IETI Transactions on Engineering Research and Practice

http://ieti.net/TERP/

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

Research on Container Cloud Task Classification Algorithm based on Improved Random Forest

Mengnan Qiu[1, a],Xiaolan Xie[1,2,b],Xiaohua Tao[1,c],Ke Xu[1,d]and Yarong Liu[1,e]
[1]College of Information Science and Engineering,
Guilin University of Technology,
Guilin, Guangxi Zhuang Autonomous Region, China
[2]Guangxi Universities key Laboratory of Embedded Technology
and Intelligent System, Guilin University of Technology, Guilin, China
[a]837315716@qq.com, [b]237290696@qq.com,[c]1156036381@qq.com,
[d]583114572@qq.com,[e]59131746@qq.com

Abstract: Nowadays, the emerging container technology can solve the problem of data processing conveniently and quickly, which has greatly promoted the development of cloud computing. The task assignment of cloud platform has been a research hotspot in recent years, and task classification is an indispensable step in task assignment. How to apply container technology to task classification is worthy of further study. This paper proposes a container cloud task classification algorithm based on improved random forest. Firstly, the fireworks algorithm is used to continuously and repeatedly set the parameter combination of random forest according to the classification accuracy of random forest. Through a finite number of iterations, an ideal random forest classification model is obtained. Secondly, ac-cording to the improved random forest classifier training container cloud task classification model, the classification accuracy and the optimal random forest classification model are returned. Finally, the effectiveness of the improved algorithm is verified by CloudSim container cloud simulation software. The experimental results show that the classification effect of FWA-RF is better than BP neural network, original random forest classification algorithm and random forest algorithm based on the genetic algorithm optimization.

Keywords: Container cloud; Classification algorithm; Fireworks algorithm; Random forest; CloudSim

1. INTRODUCTION

Due to the fast deployment and light weight of containers, more and more Internet companies and manufacturing companies choose container clouds as a deployment platform, making container clouds face greater challenges in task classification. When the container cloud platform receives a large number of task requests, because the container cloud bottom layer adopts the container virtualization technology, the overall execution speed of the task is accelerated, and the classification accuracy rate of the container task is put forward higher.

With the development of artificial intelligence, machine learning has received much attention with its outstanding performance in this field. Among them, Random Forest (RF) is one of the most widely used classifiers in massive data processing. It has excellent performance in response speed and accuracy, and is often designed into different models according to different application requirements.

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

Such as classification models, predictive models, regression models, etc., and thus have received wider attention and application.

The research on the improved algorithm of random forest is not limited to some optimization combination methods, and the research on internal structure is also constantly improved. These studies are mainly concentrated in three aspects: the number of decision trees, the way in which features are selected, and the number and manner of sampling. Oshiro and others in the literature [1] by exponentially increasing the number of trees, the experimental results show that the accuracy of random forest classification does not increase with the number of trees, and the paper also shows that when the tree When the number exceeds a certain threshold, the accuracy of the random forest does not rise and fall. Cerrada M et al. [2] transformed the problem of random feature selection in random forests into an optimization problem, and solved the optimization problem using genetic algorithm. The result of the solution is used as the input attribute set of individual classification in random forest. In the literature [3], Adnan used the genetic algorithm to pruning the trees of random forests, and selected the subtrees with high accuracy to form the final random forest classification model. Mishina Y proposed in the literature [4] to set weights for each classifier and each sample in the random forest integration. The weights of the samples that have been correctly classified are low, and vice versa, the weights are high, thereby increasing the sample of misclassification. The probability of being selected twice. Ma L et al. [5] used the CURE algorithm to cluster the samples first, and then randomly generated an artificial sample between the target sample and the central sample to increase the number of training samples.

Firework Algorithm (FWA) was proposed as a novel optimization algorithm by Tan et al. of Peking University in 2010 [6]. The fireworks algorithm not only has an advantage in the global search process, but also achieves a good convergence effect in local search. Since 2013, Zheng [7] et al. proposed an enhanced fireworks algorithm for the defects in fireworks algorithms. The algorithm and its improved algorithm are widely used in classification problems in various fields [8-11].

In recent years, the application of fireworks algorithms in classification models has also begun to emerge. In 2015, Zhu Xiaodong et al. [12] studied the structure and parameters of fuzzy system model by using fireworks algorithm, and proposed a fuzzy system modeling method based on fireworks algorithm and differential evolution algorithm. In 2017, Lu Yonghe et al. used the binary coding method in literature [13] to transform the feature selection problem in the classification algorithm into the discrete space combination optimization problem, and solved the optimization problem by using the fireworks algorithm. In 2018, Shen Yongliang [14] proposed an improved fireworks algorithm-based SVM feature selection and parameter optimization algorithm. For feature selection problems, a binary-encoded fireworks algorithm was used, and an SVM based on RBF kernel function was used. Reduce the accuracy of classification while reducing the feature data.

