

Retraction

Retracted: Parallel Classification Algorithm Design of Human Resource Big Data Based on Spark Platform

Security and Communication Networks

Received 26 December 2023; Accepted 26 December 2023; Published 29 December 2023

Copyright © 2023 Security and Communication Networks. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] W. Zhouhuo, "Parallel Classification Algorithm Design of Human Resource Big Data Based on Spark Platform," *Security and Communication Networks*, vol. 2021, Article ID 5811918, 10 pages, 2021.

Research Article

Parallel Classification Algorithm Design of Human Resource Big Data Based on Spark Platform

Wang Zhouhuo 

College of Economics and Management, Shaoyang University, Hunan 422000, China

Correspondence should be addressed to Wang Zhouhuo; 2723@hnsyu.edu.cn

Received 29 October 2021; Revised 22 November 2021; Accepted 13 December 2021; Published 26 December 2021

Academic Editor: Jian Su

Copyright © 2021 Wang Zhouhuo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to solve the problem of large data classification of human resources, a new parallel classification algorithm of large data of human resources based on the Spark platform is proposed in this study. According to the spark platform, it can complete the update and distance calculation of the human resource big data clustering center and design the big data clustering process. Based on this, the K-means clustering method is introduced to mine frequent itemsets of large data and optimize the aggregation degree of similar large data. A fuzzy genetic algorithm is used to identify the balance of big data. This study adopts the selective integration method to study the unbalanced human resource database classifier in the process of transmission, introduces the decision contour matrix to construct the anomaly support model of the set of unbalanced human resource data classifier, identifies the features of the big data of human resource in parallel, repairs the relevance of the big data of human resource, introduces the improved ant colony algorithm, and finally realizes the design of the parallel classification algorithm of the big data of human resource. The experimental results show that the proposed algorithm has a low time cost, good classification effect, and ideal parallel classification rule complexity.

1. Introduction

The development and popularization of modern information technology, such as internet technology and computer technology, have promoted the emergence of the new concept of big data and brought society into the era of big data. Doing a good job in the research on the concept and characteristics of big data has become an inevitable choice in future development [1]. Under the big data background, the human resource management is facing a brand new opportunity and the challenge of how through the effective adjustment causes the human resource management to be able to adapt the time development request is the question which the correlation managers must take seriously. Online recruitment has become an important part of human resource recruitment. On the one hand, it can screen resumes for human resources, find the best job fit personnel, and save time and cost [2]; on the other hand, it can improve job search efficiency by providing job recommendations for job seekers. Online recruitment platform needs to be filled in the resume based on the information and the needs of the

company's human resources to match the position. It is an important standard to test the recommendation algorithm of human resources to solve the high adaptability of massive data, improve the accuracy of matching, and ensure good feedback efficiency [3, 4].

For the classification of big data, reference [5] proposes a design method of big data parallel multilabel k-nearest neighbor classifier based on the Apache Spark framework. In order to reduce the cost consumption of existing MapReduce schemes by using other memory operations, firstly, combined with the parallel mechanism of Apache Spark framework, the training set is divided into several partitions, then found the k-nearest neighbor of each partition of the sample to be predicted in the map stage, and further determined the final k-nearest neighbor according to the results of the map stage in the reduce stage; finally, the nearest neighbor tag sets are aggregated in parallel, and the target tag set of the samples to be predicted is output by maximizing the a posteriori probability. Yang and Zhu [6] propose an intelligent classification method of low occupancy big data under cloud computing. The Bayesian algorithm is used to

construct the intelligent classification model so that the fault tolerance can be minimized through the naive Bayesian intelligent classifier in the subsequent classification, and the compression function and feature selection are constructed to train the intelligent classification model with the same degree of discrimination as the source data and classify the features of the source data through the trained classification model. Finally, the purpose of big data intelligent classification under cloud computing is completed. Bensaïd and Alimi [7] proposed that big data feature selection plays an important role in learning and classification tasks. The goal is to select relevant nonredundant features. Considering a large number of characteristics in real applications, the FS method using batch learning technology cannot solve the problem of big data, especially when the data arrive in order. Therefore, this study proposes an online feature selection system to solve this problem, that is, an online feature selection method based on multiobjective automatic negotiation, so as to improve the classification performance of the ultra-high-dimensional database. In reference [8], a tree crown MapReduce hybrid parallel language fuzzy rule for big data classification in the cloud is proposed. A parallel language fuzzy rule (lfr-cm) framework based on tree crown mapping reduction is introduced. The lfr-cm framework uses the tree crown MapReduce function to classify big data. However, the above existing literature methods have the problems of chaotic data clustering results and high classification time overhead. Therefore, a new human resource big data parallel classification algorithm based on the Spark platform is designed.

According to the proposed method, it adopts selective integration method to study the unbalanced human resource database classifier in the process of transmission, introduces the decision contour matrix to construct the anomaly support model of the set of unbalanced human resource data classifier, identifies the features of the big data of human resource in parallel, repairs the relevance of the big data of human resources, introduces the improved ant colony algorithm, and finally realizes the design of the parallel classification algorithm of the big data of human resource.

