Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 4140707, 12 pages
https://doi.org/10.1155/2019/4140707

Research Article

An Improved Random Forest Algorithm for Predicting Employee Turnover

Xiang Gao,[1] **Junhao Wen**,[2] **and Cheng Zhang**[1]

[1]*College of Computer Science, Chongqing University, Chongqing 400044, China*
[2]*College of Big Data & Software Engineering, Chongqing University, Chongqing 400044, China*

Correspondence should be addressed to Junhao Wen; jhwen@cqu.edu.cn

Received 13 October 2018; Revised 18 December 2018; Accepted 13 March 2019; Published 17 April 2019

Academic Editor: Nicholas Chileshe

Employee turnover is considered a major problem for many organizations and enterprises. The problem is critical because it affects not only the sustainability of work but also the continuity of enterprise planning and culture. Therefore, human resource departments are paying greater attention to employee turnover seeking to improve their understanding of the underlying reasons and main factors. To address this need, this study aims to enhance the ability to forecast employee turnover and introduce a new method based on an improved random forest algorithm. The proposed weighted quadratic random forest algorithm is applied to employee turnover data with high-dimensional unbalanced characteristics. First, the random forest algorithm is used to order feature importance and reduce dimensions. Second, the selected features are used with the random forest algorithm and the F-measure values are calculated for each decision tree as weights to build the prediction model for employee turnover. In the area of employee turnover forecasting, compared with the random forest, C4.5, Logistic, BP, and other algorithms, the proposed algorithm shows significant improvement in terms of various performance indicators, specifically recall and F-measure. In the experiment using the employee dataset of a branch of a communications company in China, the key factors influencing employee turnover were identified as monthly income, overtime, age, distance from home, years at the company, and percent of salary increase. Among them, monthly income and overtime were the two most important factors. The study offers a new analytic method that can help human resource departments predict employee turnover more accurately and its experimental results provide further insights to reduce employee turnover intention.

1. Introduction

Employee turnover is "the movement of people into and out of employment within an organization" [1]. A large number of departures in a relatively concentrated period of time will create difficulties for management. If the pace of recruitment and training cannot be maintained, important positions may remain vacant, resulting in daily work that is not carried out, and a lack of advancement of the company's development goals and plans. Thus, if we can predict which employees are likely to leave in the future, we can plan ahead, take steps to reduce the likelihood of departure, and train and recruit new employees in advance, particularly, avoiding the loss of key positions and the leaking of core information. In this way, employee turnover forecasting can play an important role in the steady development of an enterprise.

There are many effective algorithms, both qualitative and quantitative, used to stabilize a company by predicting employee turnover. Our focus is on quantitative algorithms. Among these, some algorithms based on the social network method have been applied to employee turnover prediction. For example, Feeley et al. [2] proposed employee turnover prediction based on a friendship network. Gloor [3] proposed a prediction based on an e-mail network. Some classic data mining algorithms have also been used in the field of employee turnover prediction. For example, Hong et al. [4] applied support vector machines (SVM) to the prediction of employee turnover and Kao et al. [5] introduced a decision-tree algorithm to predict turnover among nurses.

However, the methods in these studies referenced above have various limitations in generating accurate employee-turnover predictions such as the following:

(1) Inadequate consideration of unbalanced data. The employee turnover data used in these studies represent only a small portion of all the employees.

(2) Ineffective processing of high dimensional data. There are many feature dimensions among employees and these include static as well as dynamic ones.

(3) Lack of ranking. As the goal of the prediction is to prevent employee turnover, we need to identify and rank its features in terms of importance.

In this study, we propose a weighted quadratic random forest algorithm (WQRF) based on the traditional random forest algorithm combined with data characteristics. By calculating the F-measure of each tree and introducing weighted voting, the WQRF can solve the problem of unbalanced data and rank the features to reduce dimensionality. Compared with random forest, C4.5, Logistic, BP, and other algorithms, the proposed algorithm shows significant improvement in terms of various performance indicators in the area of employee turnover forecasting.