Based on the above analysis, the development of fireworks algorithm has become more and more mature, and its application in the classification model is more and more extensive, which can be applied to the parameter optimization of random forests to improve the classification accuracy of random forests. At present, no random forest algorithm has been found in the classification of container cloud tasks. Therefore, it is of great significance to study how to improve the random forest

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

classification model under the container cloud environment. Therefore, this paper studies the container cloud task classification algorithm based on improved random forest.

2. RANDOM FOREST BASED ON FIREWORKS ALGORITHM OPTIMIZATION

Since the random forest has only two parameters, it is related to the accuracy of the entire classification model. It can be seen that this combination of parameters plays a pivotal role in the classification of random forests. But so far there is no effective way to guide the setting of this parameter combination. Therefore, this paper proposes a random forest improvement algorithm based on fireworks algorithm (FWA-RF).

2.1. Fireworks Algorithm

The fireworks algorithm relies on the feedback value of the objective function to continuously update the scope and number of explosions, which is highly explosive. The algorithm is mainly composed of three parts: explosion strategy, Gaussian mutation strategy and selection strategy. First, the population is initialized, N fireworks are generated, the fitness values of the N fireworks are calculated, and the most fitness value and the optimal position are retained. Then, according to the explosion strategy, a new explosion operator is generated at the position of the N fireworks, and the optimal fitness value and the optimal position are updated. Then, according to the Gaussian mutation strategy, M fireworks are randomly selected from N fireworks to generate mutation operators, and the most fitness value and optimal position are updated. Finally, according to the selection strategy, N operators are reselected in the fireworks operator, the explosion operator, and the mutation operator to enter the next iteration. Repeat the above steps until the termination condition is reached to stop the iteration. The specific algorithm flow is shown in Figure 1.

2.1.1 Explosion Strategy

The process of finding the best fireworks algorithm is the process of finding the optimal solution within the scope of the fireworks explosion. The principle of fireworks algorithm explosion is that the better the fitness value, the smaller the radius generated by the individual, the more individual sparks are generated, and the local search is strengthened. On the contrary, the individual with lower fitness value produces less in a larger radius. Individuals that increase global search capabilities. Here, the calculation method of the explosion radius and the number of explosions is as shown in Eq. 1 and Eq. 2.

$$A_i = \hat{A} * \frac{f(x_i - y_{min} + \varepsilon)}{\sum_{i=1}^{N}(f(x_i) - y_{min}) + \varepsilon} \tag{1}$$

$$S_i = \hat{S} * \frac{y_{max} - f(x_i) + \varepsilon}{\sum_{i=1}^{N}(y_{max} - f(x_i)) + \varepsilon} \tag{2}$$

Where A_i, $S_i$represents the radius of the ith individual explosion and the number of sparks produced. $\hat{A}$, $\hat{S}$represent two constants, respectively. $y_{min} = \min(f(x))$, $y_{max} = \max(f(x))$, $x \in X(t)$,

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

X(t) Represents the population of the tth generation. At this time, the position calculation method of the new explosion spark is as shown in Eq. 3.