The main contributions of this study can be described as follows:

- (1) We study the key technology of parallel classification algorithm design of big data of human resource, which is now very important but always taken for granted
- (2) We study a new method, which is called Spark platform, and we use it to deal with the parallel classification algorithm, which can improve the performance

2. Big Data Clustering and Mining Based on Spark Platform

2.1. Advantages of Spark Platform Clustering. The Spark platform [9] is an important platform for big data computing following the Hadoop platform and represents a significant advantage in dealing with large-scale data. Like the Hadoop

platform, the Spark platform uses the MapReduce computing model for efficient data processing, which solves the time problem of parallel computation of large-scale data but fails to effectively reduce the time cost of complex computation, such as the time cost of iterative computation. In complex computing, iteration is a common computing method. The Spark platform can solve this kind of problem and improve the efficiency of data access by putting iterative computation in memory. The Spark platform's ability to iterate in memory, most importantly because of the use of flexible distributed data sets (RDDs), is also the platform's greatest advantage. All the node memory of the distributed system carries the function of data storage, and there is no need for data access between the external memory, which speeds up the data processing process. In clustering operation, data samples are stored in the memory of each node of the distributed system. Compared with the external memory, the iterative computation is completed in the memory of the node, which saves the time of frequent data interaction from different storage devices.

2.2. Clustering Process Design. The memory of all Spark nodes bears the storage cost of data sample calculation. In actual clustering, the update of the cluster center and distance calculation is completed in the RDD of the memory of all Spark nodes. The iterative calculation of the clustering process also goes through the two steps. The longest iteration in the clustering process is completed in memory, which improves the data input and output efficiency. Therefore, the clustering algorithm based on the Spark platform has a time advantage. The clustering process based on the Spark platform is shown in Figure 1.

2.3. Big Data Frequent Itemset Mining. Big data frequent itemsets are user-defined itemsets whose number is greater than or equal to that in the data set. Known data set $D = \{T_1, T_2, \dots, T_v\}$, any $T_i (i = 1, \dots, v)$ represents an event based on itemset Γ , including event $i d$ and its opposite itemset X . For an event, assuming that there is an itemset Y and $Y \subseteq X$, the event T supports the itemset Y . In set D , all event sets that support itemset Y are called the coverage of Y , which is usually represented by $\text{cover}(Y, D) = \{T | T \subseteq D \wedge X \in T\}$. The absolute support of Y is the size of the set covering the event, which is described as $\Lambda(X, D) = |\text{cover}(X, D)|$. Its relative support is the probability that Y occurs in D , that is, the ratio of the absolute support of Y to the data set D , which is described as $\Lambda(Y, D) = \Lambda(Y, D)/|D|$. The minimum support threshold is known. Assuming $\Lambda_r(Y, D) \geq \lambda_r$, the itemset Y belongs to the frequent itemset.

In the process of mining large data frequent itemsets, there is an a priori property that can be used to compress the search space. A priori property belongs to all nonempty subsets of frequent itemsets [10, 11]. Combined with this definition, if the itemset X is not frequent, it means that the itemset does not meet the requirements of minimum support. At this time, by adding any items X to B , the frequency

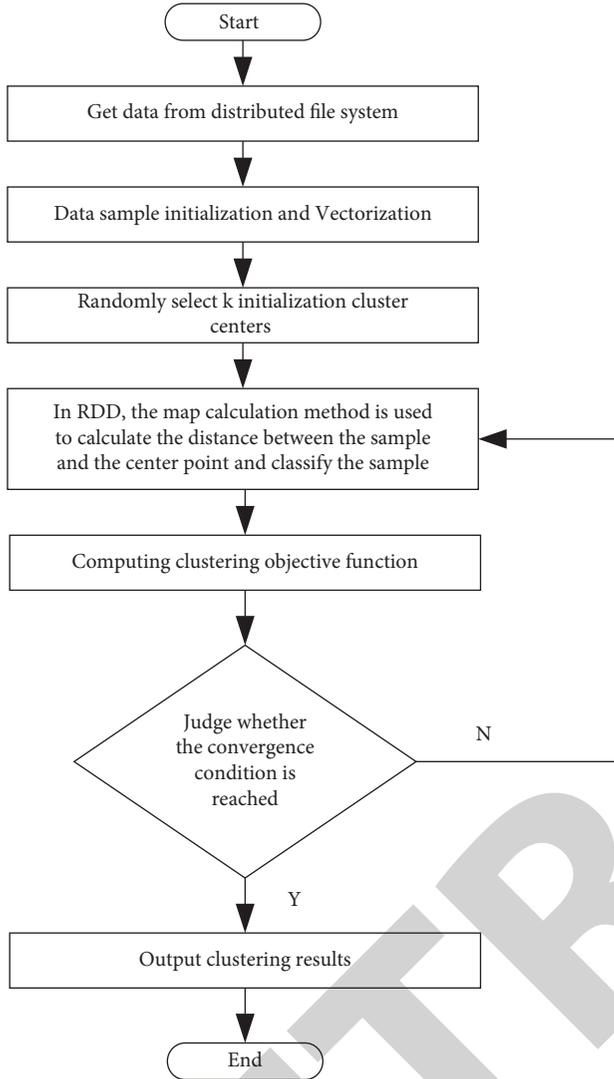


FIGURE 1: Spark platform clustering process.

of the new itemset $(X \cup A)$ is lower than that of X , which means that $(X \cup A)$ is not frequent.

When mining frequent itemsets, in addition to the strategy of compressing the space and reducing the existence of candidates through the characteristics of frequent itemsets, traversing the search space also plays a key role in the statistics of candidate support and the organization of data structure.

K-means clustering method [12] uses the FG-growth algorithm to mine the frequent itemsets of big data, eliminate redundant itemsets, and obtain the core frequent itemsets. At the same time, the core itemsets command the centroid of the original clustering of the K-means algorithm and the number of clusters. Using the FG-growth algorithm to mine frequent itemsets reduces the complexity of the K-means algorithm and the influence of outliers on the K-means clustering algorithm.

