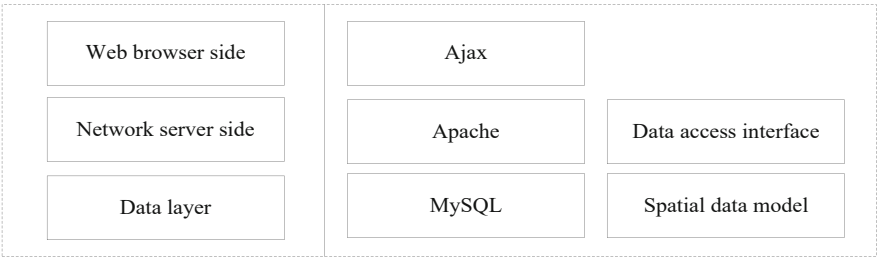

Fig. 1. Schematic diagram of the prototype system framework structure

As shown in Fig. 1, the system consists of three parts: a web front-end, a network server, and a database. The web front-end is responsible for collecting data, submitting the collected data to the server, and visualizing the data from the server. After receiving the data uploaded by the web front-end, the network server performs formatting on the data. The database is responsible for storing the collected data and providing data access services.

The prototype system is divided into four modules according to its functions: interaction, data collection, data storage, and data access, each with a detailed functional design. Among them, the interaction and data collection functions are mainly implemented on the web front-end. The functional structure design of the system is shown in Fig. 2.

As shown in Fig. 2, users can easily collect data through a browser, and the collected data content is completely customized by the user, enhancing the interactivity of the data and enabling ordinary users to transform from data users to data providers, enriching the sources of data. The use of free open source software greatly reduces the cost of

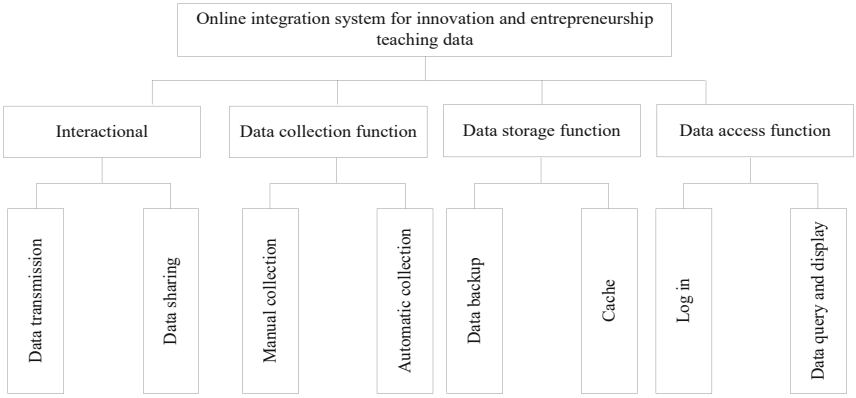

Fig. 2. Functional structure diagram of the prototype system

data collection and management. In addition, a MYSQL based database has been built to improve the efficiency of data storage and access. Provided a data access interface, enabling the collected data to be more widely used, thereby achieving online integration of innovation and entrepreneurship teaching data.

3 Classification of Innovation and Entrepreneurship Teaching Data

Based on the online integration results of innovation and entrepreneurship teaching data, conduct research on the classification of innovation and entrepreneurship teaching data.

3.1 Decision Tree Principle

Decision tree is a commonly used machine learning algorithm for classifying and predicting data. Its principle is to classify data through a series of branch conditions [8]. The decision tree can be seen as dividing some regions in the data sample space, with each region corresponding to a category. This division process is the process of constructing the decision tree.

The process of constructing a decision tree is mainly divided into three steps: selecting attributes, partitioning data, and constructing the tree.

(1) Selecting attributes: At each node of the decision tree, it is necessary to select an optimal attribute as the partitioning criterion to minimize the differences between different categories of data in the current node.
(2) Partition data: After determining the partition attributes, the data needs to be divided into different subsets based on the values of the attributes. This process generally involves determining multiple attributes, which are equivalent to the question of 'yes'. Continue to select attributes for partitioning in each subset until all data in the subset belongs to the same category.