$$x_{ij}^{k} = x_{i}^{k} + rand(-1,1) * A_i \tag{3}$$

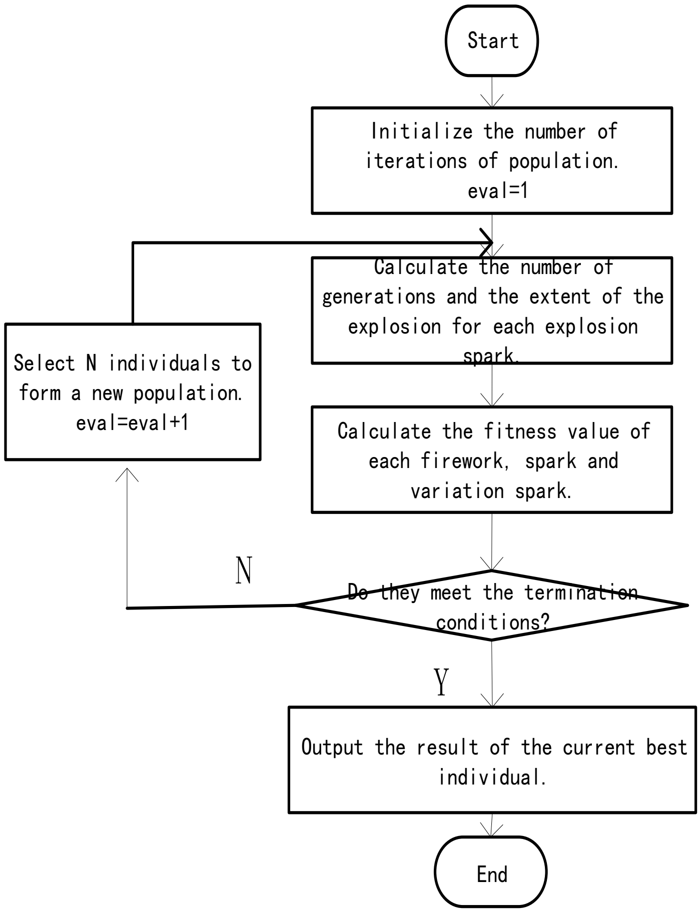

Figure 1.The flow chart of fireworks algorithm

2.1.2 Gaussian mutation strategy

In order to increase the diversity of the population, the fireworks algorithm introduces a Gaussian variation spark. The Gaussian variation spark is generated according to equation (4) within the feasible range, is not limited by the radius of the fireworks, and has the ability to jump out of the local optimum. In order to avoid the new individual generated after the individual mutation near the zero value is still close to 0, the enhanced fireworks algorithm changes the variation mode from Eq. 4 to Eq. 5.

$$X_i = X_i * normrnd(1,1) \tag{4}$$

$$X_i = X_i + normrnd(1,1) * (X_i - X_{best}) \tag{5}$$

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

Where, $normrnd(1,1)$ represents the generation of a random number obeying a normal distribution, and X_{best} represents the position of the optimal individual.

2.1.3 Selection Strategy

The fireworks algorithm needs to select N individuals from the fireworks population, the explosive population, and the mutant population to form the next generation population. In addition to retaining the best fitness value, it is necessary to randomly select N-1 individuals from the remaining individuals. The N-1individuals are selected according to the spatial distance of the spark. The principle of selection is small according to the probability that the spatial distance is close. On the contrary, the probability that the individuals with distant spatial distances are selected is large, so as to maintain the diversity of the population. The spatial distance calculation formula is as shown in Eq. 6.

$$\mathrm{D}(x_i) = \sum_{j=1}^{N} |x_i - x_j| \tag{6}$$

Where $\mathrm{D}(x_i)$represents the distance between thei-th individual and other individuals. Then the next generation of the population is selected according to the way of roulette. The calculation formula of the roulette probability is as shown in Eq. 7.

$$P_i = \frac{D(x_i)}{\sum_{j=1}^{N} D(x_j)} \tag{7}$$

Among them$P_i$indicates the probability that thei-th individual is selected. It can be seen from equation (7) that the denser the spatial distribution of the search individual is, the $\mathrm{D}(x_i)$smaller the probability that the individual is selected, and vice versa.

2.2. Random Forest Improvement Algorithm

2.2.1 Basic Principles of Random Forest

The random forest is integrated with multiple decision trees $\{\mathrm{h}(\mathrm{X}, \theta_n), \mathrm{n} = 1,2, \dots, \mathrm{N}\}$, where Nis the number of decision trees in the integrated classifier, X is the sample data set, and $\theta_n$is the n-th Decision tree classifiers. Selecting multiple decision tree integration methods can avoid a single decision tree from being “arbitrary”. When using a total sample set to train a decision tree, the decision tree classification model is used to fit the classification path of all the samples participating in the training, and the decision tree. The final structure becomes quite large and complex, and when a new sample appears, the probability that the sample will be correctly classified will decrease. Random forests use the "double random mechanism" to avoid overfitting and build multiple simple and weak decision trees, because the similarity between strong decision trees will also increase, and the meaning of integration will be lost. Finally, through the voting results of the simple decision tree set, the final category of the new sample is obtained, which avoids the “prejudice” of the single decision tree and embodies the principle of fairness and justice of the random forest. The logical forest structure diagram of random forest is shown in Figure 2. The overall implementation of the RF algorithm is shown in Algorithm 1.