The FG-growth method stores the target data in the highly compressed FP-tree data structure, which reduces the number of traversals. The method constructs FP-tree by

counting the element items and mining the frequent itemsets by using the incremental method.

The FP-tree belongs to a tree whose root node is empty, and its build rules are as follows:

Rule 1: enter the itemset. If the itemset already exists, change the count value in the previous path

Rule 2: if the itemset does not exist, you need to create a new path and change the count value of the known path

To make FP-tree access more efficient, a header pointer table is added, which points to the first instance of a known type, which is interconnected in the same type. When accessed, the elements of a known class can be efficiently accessed based on the header pointer table. To better describe the process of building a tree, the process is described in Figure 2.

The conditional pattern base of large data frequent itemsets is mined in FP-tree, and the mining results are used to establish conditional FP-tree. Conditional FP-tree is a part of FP-tree. Mining big data frequent itemsets by this method is a recursive process. The basic process is as follows.

The condition bases of the element item t are $\{z, x, y, s\}$: 2 and $\{z, x, y, r\}$: 1, and elements s and r meet the requirements of minimum support, but itemsets $\{t, s\}$ and $\{t, r\}$ do not meet the conditions of minimum support, so they do not belong to frequent itemsets. The itemsets $\{t, z\}$, $\{t, x\}$, and $\{t, y\}$ belong to frequent itemsets. Then, the conditional pattern base of $\{t, z\}$ is obtained, and the conditional pattern base is recursively processed to obtain all frequent itemsets.

The number of frequent itemsets can describe the number of clusters. Any core frequent itemsets can be the original centroid of K-means [13, 14]. In this algorithm, the threshold is adjusted to know the number of clusters. If the clustering granularity is required to be very high, the threshold can be increased or decreased. The main flow of improved K-means clustering mining large data frequent itemsets is as follows:

Step 1: the FP-growth algorithm is used to command the generation of original cluster centroid and quantity

Step 2: the original cluster centroid and quantity are taken as the input of K-means, and the mining of large data frequent itemsets is completed

3. HR Big Data Parallel Classification Algorithm

3.1. Parallel Recognition of Unbalanced Human Resource Data Based on Fuzzy Genetic Algorithm. In the process of effective parallel feature recognition under the big data classification of human resources, a feature model of unbalanced human resource data is constructed. The training samples of unbalanced human resource data are smoothly updated under the state of genetic iteration, and the central point of the transmitted unbalanced data cluster is updated according to the principle of the small square function value. The obtained power spectrum density function of the transmitted unbalanced data is taken as the feature to optimize the

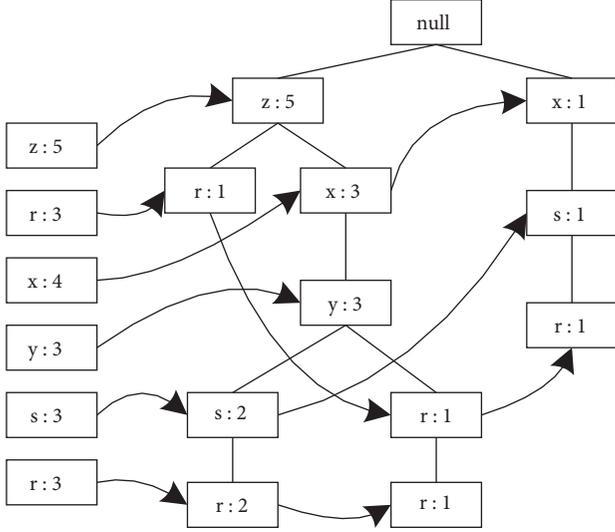


FIGURE 2: FP-tree structure.

features of the transmitted unbalanced data [15]. The process is as follows.

Suppose that $X' = \{x'_1, \dots, x'_n\}$ represents the lower set of human resource big data classification, n' represents the number of data sets X' transmitted, each element in X' is an appropriate amount of p' dimension, X' has categories, $v_\phi = \{v_{\phi 1}, \dots, v_{\phi p'}\}$ represents the center of the ϕ category, the input vector under the human resource big data classification obtained at time t is $\{x_{t1}, \dots, x_{tm}\}$, and the corresponding data type of the vector can be expressed as y_ϕ . Generally, it is defined as normal human resource big data classification when $y_\phi = 1$ and abnormal unbalanced data when $y_\phi = -1$. The effective parallel recognition model of features under human resource big data classification is given by using the following formula:

$$y_\phi = f(x_1, \dots, x_{m'}). \quad (1)$$

The training samples of transmission unbalanced data are extracted through the feature effective parallel recognition model given in the above formula, and the frequency domain model of unbalanced data is constructed as follows:

$$z(t) = j \cdot a(t) + \frac{v_\phi \cdot y_\phi}{\phi(t)}, \quad (2)$$

where $a(t)$ represents the corresponding instantaneous amplitude of unbalanced data complex signal $z(t)$ and $\phi(t)$ represents the frequency domain resonance amplitude of unbalanced data. The ϕ unbalanced data sample belongs to class j by default, and the corresponding maximum membership is marked as $F_{\phi j}$.