(3) Building a tree: Based on the partitioning results, construct a tree structure to represent the classification model. Decision trees can be constructed and optimized using algorithms such as depth first, breadth first, and pruning to achieve better performance and generalization ability.

Decision trees are widely used in data classification, such as financial fraud detection, disease diagnosis, sentiment analysis, etc. Due to the simple and easy to understand process of constructing decision trees, clear classification rules can often be generated, making it easier to analyze and interpret data.

3.2 Construction of Decision Tree Model

According to the decision tree principle, when conducting online integrated classification of innovation and entrepreneurship teaching data, it is necessary to consider the problem from a global perspective, find the optimal data content, and improve the accuracy and efficiency of data classification. The decision tree algorithm is an important part of the current development process of artificial intelligence technology. It can comprehensively mine sample data without rules and orders, form the corresponding mathematical analysis model, obtain the most basic data classification rules, and then predict and classify various data sets. In the framework of blockchain, a diversified decision tree model can be constructed based on the distribution characteristics and actual situation of each node to synchronize and efficiently classify data [9]. To ensure the effectiveness of constructing the decision tree model, it is necessary to first generate candidate splitting points, specify the best data splitting time in each node, and then treat the best data classification points as the core part to achieve the purpose of online integrated classification of innovation and entrepreneurship teaching data. Finally, the model is iteratively updated to optimize its performance. In this process, node splitting is the basis for information gain rate analysis and information entropy processing in the online integration and classification of innovation and entrepreneurship teaching data, so it is necessary to focus on node splitting methods, as shown in Fig. 3.

In the overall practical operation process, the number of innovation and entrepreneurship teaching data and training samples waiting for classification should be set to N, consistent with the number of nodes in the blockchain. The overall number of training set types should be Z, labeled as $N_1, N_2, \ldots, N_z$. Assuming that any attribute of the training set is R, this attribute involves attribute data values, labeled as $N_1, N_2, \ldots, N_z$, and the number of samples that correspond to attribute data values R_1 is N_{R_i}, under the framework of blockchain, the information entropy of the training sample set in the innovation and entrepreneurship teaching data set can be calculated according to formula (1):

$$D(W) = \sum_{i=1}^{Z} t_i \log_2 t_i \tag{1}$$

Add attribute R to the feature data set and calculate according to formula (2):

$$D(R, W) = \sum_{i=1}^{Z} \frac{N_{R_i}}{N_i^R} \log_2 \frac{N_{R_i}}{N_i^R} \tag{2}$$

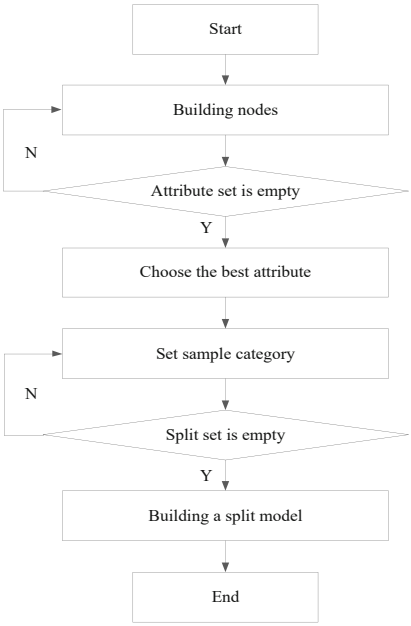

Fig. 3. Schematic diagram of data node splitting

After completing the above operation, use formula (3) to represent the information gain rate:

$$E(R, W) = \frac{\Delta(R, W)}{\sum_{i=1}^{Z} \frac{N_{R_i}}{N_i^R} \log_2 \frac{N_{R_i}}{N_i^R}} \tag{3}$$

In essence, scientific and reasonable selection will directly affect the application efficiency of the decision tree model. If there is no information entropy attribute or information gain efficiency attribute in the process of building the model, it will lead to classification errors. Therefore, in practical work, it is necessary to ensure the integrity of information entropy and information classification attributes of the overall decision tree model, and scientifically optimize and process various decision tree models, Prevent misclassification from occurring at the fundamental level. Meanwhile, due to the possibility of local node splitting and obstruction during the application process of the model, in order to avoid affecting the normal data splitting of other nodes, it is necessary to focus on optimizing the decision tree algorithm technology. Within the overall framework of the blockchain, buffer empty nodes should be pre-set, and the number of node attributes used for data splitting should be set to m. During the optimization process, $m + 1$ sub node should be set, Once there is a problem of insufficient attributes in the model, this sub node can split the data into empty nodes based on the currently known attributes to avoid classification interruption.