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

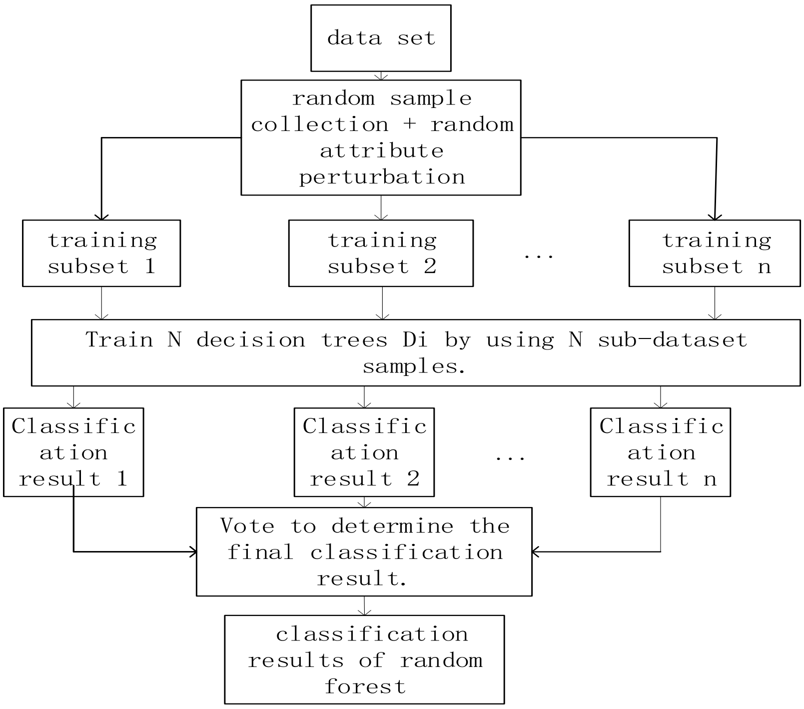

Figure 2.The logical structure of random forest

Algorithm 1： Random forest classification algorithm

Input:Data set is D.The number of base classifiers is T. The number of samples per base classifier is S.

Output： Classification accuracy and classification model.

Step 1: Divide the training data set D_train and the test data set D_test.

Step 2： Repeat steps (1) through (5) below until the termination condition is met.

（1） For i = 1：T

（2） S samples are randomly selected from the data set D_train according to autonomous sampling.

（3） D_traini；Select K attributes from the extracted sample set randomly to form a sample set D_train$_i$.

（4） Construct a base classifier $M_i$by using a learning method of the sample set $D_i$and C4.5.

（5）Endfor

Step 3: Return the classification model Forest.

Step 4: Classify D_test by using the classification model.

Step 5: Count the voting results of Forest.

Step 6: Return the classification accuracy and Forest.

2.2.2 Improved Algorithm

Suppose $\{\mathrm{h}(\mathrm{X},\theta_n),\mathrm{n}=1,2,\ldots,\mathrm{N}\}$ is a random forest classification model. For a given data set $\mathrm{D}=\{\mathrm{X},\mathrm{Y}\}$, set a marginal function, such as Eq. 8 Shown as follows.

$$mg(X,Y)=av_k I(h_k(X)=Y)-\max_{j\neq Y} av_k I(h_k(X)=j) \tag{8}$$

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

Where I(*) represents the indication function, and the number of times the condition*is satisfied is counted. The marginal function measures the degree to which the average number of correct votes for all samples in data set Xexceeds the wrong vote (the category with the highest false vote in other categories).The larger the value of mg, the higher the classification accuracy of the model. Therefore, the generalization error is calculated as shown in Eq. 9.

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{9}$$

Where $mg(X,Y) < 0$ indicates that the classifier misclassifies a certain sample.$P_{X,Y}$ indicates the proportion of the sample with the wrong classification to the total sample.

Breiman has proved in the literature [15] that according to the law of large numbers, when the subtrees in the random forest are infinite or sufficient, the generalization error will converge to an accurate value. However, in real production, if there are infinite number of subtrees in a random forest, a lot of computing resources and time will be wasted. If the sub-tree has too few training samples, the base classifier will not be able to classify. If the sample size is too large, the similarity between the base classifiers will be increased, which will not achieve the purpose of integration. In order to solve the above problems, the fireworks algorithm is used to continuously and repeatedly set the parameter combination of random forest according to the classification accuracy of random forest, in order to obtain an ideal random forest classification model through a finite number of iterations.

First, when using the fireworks algorithm to optimize random forests, a random forest classification accuracy model needs to be established, as shown in Eq. 10.

$$\mathrm{f}(\mathrm{T},\mathrm{S}) = \frac{TS_{true}}{TS_{all}} \tag{10}$$

Among them, $\mathrm{f}(\mathrm{T},\mathrm{S})$represents the accuracy of the random forest classification model. TS_{true} and TS_{all} respectively indicate the number of correctly classified samples and the total number of training samples in the data set.