The mean a of the membership degree of all samples of each type of transmission unbalanced data belonging to this category is solved and obtained. $E_j = \sum F_{\phi j} / K_j$ and K_j represent the total number of transmission unbalanced data samples of category j . The point set in the high-density area of forged transmission samples is selected as the unbalanced data cluster center set S' , the maximum value s'_ϕ in S' is selected, and this value is regarded as the first cluster center

of transmission unbalanced data. Assuming that the initial frequency mean of the transmission unbalanced data is z_1 and the corresponding standard deviation is $\hat{\sigma}$, the unbalanced data training samples are updated and smoothed in the genetic iteration state, and the state space update iteration is carried out by using the following formula:

$$\hat{\mu}_{\kappa+1} = s'_\phi z_1 + \hat{\beta}_\kappa (1 - \hat{\sigma}) \hat{\mu}_\kappa. \quad (3)$$

The center point of the training sample cluster of transmission unbalanced data is updated by taking the minimum value $\hat{\mu}_{\kappa+1}$ of the square difference function as the criterion, and the power spectral density function $\hat{\beta}_\kappa$ is solved to obtain the unbalanced data. This function is regarded as the characteristic of unbalanced data. Suppose L means that there are n'' samples in the transmission unbalanced data set to form the unbalanced data training set, when transmitting unbalanced data, the corresponding sample category $\omega_{m'}$ of c categories is known. The following solution variables are expressed as Q' and X_ϕ to describe the solution of variable Q' in space. For the corresponding functions of X_ϕ , $f_{m'}$ can be expressed as X_ϕ population fitness function. The essence of the optimal solution obtained is to make $X_{m'}$ and function value $f_{m'}$ the maximum value. The uniform random method is used to generate two intersection points $sw(u)$ for transmitting unbalanced human resource data, and the area covered by the intersection of these two points is set as the matching area.

$$\varepsilon_t(\phi, j) = \frac{\alpha_t(j) b_j(o_{t+1}) F_{\phi j}}{\sum_{\phi=1}^N \sum_{j=1}^N \alpha_t(j) b_j(o_{t+1}) F_{\phi j}}, \quad (4)$$

where $\alpha_t(j)$ represents the communication path deviation of the monitoring node transmitting unbalanced human resource data, and $b_j(o_{t+1})$ represents the Gaussian process with a mean value of 0 and variance of 1. The information fusion filter function of transmission unbalanced data node is designed:

$$H(f) = \sqrt{S_d(f)} / \varepsilon_t(\phi, j). \quad (5)$$

Among them, $S_d(f)$ is used to describe the Doppler power spectrum. A genetic algorithm is used to iterate the characteristics of transmission unbalanced data to realize the feature optimization of transmission unbalanced data. The genetic iterative search dispersion formula of transmission unbalanced data is given by using the following formula:

$$p_{r\phi}(t) = p(t) \cdot h_\phi(t) + f, \quad (6)$$

where $h_\phi(t)$ represents the variation parameter of $h_\phi(t)$ in the process of transmission unbalanced human resource data feature mining and f represents the characteristic response function of transmission unbalanced human resource data.

3.2. Effective Parallel Recognition Method of Human Resource Big Data Features. In the process of effective parallel feature recognition under the big data classification of human resources, based on the result of feature optimization of

unbalanced human resource data, the selective integration method is adopted to study the unbalanced database classifier in the transmission process, and the model of abnormal support degree of the transmitted unbalanced database classifier is constructed by introducing the decision profile matrix [16, 17]. The support entropy is used to measure the category support degree of the classification decision matrix of the transmitted unbalanced data and to solve the fuzzy difference degree problem among each classifier set. The specific process is described as follows.

Suppose that F represents a feature vector of unbalanced data, $\Omega = \{\varepsilon_1, \dots, \varepsilon_v\}$ represents a set of labels of unbalanced data, and the output support of classifiers D_v in the integrated learning method is $D = \{D_1, \dots, D_L\}$. $d_{v,o}(x)$ means that the classifier D_v assumes that the nonequilibrium data F originate from the support of category v_o , and the greater the support, the more likely it is to become a category label v_o . Input F for a given unbalanced HR data sets the output of classifier [18] to the support matrix $D(F)$ of a base classifier, where

$$D(F) = \begin{bmatrix} d_{1,1}(f) & d_{1,1}(f) & d_{1,v}(f) \\ & \vdots & \\ d_{v,1}(f) & d_{v,o}(f) & d_{v,v}(f) \end{bmatrix}. \quad (7)$$

The output value of the classifier $D_v(f)$ is described by the decision-making profile matrix of the unbalanced human resource data in the transmission process, and the values $d_{v,o}(f)$ and $d_{v,v}(f)$ represent the elements of the matrix, and the columns represent the corresponding support degree of the classifier D_v output class v_o . The sparse matrix is adopted to describe the relationship combination between the base classifier and the integration classifier of the unbalanced data transmission, where

$$R(F) = \begin{bmatrix} r_{1,1}(x) & r_{1,1}(x) & r_{1,v}(x) \\ & \vdots & \\ r_{\psi,1}(x) & r_{\psi,o}(x) & r_{\psi,v}(x) \end{bmatrix}. \quad (8)$$

Among them, $r_{v,o}(x)$ and $v = 1, \dots, \psi, o = 1, \dots, v$ are used to describe whether the base classifier o exists in the v classifier and $r_{\psi,v}(x)$ is used to describe whether the base classifier v exists in the ψ classifier.

The method of variance measure is used to select and integrate the unbalanced data in the process of transmission. The precondition of selecting and integrating is to evaluate the support of each classifier set to the unbalanced human resource data set and calculate by using the following formula:

$$g(C_i) = - \sum_j \frac{d_{ik}}{S_i} \ln \left(\frac{d_{ij}}{S_i} \right). \quad (9)$$

Among them, C_i represents the set of i -base classifiers for unbalanced data, d_{ij} describes the support of i -base classifiers for the attributes of the class j of unbalanced data, \bar{n} represents the number of sets of base classifiers for unbalanced data, \bar{m} represents the number of categories of attributes to be classified for unbalanced data, and S_i

represents the total support of each set of classifiers for a single instance of unbalanced data classification.