To test the effectiveness of the method, we apply it to an employee dataset of a branch of a communications company in China. The performance of the dataset with the proposed algorithm is proven to be effective. Many experiments have shown that this prediction algorithm has obvious advantages for enterprise employee turnover prediction.

In sum, the significant contributions of this study are as follows:

(1) It proposes a detailed process of WQRF, which is effective in predicting high dimensional unbalanced data.

(2) It proposes a method of employee turnover prediction based on WQRF and verifies it using actual Chinese enterprises.

(3) It points out the main factors of employee turnover in Chinese communication enterprises and provides guidance for the human resource department to help reduce the turnover rate of employees. It also offers value as a reference for the management of employees in other industries.

This rest of the paper is organized as follows. Section 2 presents related studies on employee turnover prediction, some models used in employee turnover prediction, and the factors underlying employee turnover. The proposed approach and feature selection are explained in Section 3 and the experiments and results are illustrated in Section 4. Finally, the conclusions are presented in Section 5.

2. Literature Review

2.1. Employee Turnover Prediction Models. Research on employee turnover prediction has been conducted for several decades. Many useful models have been proposed in both theory and practice. In recent years, some statistical analysis techniques have been used to predict or analyze employee turnover intention. Chien [6] applied the two-phase cluster analysis method to predict employee turnover intention. The study added a self-organized mapping graph into the cluster analysis to find clusters of employee turnover intention features. Wu [7] used a decision-tree to predict employee turnover intention. The study calculated the relationship among job-satisfaction and work arrangements and employee turnover intention; it created a classification method to identify and group potential turnover employees. Mitchell et al. [8] initiated an embedded approach to turnover and Holtom et al. [9] developed social network metrics for oscillation and average response time to identify changes in the communication behaviors of managers who were about to quit their jobs.

2.2. The Factors. Human resource departments need to identify the personal employee characteristics, the work environment factors, and the job attitudes that are most likely to relate to employee turnover. Without such information, turnover control efforts are likely to be fragmented and possibly misguided.

Jung and Yoon [10] surveyed employees in deluxe hotels and found that job satisfaction was related to turnover intent. They investigated how prioritizing an error management culture (EMC) could lead to higher job satisfaction and lower turnover intent. Labrague et al. [11] analyzed nurse turnover in the Philippines and found that age, job satisfaction, and job stress were the most influential factors for turnover intention. Tran [12] investigated high school principals' turnover intentions and found that pay satisfaction and school achievements were negatively associated with turnover intention.

However, the accuracy of these algorithms is not strong enough and the features of employee intention seldom work well in these models. To address this, our study aims to enhance the ability to forecast employee turnover and introduce a new method based on an improved random forest algorithm. Our study offers a new analytic method that can help human resource departments better predict employee turnover and identify the key factors influencing employee turnover intention.

3. Methodology

3.1. Random Forests. The implementation of WQRF is based on the traditional random forest (RF) algorithm. RF is a combination algorithm proposed by Breiman [13] in 2001 where if the predicted result is a discrete value, it is a random forest classification, and if it is a continuous value, it is a random forest regression. Many empirical studies have confirmed the theory that the random forest algorithm has a high prediction accuracy with good tolerance for abnormal value and noise.

The RF classification algorithm is used in two phases. First, the RF algorithm extracts subsamples from the original samples using the bootstrap resampling method and creates decision trees for each sample. Second, the algorithm classifies the decision trees and implements a simple vote, with the largest vote of the classification as the final result of the prediction. The RF algorithm always includes three steps as follows:

(1) Select the training set. Use the bootstrap random sampling method to retrieve K training sets from the original dataset (M properties), with the size of each training set the same as that of the original training set.