3.3 Decision Tree Based on Differential Grey Wolf Optimization

In the data classification problem of innovation and entrepreneurship teaching, traditional classification algorithms cannot handle the uncertain and imprecise information in attributes well, and fuzzy theory can handle this to some extent. Therefore, the fuzzy decision tree obtained by combining fuzzy theory with decision trees has better applicability. When encountering two problems in innovation and entrepreneurship teaching data: data imbalance and missing data types, it can lead to model misjudgment and reduced classification accuracy. To solve this problem, a data category weight modification strategy is adopted, which reduces the weight of unimportant data and increases the weight of important data. The change in weight is based on the best choice of differential grey wolf optimization [10].

3.3.1 Feature Selection

On the basis of splitting the decision tree results, a decision tree is further constructed using the differential grey wolf optimization algorithm for the classification of innovation and entrepreneurship teaching data. The construction of a decision tree is usually divided into three processes: feature selection, decision tree generation, and decision tree pruning. Feature selection is an important step before constructing a decision tree. If features are randomly selected, the learning efficiency of the established decision tree is very low. The generation of a decision tree starts with an empty tree, where the samples to be classified enter from the root of the tree. At each node of the tree, different paths are selected to gradually descend to the bottom by judging a certain attribute of the samples, and their category is determined. To prevent overfitting of the decision tree, an attempt is made to eliminate redundant nodes after constructing the entire tree, which is called pruning [11]. The pruning process involves checking nodes with the same parent node to determine whether the information gain after merging them will be less than a specified value. If so, merge these nodes.

The usual criteria for feature selection are based on information gain theory, which provides criteria for selecting candidate attribute lists between each decision node. Select the feature with the highest information gain by calculating the information gain of each feature:

$$G(Z, R_t) = P(Z) - P_{R_t}(Z) \tag{4}$$

In the formula, R_t represents the sample features; $P(Z)$ represents the entropy of the feature; $P_{R_t}(Z)$ represents the information gain rate, and the expression for both is as follows:

$$P(Z) = \sum_{j=1}^{m} \frac{m\left(F_j, Z\right)}{|Z|} \log_2 \frac{m\left(F_j, Z\right)}{|Z|} \tag{5}$$

$$P_{R_t}(Z) = \sum_{h \in C(R_t)}^{m} \frac{Z_h^{R_t}}{|Z|} P\left(Z_h^{R_t}\right) \tag{6}$$

In the formula, $m(F_j, Z)$ represents the number of objects belonging to class F_j in the training set; $C(R_t)$ represents the Finite field of feature R_t; $Z_h^{R_t}$ represents the cardinality of the set of objects with a value of h for feature R_t. The information gain ratio refers to the ratio of information gain to the entropy of the value of feature R_t in the training dataset Z, which can be expressed as:

$$V(Z, R_t) = \frac{G(Z, R_t)}{Q(Z, R_t)} \tag{7}$$

In the formula, $Q(Z, R_t)$ represents the potential information generated by dividing P into k subsets, and the mathematical definition equation is:

$$Q(Z, R_t) = \sum_{h \in C(R_t)}^{k} \frac{\left|Z_h^{R_t}\right|}{|Z|} \log_2 \frac{\left|Z_h^{R_t}\right|}{|Z|} \tag{8}$$

3.3.2 Decision Tree Generation

The generation of a decision tree starts from the root node, calculates the information gain of all features at the node, selects the feature with the highest information gain as the node feature, and establishes sub nodes based on different values of the feature. For sub nodes, the above method is used for recursion. When the information gain of all features is small or there are no features to choose from, the decision tree construction is completed.