Multiple observations are used to integrate the differences of classifier class tags of the unbalanced database in the transmission process, and exponential weighting is used to identify the current support entropy of unbalanced data unit time in parallel, namely,

$$\bar{h}_t = \begin{cases} ah_{t-1} + (1-a)\bar{h}_{t-1}, & t > 2, \\ h_1, & t = 2. \end{cases} \quad (10)$$

In the formula, \bar{h}_{t-1} represents the parallel identification entropy value of unbalanced data transmitted at $t-1$ time, a represents the ratio between the dimension size of data and the overall dimension size, and h_{t-1} represents the actual entropy value of unbalanced data transmitted at $t-1$ time, from which the parallel identification entropy value \bar{h}_t of unbalanced data transmitted at t time can be obtained.

The dimension size of human resource big data classification affects the support entropy parallel recognition to a certain extent and adjusts the dimension size according to the change of a value, κ represents the support entropy standard deviation of t unit time unbalanced human resource data classifier set i , κ is used to describe the threshold of t unit time unbalanced human resource database classifier i for the effective parallel recognition difference degree of classification features, and the following:

$$\kappa = \sqrt{\frac{1}{(\psi-1)} \sum_{j=t-\psi}^i (h_t - \bar{h}_{t-1})^2}, \quad (11)$$

$$\delta_{t,i} = \bar{h}_{t,i} \pm 6\sigma_{t,i}. \quad (12)$$

Among them, $[\bar{h}_{t,i} - 6\sigma_{t,i}, \bar{h}_{t,i} + 6\sigma_{t,i}]$ represents the confidence interval range of the classification of unbalanced human resource data during the transmission of the set i of classifier t unit time.

According to the above steps, the parallel recognition of the sequence under the big data classification is completed.

3.3. Association Repair of HR Big Data. At the beginning of parallel classification of human resource big data, the weak association data are first repaired by using the real-time data in the process of association query, and then, the repairable association features are used to establish the most probable association of human resource data, so as to increase the association of massive human resource data in the database under the weak association rules, and so as to facilitate the parallel classification of human resource data in the next step.

Due to the huge amount of human resource data in the database, there are a large number of redundant human resource data, which leads to the shortage of a large amount of human resource data and little information. The existence of these weak association attributes brings great difficulty to information search, which needs to be repeated several times to complete. If the correlation is too weak, even cannot query

the required information, it seriously affects the performance of the database. In order to ensure the efficiency of the query of large database information, it is necessary to repair the relationship between a weakened association of human resources data to ensure smooth query work. The repair conditions are as follows.

A parameter is set as the judgment value of whether each HR data in the large database need to be repaired, and the value is set to m . When $m = 1$, the HR data need to be disconnected for repair, when $m = 2$, the HR data need to be weakly associated for repair, and when $m = 3$, the HR data need to be expanded and simplified for repair. By calculating the probability distribution of m in $\xi K(m)$, we can get whether the nonfixed setting is required. K represents the HR data to be queried, and the value of K represents the status of the HR data.

- (1) $K = 0$ means that there are no HR data to be queried in the database, and only the lost HR data need to be patched; then, the value of $\xi K(m)$ is set to 0.
- (2) When $K = 1$, it indicates that HR big data exist in the database. The value is randomly set, and the range $\xi_1(m) > 0$ needs to be set.
- (3) When $K \neq 0$ and $K \neq 1$ and index probability $\xi K(m) > 0$, all types of repair are required.

3.4. Application of Improved Ant Colony Algorithm in Parallel Classification of Human Resource Data. After updating the weak association rules, the TSP problem is combined with an ant colony algorithm to find an optimal algorithm for the parallel classification of human resource data.

In the parallel classification of human resource data, IF – THEN rules must be extracted from the training set, and all attribute fields in the database are traversed by an ant colony algorithm to set up the decision rules. The specific process is described as follows:

- (1) Any human resource data is selected in the database to form a training set.
- (2) The pheromone of the ant colony algorithm is initially set.
- (3) The ant is placed in the database and traversed through all the attributes in the database; then, the required attributes are selected to establish a parallel classification rule.
- (4) The attributes irrelevant to the inquiry information are filtered out in the rules established in the previous step.
- (5) Pheromones based on the maturity of established rules are added or subtracted, and if the query jumps, the process continues, otherwise step 3 will be returned.
- (6) The parallel classification rules extracted from the above steps are placed in the rule list, and when the rules in the list overwrite all the attributes of the elements in the training set, the program ends; otherwise, return to step 2. First, the pheromones

contained in each term $_{ij}$ should be initially set as follows (13).

$$\tau_{ij} = \frac{1}{\sum_{i=1}^{\alpha} b_i} \quad (13)$$

In the formula, α represents the number of human resource data attributes in the database, and b_i represents the number of elements in the set of A_i attribute values.

The pheromone on each term $_{ij}$ is updated when the ant has completed the traversal of all the attribute sets in the database. The update of pheromone is the increase or decrease in the pheromone content, and the judgment of the increase or decrease is decided by the rules constructed by ants after traversing the set of attributes. The incremental size of the pheromone is determined by the attribute coverage Q of the training set by a parallel classification rule established by the ant after traversing all attributes, as shown in type (14):

$$Q = \frac{TP}{TP - FN} \times \frac{TN}{FP + TN} \quad (14)$$

where TP represents the number of HR data covered by the parallel classification rules and judged as correct by the rules and FP represents the number of HR data covered but judged as incorrect by the rules. TN represents the number of HR data not covered by the rules and determined to be incorrect by the rules, and FN represents the number of HR data not covered by the rules but determined to be correct by the rules.