(2) Build the RF model. Create a classification regression tree for each of the bootstrap training sets to produce K decision trees to form a "forest"; these trees are not pruned. Looking at the growth of each tree, this approach does not choose the best features as internal nodes for branches but rather the branching process is a random selection of m≤M of all features.

(3) Create simple voting. Since the training process of each decision tree is independent, the training of the random forests can work in parallel, which significantly improves efficiency. The RF can be created by combining K decision trees trained in the same way. When classifying the input samples, the results depend on the simple voting of the output of each decision tree. The RF algorithm determines the samples by constructing a series of independent and distributed decision trees and determines the final category of the sample according to each decision tree.

RF has been used in a wide range of fields in recent years. Malek et al. [14] combined RF and self-organizing maps to predict pediatric fracture healing time effectively. Wang [15] applied RF to the condition monitoring and fault diagnosis of manufacturing and proposed a panoramic crack detection method based on structured RF. Malazi and Davari [16] used RF and emerging pattern algorithms to identify the complex activities of elders at home and the performance reached a high degree of accuracy with the F-measure index. Santana et al. [17] quantified the quality soil parameters based on the multivariable regression of RF, making it possible to develop a fast and automatic analysis process. Anitha and Siva [18] proposed a new computer-aided brain tumor detection method using the RF classifier.

In addition, many scholars have optimized the RF algorithm to improve the predictive ability of specific data sets. Hu et al. [19] proposed the class incremental RF (CIRF) to solve the identification problem of new activities. Abellan et al. [20] proposed a random trust RF (RTRF), which shows better performance on the noise dataset. Gomes et al. [21] proposed an adaptive RF algorithm (ARF) to classify the data stream. Genuer et al. [22] put forward an RF algorithm that would be suitable for big data analysis that solved parallel computing, online modification, and out-of-bag error problems. Zhu et al. [23] proposed the class weights RF algorithm to solve the medical class imbalance problem in data analysis.

Based on employee turnover data with high-dimensional unbalanced characteristics, this study puts forward the WQRF algorithm, which uses the RF algorithm in two phases. First, the algorithm is used to rank each feature's importance and reduce the dimensions. Second, the selected features are used with the RF algorithm and the performance evaluation results are calculated for the F-measure value of each tree to establish the different weights of the trees to generate the prediction model.

Table 1: Confusion matrix of employee turnover prediction.

		Prediction	
		Turnover	No turnover
Actual	Turnover	TP	FN
	No turnover	FP	TN

3.2. Classifier Evaluation Index. The common evaluation indices for the prediction model's performance are accuracy (ACC), recall, precision (PPV), and the area under the curve (AUC). To calculate these indices, the confusion matrix is used. In the matrix, the columns represent the prediction categories and the sum of the value in the column is the data observations in the category. In addition, the rows in the matrix represent the actual categories and the sum of the values in the rows represents the data observations in that category. In this study, our focus is on whether or not there is employee turnover, which is a binary classification.

Turnover is set as the positive category and no turnover set as the negative category. As shown in Table 1, TP denotes that the actual turnover is predicted as turnover; FN denotes that the actual turnover is predicted as no turnover; TN denotes that actual no turnover is predicted as no turnover and FP denotes that actual no turnover is predicted as turnover.

Recall denotes the true positive rate (TPR) and the equation is

$$Recall = TPR = \frac{TP}{(TP + FN)} \tag{1}$$

FPR denotes the false positive rate and the equation is

$$FPR = \frac{FP}{(FP + TN)} \tag{2}$$

Precision denotes the positive predictive value (PPV) and the equation is

$$PPV = \frac{TP}{(TP + FP)} \tag{3}$$

ACC denotes accuracy and the equation is

$$ACC = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{4}$$

The AUC denotes the area under the receiver operating characteristics curve (ROC). It is an important index for judging the advantage and disadvantage of a binary prediction model. If its value is bigger, the performance of the model is better. The differences in the points of the ROC reflect the different responses to the same signal stimulation. In addition, the x-coordinate of the ROC curve is FPR and the y-coordinate is recall.