3.3.3 Decision Tree Pruning

The purpose of pruning decision trees is to prevent overfitting. Some unnecessary classification features are removed from the generated decision trees by optimizing the loss function to reduce the overall complexity of the model. The method of pruning is to start from the leaf nodes of the tree, shrink upward, and gradually judge [12, 13]. If the loss function corresponding to the whole decision tree is smaller after cutting off a certain feature, then cut off the branches of the feature set.

Decision tree pruning is generally realized by minimizing the overall loss function of the decision tree. The loss function can be expressed as:

$$U_a(T) = U(T) + a|T| \tag{9}$$

In the formula, T represents any subtree; $|T|$ represents the number of nodes in the subtree; $U(T)$ represents the prediction error of the training data; a is used to measure the complexity of the model and the degree of fit of the training data.

3.3.4 Decision Tree Optimization

The main purpose of this article is to optimize the decision tree classifier during the process of constructing a decision tree. Therefore, a decision tree optimization strategy based on Differential Grey Wolf Optimization (DGWO) is proposed, with the accuracy

of the decision tree classifier as the objective function. By optimizing and modifying the weight of data categories, the optimal decision tree is constructed.

The differential grey wolf optimization algorithm basic idea is to introduce the idea of differential evolution on the basis of GWO, and by introducing crossover and mutation operations, it increases the algorithm's local search ability and the ability to jump out of local optima [14, 15]. In addition, DGWO also introduces the mechanism of Tabu search to make the algorithm more easily jump out of the local optimal solution, thus obtaining better global search ability.

The specific implementation process of the DGWO algorithm is as follows:

(1) Initialize the population: randomly generate a certain number of gray wolf individuals and calculate their fitness.
(2) Updating of gray wolf individuals: According to the location and fitness of gray wolf individuals, calculate the direction and distance of updating of gray wolf individuals, and update the location of gray wolf individuals.
(3) Differential evolution operation: Randomly select a certain number of gray wolf individuals as the mutation population according to a certain probability, and generate a certain number of offspring individuals by selecting the parent individual and mutation operation. By comparing the fitness of the offspring and the parent, the individuals with better fitness were selected to update the gray wolf population.
(4) Cross operation: According to a certain probability, randomly select two gray wolf individuals for cross operation. The cross operation can be performed by linear interpolation.
(5) Tabu search: Tabu search mechanism is introduced to avoid the algorithm falling into local optimal solution. This is generally achieved by setting the length of taboos, maintaining taboo tables, and other methods.
(6) Judgment of termination conditions: after multiple iterations, terminate the algorithm by judging whether the fitness in the population converges or reaches the preset number of iterations.

In a word, DGWO algorithm introduces differential evolution and Tabu search functions on the basis of grey wolf optimization algorithm, which can search the solution space more accurately and avoid local optimal solution. In addition, due to its simple principle and easy implementation, the algorithm has good application prospects in various optimization problems.

Based on the differential grey wolf optimization algorithm, the objective function for optimizing the classification of innovation and entrepreneurship teaching data is:

$$Y = \max(accuracy) \tag{10}$$

In the formula, *accuracy* represents the accuracy of the decision tree classifier.

At present, the GWO algorithm has solved many engineering problems, but it also has certain limitations and lacks the ability to find global optimal solutions that affect the convergence speed of the algorithm. The model has high complexity and is essentially nonlinear. Due to the linear decrease in convergence factor, it cannot truly reflect the actual search process. In order to overcome the limitations of conventional GWO algorithms in terms of efficiency, exploration, and development characteristics, this paper

proposes a differential grey wolf optimization algorithm. On the basis of the conventional GWO algorithm, update the optimal solution position as follows:

$$X(\theta+1)=\left(\frac{X_1+X_2+X_3}{3}\right)-\left(\frac{X_1'+X_2'+X_3'}{3}\right) \tag{11}$$

In the equation, $X_1'X_2'$ and X_3' represent the first, second, and third optimal solutions, respectively. Based on the optimization results, obtain the optimal classification results of innovation and entrepreneurship teaching data, and provide data reference for student entrepreneurship.

4 Experiments and Result Analysis

To verify the effectiveness of the online integrated classification algorithm for innovation and entrepreneurship teaching data based on decision trees, experimental research was conducted.