According to the way the fireworks algorithm finds the optimal solution, the radius AA_i of the i-th individual explosion and the number SS_i of the generation of the child combination for each explosion are calculated, as shown in Eq. 11 and Eq. 12.

$$AA_i = \hat{A} * \frac{f((T,S)_i) - y_{min} + \varepsilon}{\sum_{i=1}^{N}(f((T,S)_i) - y_{min}) + \varepsilon} \tag{11}$$

$$SS_i = \hat{S} * \frac{y_{max} - f((T,S)_i) + \varepsilon}{\sum_{i=1}^{N}(y_{max} - f((T,S)_i)) + \varepsilon} \tag{12}$$

Where $y_{max} = \max(f(T,S))$, where N is the number of initialization parameter combinations. SS_i represents the number of next-generation parameter combinations produced by the i_th parameter combination.

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

According to the explosion mode of ordinary individuals in the FWA algorithm, the combination of parameters of the random forest is updated. The combination generated by this method is called the explosion combination, and the newly generated explosion combination is calculated as shown in Eq. 13 and Eq. 14.

$$T_{ij} = T_{ij} + rand(-1,1) * AA_i \tag{13}$$

$$S_{ij} = S_{ij} + rand(-1,1) * SS_i \tag{14}$$

Where $(\mathrm{T,S})_{\mathrm{i}}$ denotes the i-th ordinary individual combination, and $(\mathrm{T,S})_{\mathrm{ij}}$ denotes the j-th explosion combination generated by the i-th ordinary individual combination explosion.

According to the variation method in the FWA algorithm, the parameter combination of the random forest is updated. The combination generated by this method is called the mutation combination, and the newly generated variation combination is calculated as shown in Eq. 15 and Eq. 16.

$$T_i = T_i + \mathrm{normrnd}(1,1) * (T_i - T_{best}) \tag{15}$$

$$S_i = S_i + \mathrm{normrnd}(1,1) * (S_i - S_{best}) \tag{16}$$

Where, $normrnd(1,1)$represents the generation of a random number obeying a normal distribution, andT_{best}, S_{best} represents the explosion radius of the optimal individual and the number of sparks generated.

In summary, the overall implementation of the FWA-RF algorithm is shown in Algorithm 2.

Algorithm 2： FWA-RF

Input:The data set is D.

Output:the optimal classification model and the highest accuracy

Step 1: Initialize N parameter combinations, calculate their classification accuracy according to Algorithm 1, and retain the optimal classification model.
Step 2:Cycle through the following steps (1) to (13) to find the optimal classification model until the termination condition is met.

（1）while(the highest number of iterations)

（2）for i= 1 : N

（3）Calculate the radius AA_i of the i-th individual explosion and the number of child combinations SS_i generated by each parameter combination according to the equation (11) and (12).

（4）for j = 1: SS_i

（5） Update the explosion combinations according to the equation (13) and (14), count their classification accuracy according to Algorithm 1, and update the optimal classification model.

（6） Endfor

（7） Endfor

（8）M combinations are randomly selected from N.

IETI Transactions on Engineering Research and Practice

http://ieti.net/TERP/

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

（9） For k = 1: M

（10） Update the mutation combinations according to the equation (15) and (16), update their classification accuracy according to Algorithm 1, and update the optimal classification model.

（11） Endfor

（12） Choose explosive combinations of the next generation.

（13）Endwhile

Step 3: Return the optimal classification model and the highest accuracy.

3. CONTAINER CLOUD TASK CLASSIFICATION ALGORITHM BASED ON IMPROVED RANDOM FOREST

In order to effectively allocate tasks in the container environment, the tasks are classified by the container cloud task at the time of task submission, and are divided into five levels with different execution time lengths. Since the task handled in this paper is a computation-intensive task, when constructing the container task data set, the selected attributes include the task submission time, the memory capacity required for task execution, the averageCPU usage time, and the task request execution time. The state of the task execution and the actual execution time of the task, the attribute set of the container task data set is composed as {Subt, reqMem, reqCpu, reqT, Stat, RunT}, then each class in the container task set is labeled with the class attribute Class, and the RunT attribute is removed, and the Class attribute is added to form the final set of attributes of the container task set, which can be represented as Attri={Subt, reqMem, reqCpu, reqT, Stat, Class}. Since the task has no task status attribute value at the time of submission, the task status attribute value of each submitted task before commit is -1.