In this study, we pay attention to the small classes (categorization features as turnover) in the problem of unbalanced

classification. The main goal is to avoid misdiagnosis and minimize misdetection. Therefore, recall, the F-measure, the AUC, and the overall ACC will be used to evaluate the performance of the classification algorithm.

3.3. Features Ranking. There are many employee features within an organization and they can be divided into several general categories: basic information, job information, job position information, education experience, training experience, qualification certification, work experience, reward and punishment, family background, compensation and benefits, performance appraisals, resident region, evaluation information, attendance information, and economic conditions. In addition, different countries, different industries, and different company scale may affect this collection of features.

If all features (dimensions) are taken into account, this would result in a significant cost in terms of time and space in the application of the algorithm, which then has a serious impact on its performance. Therefore, the RF method is first used to reduce the dimensions.

The RF method can rank the features (variables) based on their importance, so it can be used to reduce dimensions and delete less important features. The core idea is to calculate the degree of decreasing accuracy of the RF prediction by adding noise to each feature. The importance calculation for each feature X in the RF algorithm is shown as follows:

(1) For each decision-tree in the RF, use the corresponding out-of-bag (OOB) data to calculate the OOB error marked as errOOB1. Here, the OOB data refer to the rest of the data, which are never used in the process of training the decision trees. In addition, the OOB data are about one-third of all the data. Moreover, the OOB data can be used to evaluate the performance of the decision tree and the prediction error rate of this calculation is the OOB error.

(2) Add noise interference into the feature X for all OOB data randomly; then calculate the OOB error again as errOOB2.

(3) Assume that there are N trees in the RF and the important score of feature X can be calculated by the following equation:

$$I_X = \frac{\sum_{j=1}^{N} \left(errOOB2_j - errOOB1_j\right)}{N} \tag{5}$$

The importance value of the expression can be used as a measurement because the noise is randomly added to the feature and the accuracy of the outer bag is reduced significantly. This indicates that this feature has a strong influence on the classification results of the sample; namely, it has a higher degree of importance.

The objective of the feature selection is (1) to improve the accuracy of model prediction, (2) to construct a faster and lower consumption model, and (3) to make the model more explanatory. The approach is (1) to find the characteristic variables that are highly correlated with the target variable and (2) select the characteristic variables that have fewer numbers and can fully predict the value of the target variable. The steps of feature selection using RF are as follows.

Step 1. Use the current feature set to establish the model by applying the RF algorithm.

Step 2. Calculate the I value (the importance score of the variables) for each feature in the current feature set and sort each variable in descending order according to I value.

Step 3. Determine the deletion ratio and remove the unimportant index from the current feature variable to obtain a new feature set.

Step 4. Repeat Steps 1 to 3 until the remaining m features (predefined) and the feature set consisting of m features use the last selected feature set.

This study uses the RF ranking method to measure the features of the employees and to reduce the dimensions of employee turnover prediction. Maintaining the m of the most important features of employee turnover will increase the efficiency of the prediction.

3.4. The Weighted Random Forest Algorithm. Commonly, all decision trees of the RF have the same weight value while voting for the classification. However, it has a fatal defect when used with the unbalanced data classification prediction. To solve this, we introduce the weighted F-measure into the RF algorithm, which generates a better performance for employee turnover prediction by assigning different weights to different decision trees. From the viewpoint of data mining, the problem in employee turnover predictions is the binary classification of the unbalanced data.

We set "turnover" as the positive category, while "not leaving" as the negative category. It is obvious that the positive category is minor and the negative category is the major category. It is not sufficient to measure the performance of the model for accuracy with unbalanced data; for example, if a company's employee turnover rate is 2% and no one is expected to leave, the accuracy rate could be as high as 98%; however, this does not make sense here.