When using the ant colony algorithm to complete the parallel classification of human resource data, it is necessary to introduce a contribution function to avoid premature convergence of the algorithm and unreasonable allocation of pheromones.

3.5. Improvement of Parallel Classification Rules. The improvement of parallel classification rules is mainly accomplished by postprocessing the rules, which can simplify the rule list by deleting the irrelevant attributes in the rule list, make it express the attributes of human resource data in the database better, and then make the rule list to more accurately judge the attributes of querying human resource data. Moreover, by removing the rules that are not relevant to the required attributes, the algorithm can be made easier to understand and use.

Usually, after the rule list is established by the ant colony algorithm, the rule list must be simplified. For a rule in a rule list, its deletion criterion is to judge whether the rule is conducive to expressing the attribute of the information queried. If the attribute does not contribute to the attribute expression, the influence of the attribute on the quality of the rule list is negative. In the part of rule list improvement, we select the coverage parameter Q of the human resource data set mentioned in the previous section to judge the impact of rules on the list. The process is a cyclic process until each rule is judged to end after the cyclic process to simplify the parallel classification of human resource big data.

TABLE 1: Operating parameters of experimental platform.

Parameters and platform	Parameter value
Processor	P4 2.8c
Memory	584M
Operating system	Windows XP + sp2
Programing language	Java 1.5
Number of ants	9000
Number of candidate sets	100
Maximum number of cycles	50

TABLE 2: Experimental human resource data set.

Data settings	Data quantity	Discrete attribute	Continuous attribute	Number of classes
Human resource category 1	330	8	1	3
Human resource category 2	722	3	7	3
Human resource category 3	1076	4	1	3
Human resource category 4	624	22	1	3
Human resource category 5	317	10	6	3
Human resource category 6	2858	7	6	3

4. Experimental Results and Analysis

4.1. Experimental Results. In order to prove the effectiveness of the designed human resource big data parallel classification method, a simulation experiment is needed. Taking the big data intelligent classification algorithm based on cloud computing proposed in reference [6], the big data classification algorithm based on big data feature selection proposed in reference [7], and the big data classification algorithm based on parallel language fuzzy rules proposed in reference [8] as the control group, the performances of different algorithms are compared through the analysis of experimental results. First, the operating parameters of the experimental platform are set, as shown in Table 1.

Six human resource data sets from the UCI Machine Learning Repository are selected to be simulated with improved methods and traditional methods to demonstrate the accuracy and complexity of the proposed methods and to improve the effectiveness of the proposed methods. The experimental human resource data set is shown in Table 2.

Based on the obtained human resource data set, the experimental analysis is carried out. The improved algorithm and the traditional algorithm are, respectively, used for parallel classification of the human resource data set used in the experiment, and the parallel classification results are shown in Table 3.

It can be seen from Table 3 that the maximum error of the improved method is 2.13%, and the errors of the literature methods are higher than those of the improved method. This is because the improved algorithm establishes the parallel classification standard of massive human resource big data by analyzing the necessary conditions for massive human resource big data training and the distribution function of aperiodic tasks. The comparison of the complexity of algorithm parallel classification rules is shown in Table 4.

Simulation results show that the proposed algorithm is not only more accurate than the traditional parallel

classification algorithm but also more simple and efficient than the traditional method.

The classification overhead is the main index to measure the classification performance, and the classification time overhead of different algorithms is calculated. The result is shown in Figure 3.

According to the test results in Figure 3, the cost of the proposed algorithm is lower than 30 ms, but the cost of the proposed algorithm is relatively stable and does not obviously increase. It is shown that the proposed method can effectively reduce the classification delay and time overhead and has high applicability.

4.2. Human Resource Big Data Classification Effect Test. It is known that the selected human resource big data are divided into three categories: high-quality resources, ordinary resources, and poor-quality resources. The big data intelligent classification algorithm based on cloud computing proposed in reference [6], the big data classification algorithm based on big data feature selection proposed in reference [7], and the big data classification algorithm based on parallel language fuzzy rules proposed in reference [8] are used as the control group. As the control group, the algorithm designed in this study is used as the experimental group. Figure 4 shows the effect of human resource big data classification of four groups of algorithms.

According to Figure 4, there are many differences in the classification effects of different algorithms on human resource big data. The classification results of the big data intelligent classification algorithm based on cloud computing proposed in reference [6] have some isolated data, resulting in the problem of data omission in big data classification. The classification results of the big data classification algorithm based on big data feature selection proposed in reference [7] and the big data classification algorithm based on parallel language fuzzy rules proposed in reference [8] are chaotic, and the classification accuracy is not ideal. In contrast, the designed algorithm can classify three types of human resource big data with high accuracy and good application performance.

TABLE 3: Comparison of accuracy of parallel classification of big data.

Types of human resources	Complexity of the proposed algorithm (%)		Complexity of big data intelligent classification algorithm based on cloud computing (%)		Complexity of big data classification algorithm based on big data feature selection (%)		Complexity of big data classification algorithm based on parallel language fuzzy rules (%)	
	Mean	Deviation	Mean	Deviation	Mean	Deviation	Mean	Deviation
1	57.23	2.13	71.45	4.58	70.22	2.57	68.27	5.59
2	90.74	0.57	91.56	3.54	90.55	4.26	91.26	4.52
3	91.68	0.22	70.29	4.55	69.37	4.70	72.96	4.23
4	88.23	1.41	91.75	6.96	92.75	4.26	92.44	5.74
5	82.25	1.27	89.20	3.97	88.66	3.77	85.22	2.76
6	42.17	1.06	55.21	2.85	54.20	3.75	54.69	5.20

TABLE 4: Comparison of the complexity of algorithm parallel classification rules.