When training the container cloud task classification model according to the improved random forest classifier, it is mainly divided into two steps. The first step is to construct a task set with a class label. In the second step, the task set with the class label is used as the input parameter of Algorithm 2, and the optimal random forest classification model is returned. The specific implementation process is shown in Algorithm 3. When the training of the container cloud task classification model based on the improved random forest is completed, as shown in Figure 3, $\{\theta_1, \theta_2, \ldots, \theta_n\}$ represents n base classifiers, and the leaf nodes of each base classifier represent one type. The classification result, such as [10000], indicates the first classification result, that is, SLT, [01000] indicates the second classification result, that is, LT, and so on. The specific correspondence is shown in Table 1. The model is delivered to the task distribution center. When a new task is submitted to the container cloud platform, the container cloud data center invokes a random forest classification model corresponding to the task capacity level, and all the base classifiers in the model will classify the task, and then submits its own classification results to the random forest. The random forest finally analyzes the voting results of all the base classifiers, and the result is the final category of the task.

Table 1. Description of leaf node representation in container cloud task classification model based on random forest

The representation of the leaf node	[10000]	[01000]	[00100]	[00010]	[00001]
Task type	SLT	LT	NT	ST	SST

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

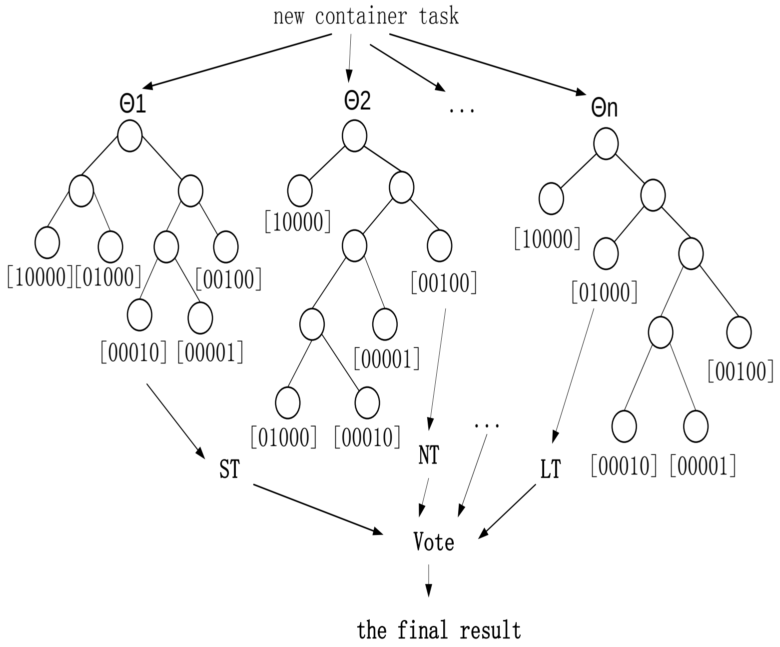

Figure 3.Container cloud task classification model based on random forest

---

Algorithm 3： Container cloud task classification algorithm based on FWA-RF algorithm

Input： The task set of container cloud is DockerTask.The number of categories is L.

Output：the optimal random forest classification model

Step1: According to the RunT attribute in the task set, task samples in the task set are divided into L categories to generate a task set DockerTask_Class with category attributes.
Step 2: The task set DockerTask_Class is used as the input parameter of Algorithm 2, and the random forest classification model is trained to return the classification accuracy and the optimal random forest classification model.

---

4. EXPERIMENT AND RESULT ANALYSIS

4.1. Experimental Environment and Settings

The experimental hardware environment is Intel(R) Core(TM) i5-3470 CPU @3.20GHz, 8G RAM; the software environment includes win10 64-bit operating system, MATLAB 2014b and CloudSim 4.0. MATLAB is mainly used to implement various algorithms used in this paper and comparative experiments, including RF, FWA-RF, GA-RF, BP neural network and other intelligent algorithms.

As the latest version of the cloud computing simulation platform, CloudSim 4.0 extends the container cloud simulation module based on the original. Cloudsim's own architecture is a multi-level model that can be roughly divided into four levels, from top to bottom: Usercode, Cloudsim, GirdSim, and SimJava. CloudSim 4.0 adds a container and container management module called ContainerCloudsim at the Cloudsim layer. Developers can add their own algorithms at this level to implement task allocation strategies as well as resource management and monitoring.

http://ieti.net/TERP/

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

4.2. Introduction to Task Data Set

The task dataset is sourced from the Grid Workloads Archive website, which provides a large number of workload tracking records for the network environment and is intended for use by industry researchers and practitioners. Dataset download link: http://gwa.ewi.tudelft.nl/datasets/gwa-t-4-auvergrid. The data set has more than 400,000 task samples, each with a total of 29 attributes. Select 2500 task samples from 400,000 pieces of data, including 500 SLT class samples, LT class samples, NT class samples, ST class samples, and SST class samples. The task sample classification is divided according to the execution time, as shown in Table 2. 100, 200, 300, 400, and 500 samples are randomly selected from each type of task sample set to form a sample set of 500, 1000, 1500, 2000, and 2500 tasks. The corresponding level of random forest classification model is trained according to different levels of sample sets.