Since we focus on minor categories, even if all the minor categories are falsely classified as major categories, the accuracy is still very high but the model has no value for employee-turnover research. In this study, it is better to misjudge an employee who has no intention of leaving as a possible departure than to overlook a person with a genuine intention of leaving. This research combines the two evaluation indices of precision and recall and uses the harmonic mean of the F-measure to evaluate the performance of each decision tree and calculate the weight of the vote. Compared with the common RF algorithm, this updated algorithm improves the performance of the unbalanced data classification.

The F-measure here refers to F1 (namely, α is set as one) and its formula is as shown in (6). The F-measure combines the results of precision and recall and the higher the F-measure, the better the classification performance. The

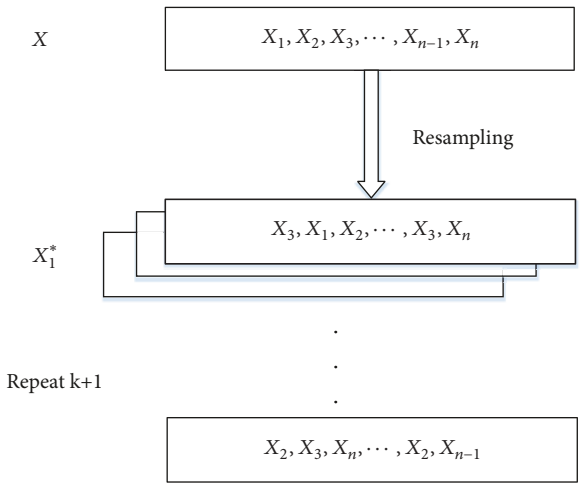

Figure 1: Schematic diagram of bootstrap resampling method.

precision rate and recall were discussed above under the Classifier Evaluation Index section.

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

The algorithm follows the steps as follows.

Step 1. Confirm the training set, validation set, and test set. Randomly extract K+1 datasets from the original dataset, with known categories, by the bootstrap method. The capacity of each dataset is n, namely, the same as the original dataset. Among the K+1 datasets, K sets are used as training sets and the remaining set is used as the validation set. The sample that has not been drawn represents about a third of the total sample of the original training set and constitutes the test set. The details of the bootstrap reextraction method are shown in Figure 1.

Step 2. Construct the RF classifier. Input K training sets and use the RF algorithm to build the model; apply the combination classifier composed of K classification decision trees.

Step 3. Get the weight value by calculating the F-measure of the subclassifier. Input the validation set and classify each sample in the validation set by regarding each decision tree in the forest as an independent classifier. Then get the TP, FP, FN, TN, precision rate, and recall rate values of each classifier. Next, calculate the F-measure, corresponding to the weight of each subclassifier. The weight of the subclassifier for j is as shown in (7).

Step 4. Input the test set to evaluate the performance of the model.

Step 5. Input the unclassified samples. Classify the samples by the F-measure weighted random forest. The result H depends on the weighted vote of the classification results of each subclassifier. The classification result of subclassifier j is $h_j(x)$ and the classification weight result of subclassifier j is W_j.

The final classification decision can be expressed as (8)

$$W_j = F_j = \frac{2TP_j}{2TP_j + FP_j + FN_j} \quad (7)$$

$$H(x) = \arg\max_Y \sum_{j=1}^{K} W_j I\left(h_j(x) = Y\right) \quad (8)$$

In (8), H(x) represents the combined classification model obtained by the weighted RF algorithm. h_j is the subclassifier (i.e., single decision tree), Y represents the output variables (i.e., the classification type), and function I( ) is the indicator function.

In the employee turnover prediction question, parameter Y has two options: leaving and not leaving. Therefore, in (8), when Y denotes leaving, all the weighted values of the subclassifiers, classified as turnover, will be added together as the score of H(x). On the other hand, when Y denotes not leaving, all the weighted values of the subclassifiers, classified as not leaving, will be added together as the score of H(x). The comparison of the two scores and the corresponding value of Y of the max values of the two scores represent the prediction classification values of the combined classification model. Figure 2 presents the structure of the weighted random forest.