4.1 Experimental Environment Configuration

The hardware environment used in this experiment is Intel i5–6500 CPU, 12G system memory, and 500G available hard disk space. The software environment used in this experiment is Windows 10. Matlab is mainly used to normalize experimental results for easy plotting.4.2Experimental results.

4.2 Analysis of Experimental Results

In order to enhance the credibility of the experimental results, the methods of reference [6] and reference [7] were used as comparative methods for comparative analysis with the proposed methods. Firstly, the data classification accuracy of the three methods was compared, as shown in Fig. 4.

By analyzing Fig. 4, it can be seen that as the number of data samples increases, the data classification accuracy of the three methods shows a gradually decreasing trend. Among them, the highest data classification accuracy of the proposed method is 94.5%, while the highest data classification accuracy of the methods in reference [6] and [7] are 84.1% and 79.8%, respectively, which are significantly lower than the proposed method. This indicates that the proposed method has a high accuracy in data classification and can accurately classify different types of innovation and entrepreneurship teaching data.

Secondly, the data misclassification rate was used as an experimental indicator to further validate the data classification effect of the proposed method, as shown in Fig. 5.

Analyzing Fig. 5, it can be seen that as the number of data samples increases, the data misclassification rates of all three methods show a gradual increasing trend. Among them, the highest data misclassification rate of the proposed method is only 3.8%, while the highest data misclassification rates of the methods in reference [6] and [7] are 11.9% and 13.3%, respectively, which are significantly higher than the proposed method. From the above comparison results, it can be seen that the proposed method can effectively reduce

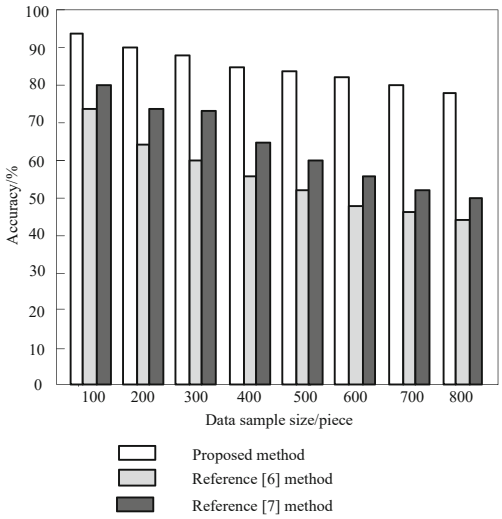

Fig. 4. Comparison Results of Data Classification Accuracy

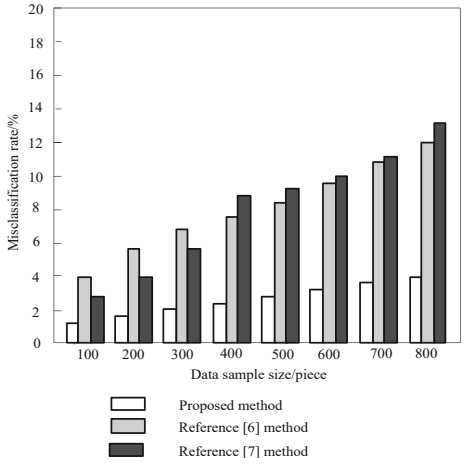

Fig. 5. Comparison Results of Data Misclassification Rate

the data misclassification rate, further verifying the reliability of its data classification results.

Finally, the data classification efficiency was used as an experimental indicator to compare the application effects of the proposed method, reference [6] method and reference [7] method. The results are shown in Fig. 6.

From Fig. 6, it can be seen that the data classification time of the proposed method is significantly lower than that of the methods in reference [6] and [7], with a classification time consistently below 15 s. However, the data classification time of the methods in

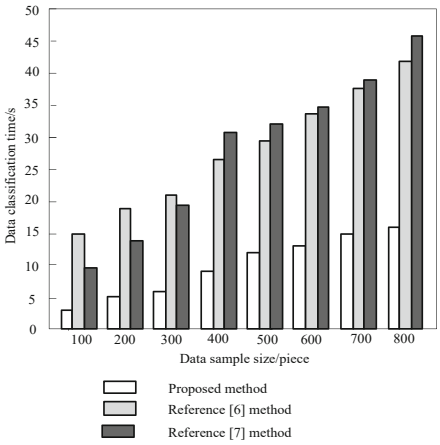

Fig. 6. Comparison Results of Data Classification Efficiency

reference [6] and [7] has reached over 40 s. This indicates that the data classification efficiency of the proposed method is higher, and it can achieve rapid division of innovation and entrepreneurship teaching data in a shorter time.