Types of human resources	Complexity of the proposed algorithm (%)	Complexity of big data intelligent classification algorithm based on cloud computing (%)	Complexity of big data classification algorithm based on big data feature selection (%)	Complexity of big data classification algorithm based on parallel language fuzzy rules (%)
1	8.5	15.41	56.7	42.41
2	5.3	25.23	154	52.05
3	2.6	45.79	38.62	42.17
4	8.5	36.54	19.6	44.89
5	7.6	44.20	18.7	50.20
6	9.2	30.47	36.8	30.47

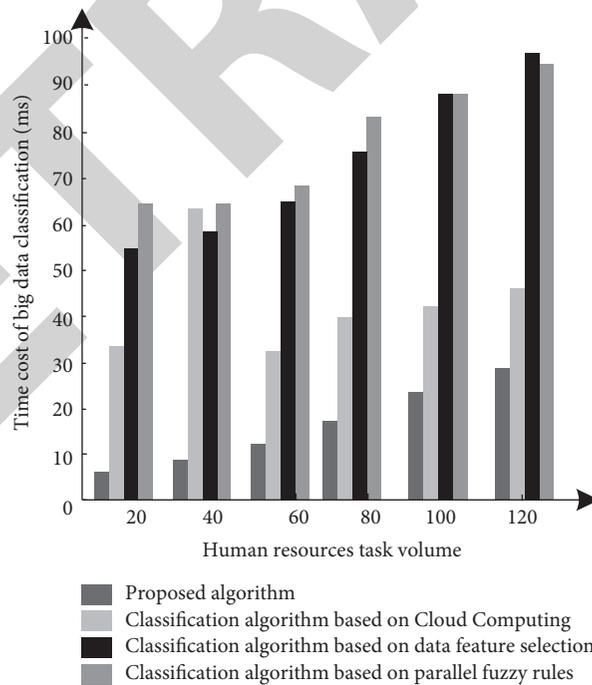


FIGURE 3: Comparison of classification overhead for different algorithms.