3.5. The WQRF Method Overview. Here, by introducing the combination forecasting theory into this field of data mining, the current employee data (leaving and not leaving employees and the leaving status labels) are trained and modeled to predict whether an employee will leave the job in the future.

The WQRF algorithm is proposed to build the prediction model. The original data (historical) are divided into the training set and the validation set randomly. In addition, the not selected data (OOB) are used as the test set. In the first calculation, the training set is used to rank the features by importance applying the RF algorithm. The m of the most important features for employee turnover are selected. In the second calculation, these m features are included in the weighted RF algorithm with the training set to build the prediction model of employee turnover. The weights in the weighted RF algorithm are obtained by using the validation set to calculate the F-measure of each tree. For any employee, the m features are obtained and input into the prediction model, so the intention of the employee to depart in the future can be predicted. The prediction model can be evaluated by a 10-fold cross validation method to obtain the accuracy, sensitivity, precision, and AUC. In the next section, the experiment will validate the performance of the algorithm proposed here; it will show that it is actually better than common predictive algorithms such as the RF, the decision

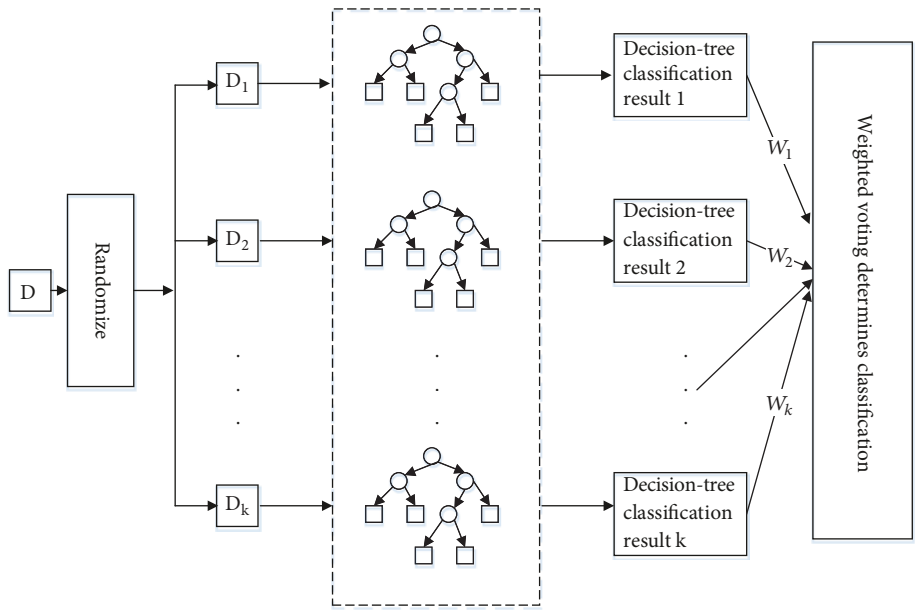

Figure 2: Structure of weighted random forest.

tree (e.g., the C4.5 algorithm), the Logistic regression (Logistic), and back propagation (BP) neural network. Figure 3 is a schematic diagram of the WQRF algorithm.

4. Experiments

4.1. Data Overview. We use the real dataset of employees of a branch of a communications company in China in our experiment. There are 2000 employees with 32 features in the dataset. Among these, 271 employees are marked as turnover, meaning they are marked as "yes" in terms of the "attrition" features. The proportion of turnover is 13.5% and obviously represents an unbalanced data problem.

Here, we select the most common and public employee features and list these 32 features and their ranges in Table 2.

In the experiment, we use Python with NumPy, SciPy, SciKit-learn, Pandas, and Matplotlib and develop three tools: one tool is for feature ranking based on the RF, one is a visual tool for analysis of the feature variables and the target variable, and one is a modeling tool for the RF based on the weighted F-measure.