Table 2. Classification of tasks by duration of execution

The execution time of tasks	(20000,25000)	(15000,20000)	(10000,15000)	(5000,10000)	(0,5000)
Category label	SLT	LT	NT	ST	SST

The composition of the task sample attribute in each task set is Attri={Subt, reqMem, reqCpu, reqT, Stat, Class}, according to the submission time of the task submitted to the container cloud platform, the request execution time, the requested memory amount, and the processing. The number of devices and the execution state of the task are classified into the task tasks received by the data center by using the FWA-RF classification algorithm, and are divided into tasks of different durations, namely, one of SST, ST, NT, LT, and SLT.

4.3. Experimental Results and Analysis

In order to verify the effectiveness of the fireworks algorithm optimized random forest FWA-RF to classify tasks in the container cloud environment, the BP neural network and the original random forest classification algorithm, and the genetic algorithm optimized random forest algorithm are used to compare different levels. The task set is classified and tested. To be fair, the number of iterations of the FWA-RF algorithm and the GA-RF algorithm is unified to 300 times, and the number of training samples per base classifier is 10 to the random number generated between the total number of samples, the training task set and the test task set. The division ratio is 3:2. In order to control the length of the training time, the number of trees will not exceed 300 trees per training. The specific parameter settings are shown in Table 3. The first row indicates different levels of data sets, and the second and third rows indicate classification. The setting range of the number of base classifiers in the device and the setting range of the number of samples to be used when training each base classifier.

Table 3. Parameter settings in random forest

Data set	500	1000	1500	2000	2500
The number of trees	(0,100)	(0,200)	(0,200)	(0,300)	(0,300)
The number of samples	(0,500)	(0,1000)	(0,1500)	(0,2000)	(0,2500)

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

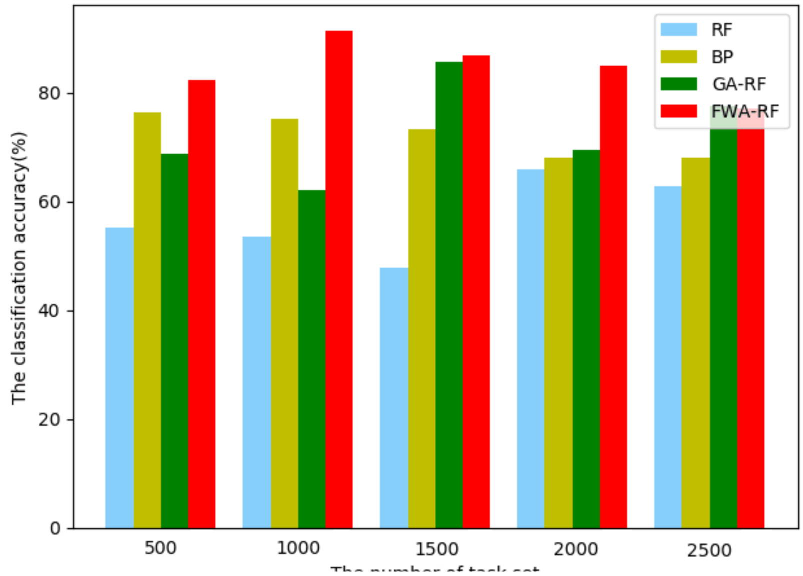

Figure 4.Classification accuracy of four classification algorithms on different task sets

The FWA-RF classification algorithm is used to classify the task set generated in Section 4.2, and the experimental results with RF algorithm, BP neural network, and GA-RF algorithm are shown in Figure4. First, compare the classification effects of BP neural network and random forest. From a vertical point of view, that is, from the classification results of different classification algorithms on the same data set, the classification accuracy of BP neural network is generally higher than RF, but from a horizontal perspective, it is from According to the performance of each algorithm on different datasets, the BP neural network classification algorithm performs almost on any dataset, and basically maintains a correct rate of about 70%. It can be seen that the BP neural network has a small uplift space. In comparison, the fluctuation of RF is relatively large, and the overall trend shows an upward trend, especially when the data set capacity reaches 2000 and 2500, the RF is basically the same as the BP neural network, which shows that the RF ratio is BP. More space for development and optimization.