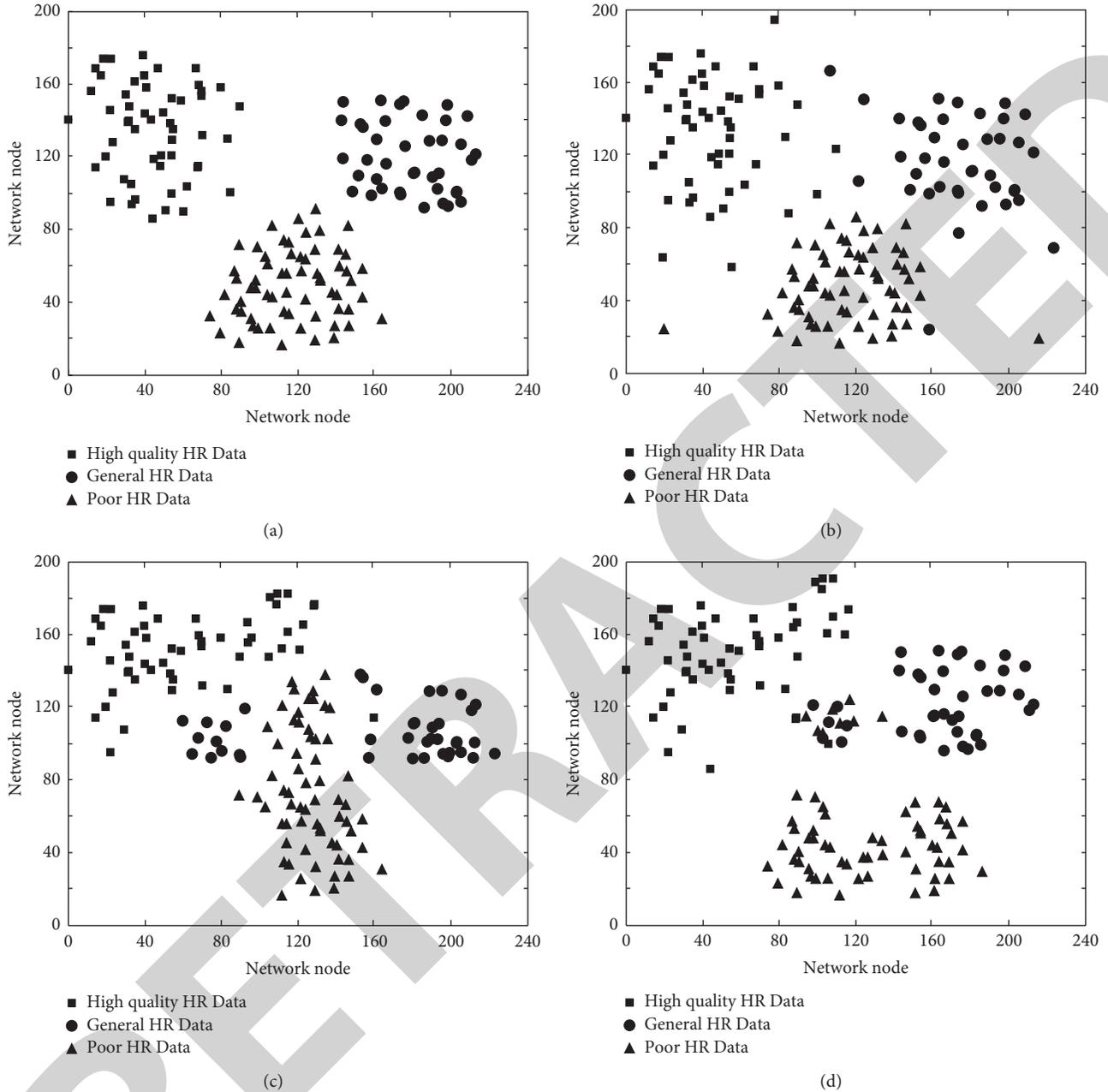


FIGURE 4: HR data classification effect. (a) Designed algorithm. (b) Big data intelligent classification algorithm based on cloud computing. (c) Big data classification algorithm based on big data feature selection. (d) Big data classification algorithm based on parallel language fuzzy rules.

5. Conclusions

At present, the widely used human resource big data classification algorithms have the problems of unsatisfactory data clustering effect and long classification time. In order to improve this disadvantage, a parallel human resource big data classification algorithm based on the Spark platform is designed. Through the calculation of cluster center update and distance on the Spark platform, the clustering effect of big data is optimized. The unbalanced big data are filtered out, the unbalanced human resource data classifier is learned in the transmission process through selective integration, the parallel recognition of human

resource big data features is completed, and the efficiency of parallel classification of big data is improved. The correlation between human resource big data is repaired, and the efficient classification of human resource big data is realized. The experimental results show that the designed algorithm has a better classification effect, can complete the big data task in a shorter time, and has reliable practicability.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] T. C. Chuah and Y. L. Lee, "Intelligent RAN slicing for broadband Access in the 5G and big data era," *IEEE Communications Magazine*, vol. 58, no. 8, pp. 69–75, 2020.
- [2] C. D'Angelo, D. Gazzaroli, and C. Gozzoli, "Organisational welfare in Italian SMEs: the process of valorising human resources," *Sustainability*, vol. 12, no. 22, p. 9318, 2020.
- [3] L. A. Cabrera-Diego, M. El-Bèze, J.-M. Torres-Moreno, and B. Durette, "Ranking résumés automatically using only résumés: a method free of job offers," *Expert Systems with Applications*, vol. 123, pp. 91–107, 2019.
- [4] M. Safran and D. Che, "Efficient learning-based recommendation algorithms for top-N tasks and top-N workers in large-scale crowdsourcing systems," *ACM Transactions on Information Systems*, vol. 37, no. 1, pp. 2.1–2.46, 2019.
- [5] Y. Cao, N. Wang, and Z. C. Xu, "Network big data classification processing method based on Spark and distributed KNN classifier," *Application Research of Computers*, vol. 36, no. 11, pp. 3274–3277, 2019.
- [6] X. Q. Yang and Y. Q. Zhu, "Research on-line monitoring method of noise pollution in robot master-slave control system," *Computer Simulation*, vol. 38, no. 6, pp. 301–305, 2021.
- [7] F. Bensaid and A. M. Alimi, "Online feature selection system for big data classification based on multi-objective automated negotiation," *Pattern Recognition*, vol. 110, no. 1, Article ID 107629, 2020.
- [8] V. Vennila and A. R. Kannan, "Hybrid parallel linguistic fuzzy rules with canopy MapReduce for big data classification in cloud," *International Journal of Fuzzy Systems*, vol. 21, no. 3, pp. 809–822, 2019.
- [9] B. Hosseini and K. Kiani, "A big data driven distributed density based hesitant fuzzy clustering using Apache spark with application to gene expression microarray," *Engineering Applications of Artificial Intelligence*, vol. 79, pp. 100–113, 2019.
- [10] A. Belhassena and H. Wang, "Trajectory big data processing based on frequent activity," *Tsinghua Science and Technology*, vol. 24, no. 3, pp. 317–332, 2019.
- [11] H. Fadishei and A. Soltani, "The curse of indecomposable aggregates for big data exploratory analysis with a case for frequent pattern cubes," *The Journal of Supercomputing*, vol. 76, no. 1, pp. 688–707, 2020.
- [12] B. S. Aski, A. T. Haghighat, and M. Mohsenzadeh, "Evaluating single web service trust employing a three-level neuro-fuzzy system considering k-means clustering," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 1, pp. 1–15, 2021.
- [13] F. Yoseph, N. Malim, M. Heikkil, and B. Adrian, "The impact of big data market segmentation using data mining and clustering techniques," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 1, pp. 1–15, 2020.
- [14] S. Sumalatha and R. Subramanyam, "A MapReduce solution for incremental mining of sequential patterns from big data," *Expert Systems with Applications*, vol. 133, pp. 109–125, 2019.
- [15] H. Amazal and M. Kissi, "A new big data feature selection approach for text classification," *Scientific Programming*, vol. 2021, no. 2, pp. 1–10, 2021.
- [16] S. Hassan, A. M. Hemeida, S. Alkhalaf, A.-A. Mohamed, and T. Senjyu, "Multi-variant differential evolution algorithm for feature selection," *Scientific Reports*, vol. 10, no. 1, Article ID 17261, 2020.
- [17] R. Salman, A. Alzaatreh, H. Sulieman, and S. Faisal, "A bootstrap framework for aggregating within and between feature selection methods," *Entropy*, vol. 23, no. 2, p. 200, 2021.
- [18] R. Hou and J. Li, "Multi-dimensional big data intelligent classification retrieval simulation under multi-layer perception learning," *Computer Simulation*, vol. 37, no. 5, pp. 345–348, 2020.