4.2. Feature Selection. The first step in the WQRF here is to apply RF for feature selection. The feature of "EmployeeNumber" is removed from the 32 features because it is not related to the target variable. In order to call the RF algorithm in Python, all features are converted to float values.

Using the tool for feature ranking based on the RF, we input the employee data into 2000 rows with 30 independent features and one dependent variable. According to their importance, we rank the 30 features in descending order as shown in Table 3.

After removing the lower scoring features one by one, we find that the accuracy, F-measure, and AUC perform better when only the top 15 features are retained.

Of the 30 features, 15 are removed and the data are reorganized into 2000 rows, 15 independent features, and 1 target variable. The 15 features based on importance are selected as the features of the employee turnover prediction problem and the operation dimensionality are reduced significantly.

Using the visual tool to examine the relation between the features and the target variable, we analyze six features scored earlier and find some interesting results. Figure 4 shows the relation between overtime and attrition and reveals that as overtime increases so does turnover. Figure 5 shows the association among age, years at the company, and attrition. It indicates that the highest turnover is associated with employees who have worked for a relatively few years at the company and are younger. Figures 6 and 7 show the associations between monthly income and attrition, and among percent of salary increase, distance from home, and attrition, respectively. In the figures, turnover is identified in orange.

4.3. Weighted Random Forest Forecast. The second step in the WQRF algorithm is to add weight to the simple voting mechanism of the RF; this will achieve a better classification prediction for employee turnover issues. In order to evaluate the proposed WQRF algorithm based on the weighted F-measure, we apply it to the employee dataset and we set the value of "Attrition" to mean the status of turnover, where "yes" means "leave" (turnover) and "no" means "not leave" (no turnover).

By random sampling with replacement, we obtain 101 datasets. The capacity of each dataset is consistent with the original dataset and the count is 2000 rows. We select 100 as