Then compare the classification effect of FWA-RF algorithm and GA-RF algorithm. Both of these algorithms use intelligent algorithms to optimize random forests. In order to prove the effectiveness of optimizing random forest based on FWA algorithm, the GA-RF algorithm proposed in the literature [69] is used as a comparative experiment in this paper. As shown in FIG. 6, the classification effects of the FWA-RF algorithm and the GA-RF algorithm are first compared. Although the classification results of the FWA-RF and GA-RF algorithms are similar on the 1500-capacity dataset and the 2500-capacity dataset, the FWA-RF is significantly better than the GA-RF on the other three datasets. The accuracy is higher, so it can be seen that the overall optimization effect of FWA-RF on different task sets is better than GA-RF.

In summary, the classification effect of FWA-RF performs best in these four classification algorithms.

5. CONCLUSION

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

This paper proposes a container cloud task classification algorithm based on improved random forest. Firstly, the fireworks algorithm is used to continuously and repeatedly set the parameter combination of random forest according to the classification accuracy of random forest. Through a finite number of iterations, an ideal random forest classification model is obtained. Secondly, according to the improved random forest classifier training container cloud task classification model, the classification accuracy and the optimal random forest classification model are returned. Finally, the feasibility and effectiveness of the container cloud task classification algorithm in this paper are proved by experiments. However, the traditional fireworks algorithm has the disadvantage of being easy to fall into the local optimum. This will cause the parameter setting of random forest to be repeated continuously. In the future, the algorithm can be improved to further improve the accuracy in the classification process.

Acknowledgements

This research work was supported by the National Natural Science Foundation of China (No.61762031), Science and Technology Major Project of Guangxi(No.2018AA32003), The Key Research and Development Program of Guangxi (No.AB17195029, No.AB18126006), Innovation Project of Guangxi Graduate Education (No.YCSW2017156, No.YCSW2018157), Subsidies for the Project of Promoting the Ability of Young and Middle-aged Scientific Research in Universities and Colleges of Guangxi (No.KY2016YB184).

References

[1]Oshiro,Pedro Santoro Perez,et al. How Many Trees in a Random Forest[C]. World congress on the frontiers in intelligent data and signal analysis, International conference on machine learning and data mining. Berlin, Germany, 2017.154-168.

[2]Cerrada M, Zurita G, Cabrera D, et al. Fault diagnosis in spur gears based on genetic algorithm and random forest[J]. Mechanical Systems and Signal Processing, 2015, 70-71:87-103.

[3]Adnan M N, Islam M Z. Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm[J]. Knowledge-Based Systems, 2016, 110:86-97.

[4]Mishina Y, Murata R, Yamauchi Y, et al. Boosted Random Forest[J]. International Conference on Computer Vision Theory and Applications, 2015, 594-598.

[5]Ma L, Fan S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests[J]. BMC Bioinformatics, 2017,169-187.

[6]Tan Y, Zhu Y. Fireworks Algorithm for Optimization[C]. International Conference in Swarm Intelligence. Springer, Berlin, Heidelberg, 2010.355-364.

[7]Zheng S, Janecek A, Tan Y.Enhanced Fireworks Algorithm[C]. IEEE Congress on Evolutionary Computation. IEEE, 2013:2069-2077.

[8]YAN Laiyi. Research on Applications of Fireworks Algorithm in Image Processing[D]. Hubei University of Technology,2017.

[9]Asaju L B, Aminu A A, Peter B S. Training of neural network for pattern classification using fireworks algorithm[J]. International journal of systems assurance engineering and management, 2018,9,1, 208-215.

[10]LUO Yong, GUO Yamo, LIU Chong. Design of Fuzzy Classification System Based on Pareto Firework Algorithm[J]. Computer Engineering,2017,43(02):304-307+316.

2019, Volume 3, Issue 1, 21-34, DOI: 10.6723/TERP.201908_3(1).0003

[11]CUI Ying, SONG Guojiao, Chen Liwei, et al. Hyperspectral image classification based on fireworks algorithm for dimensionality reduction[J]. Journal of South China University of Technology (Natural Science Edition), 2017, 45(3):20-28.

[12]ZHU Xiaodong, LIU Chong, GUO Moya. Design of Fuzzy Classification System Based on Fireworks Algorithm and Differential Evolution Algorithm[J]. Journal of Zhengzhou University (Engineering Science),2015,36(6):47-51.

[13]Lu Yonghe, Chen Yongshan. Feature Selection Method Based on the Binary Fireworks Algorithm[J]. Journal of the China Society for Scientific and Technical Information,2017,36(3):249-259.

[14]SHEN Yong-liang, SONG Jie, WAN Zhi-chao. Improved Fireworks Algorithm for Support Vector Machine Feature Selection and Parameters Optimization[J]. MICROELECTRONICS & COMPUTER,2018(1):21-25.

[15]Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1):5-32.