5 Conclusion

Aiming to improve the accuracy of data classification, reduce misclassification rates, and improve classification efficiency, a decision tree based online integrated classification algorithm for innovation and entrepreneurship teaching data is proposed. Establish a prototype system to achieve online integration of innovation and entrepreneurship teaching data through this system; Based on the results of data integration, a decision Tree model is constructed, and a fuzzy decision tree is obtained by combining the fuzzy theory with the decision tree to solve the problem of data imbalance and missing data types. The difference gray wolf optimization algorithm is used to optimize the decision tree, and the decision tree is improved through feature selection, decision tree pruning and other operations to obtain the best innovation and entrepreneurship teaching data classification results. The experimental results show that the proposed method has a high data classification accuracy, with a maximum value of 94.5%, a low misclassification rate, and a high classification efficiency, indicating that the method can improve the data classification effect. In the future, research will be conducted on how to effectively integrate and integrate innovation and entrepreneurship teaching data from different sources and formats, in order to better support teaching decision-making and evaluation.

References

1. Zhang, X., Zhou, X., Zhao, C., Shao, L.: Unbalanced data classification based on hesitant fuzzy decision tree. Comput. Eng. **45**(08), 75–79+91 (2019)
2. Chen, L., Fei, H., Ding, H., Cheng, L., Zhai, J.: A data sampling method based on double decision tree. Comput. Eng. Sci. **41**(01), 130–135 (2019)

3. Lu, X., Chen, Y., Xiong, Z., Liao, B.: A fast parallel decision tree algorithm for big data analysis. J. Yunnan Univ. (Nat. Sci. Edn.) **42**(02), 244–251 (2020)
4. Wang, J., Yan, J.: Classification algorithm based on undersampling and cost-sensitiveness for unbalanced data. J. Comput. Appl. **41**(01), 48–52 (2021)
5. Zhang, Z., Wu, D.: An imbalanced data classification method based on probability threshold Bagging. Comput. Eng. Sci. **41**(06), 1086–1094 (2019)
6. Meng, Y., Zhou, Q., Shi, H., Ma, N.: Data classification method based on improved ID3 algorithm. Comput. Simul. **39**(05), 329–332+417 (2022)
7. Li, J., Wang, X.: XGBoost for imbalanced data based on cost-sensitive activation function. Comput. Sci. **49**(05), 135–143 (2022)
8. Zheng, J., Li, X., Liu, S., Li, D.: Improved random forest imbalance data classification algorithm combining cascaded up-sampling and down-sampling. Comput. Sci. **48**(07), 145–154 (2021)
9. Bao, H., Fan, X.: Simulation of dynamic classification for unbalanced big data in cloud computing environment. Comput. Simul. **37**(08), 311–314+461 (2020)
10. Liu, P., et al.: Remotely sensed data classification by collaborative processing of Landsat, radarsat-2 and topography information. Remote Sens. Technol. Appl. **34**(06), 1269–1275 (2019)
11. Liu, X., et al.: Imbalanced data classification algorithm based on ball cluster partitioning and undersampling with density peak optimization. J. Comput. Appl. **42**(05), 1455–1463 (2022)
12. Chen, L., Tang, X.: Improved sine cosine algorithm for optimizing feature selection and data classification. J. Comput. Appl. **42**(06), 1852–1861 (2022)
13. Liang, Y., Liu, X., Li, Q., Bai, Y., Ma, Y.: Classification method for unbalanced and small sample data in judicial documents. J. Comput. Appl. **42**(2), 118–122 (2022)
14. Li, A., Han, M., Mu, D., Gao, Z., Liu, S.: Survey of multi-class imbalanced data classification methods. Appl. Res. Comput. **39**(12), 3534–3545 (2022)
15. Zhou, E., Gao, S., Shen, Z.: Classification algorithm of imbalanced data based on rotation balanced forest. Comput. Eng. Des. **43**(02), 458–464 (2022)