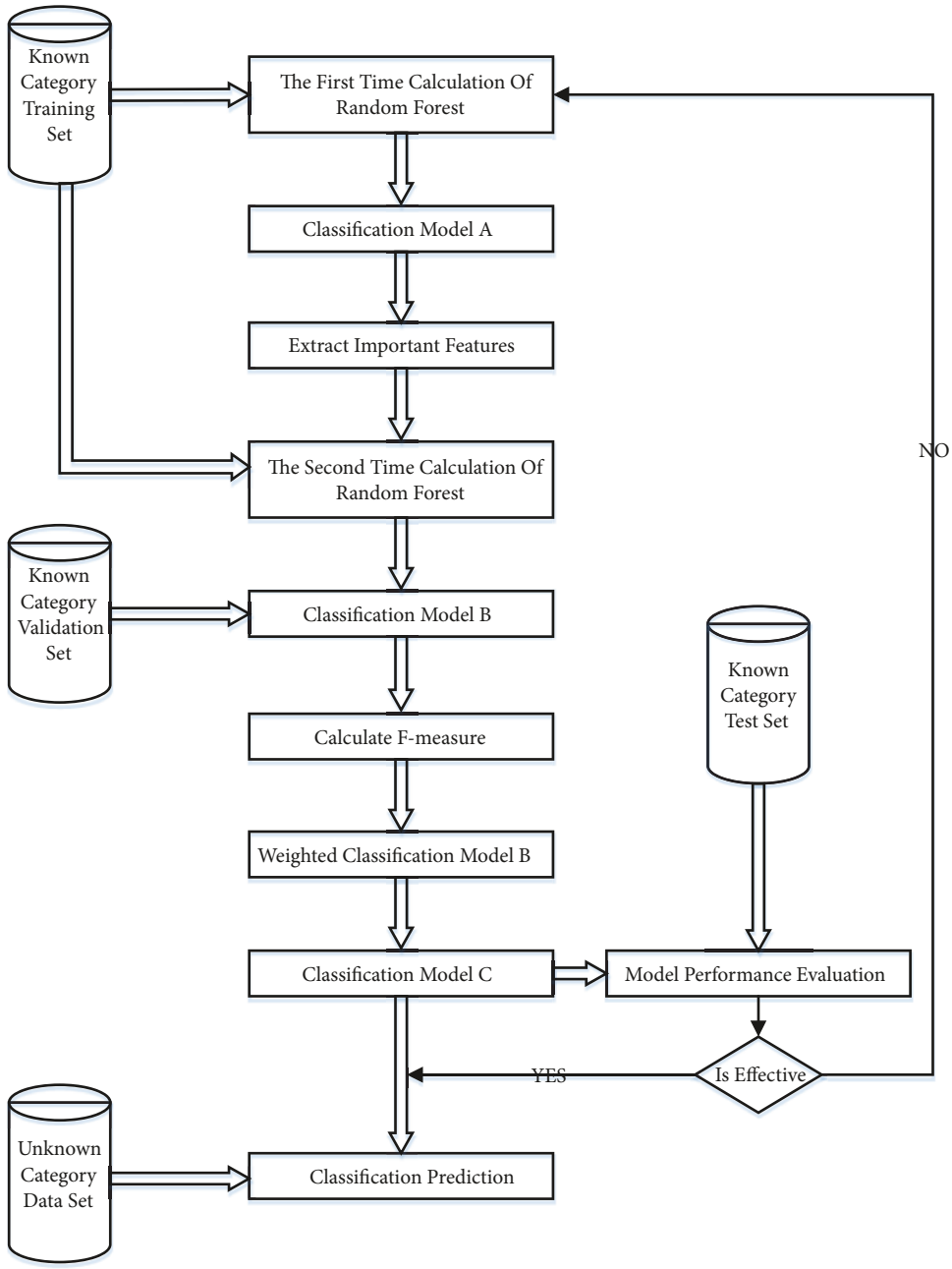

Figure 3: Flowchart of the weighted quadratic random forest algorithm.

the training set and to model the RF algorithm; we use the rest as the validation set. In this model, we get 100 F-measure values, which come from classifying the category features to yes. We add these F-measures to the model as weighted values, with the higher value of the F-measure representing more voting weight. In this way, we get a new model. Next, we get the performance index of the new model by using the 10-fold cross-validation method.

The WQRF based on the weighted F-measure is compared with the RF, the C4.5 [24], the Logistic regression, and the BP neural network; the result is that it shows better prediction performance. The evaluation index for the experiment is shown in Table 4; the 10-fold cross-validation method to predict the confusion matrix in Table 5; and the ROC curve in Figure 8.

In Table 4, we can see that the WQRF algorithm shows better performance for the indicators Recall, F-measure, ROC Area, and ACC than the algorithms RF, C4.5, Logistic, and BP. Although the comparison with RF for ACC is not obvious, the increase in the WQRF ACC value still shows better overall performance.

Table 5 is the obfuscation matrix for the prediction results of the five algorithms using the 10-fold cross-validation method. In the dataset of 271 turnover employees, the RF

Table 2: Value ranges of 32 features.

No.	Feature	Value Range
1	Attrition (class name)	Yes, No
2	EmployeeNumber	0~2000
3	Age	22~60
4	Gender	Female, Male
5	EmploymentNature	Regular Worker, Dispatched Worker
6	JobLevel	(high) 4~13 (low)
7	JobRole	Staff, Junior Management, Middle Management, Senior Management
8	MonthlyIncome	5000~97938
9	DepartmentType	Sales, Management, Technical
10	DistanceFromHome	1~20
11	Education	Academy, Bachelor, Master, Doctor
12	EducationField	Technology, Management, Economics, Other
13	RelationshipSatisfaction	Low, Medium, High, Very High
14	WorkLifeBalance	Bad, Good
15	EnvironmentSatisfaction	Low, Medium, High, Very High
16	JobSatisfaction	Low, Medium, High, Very High
17	OverTime	Yes, No
18	AvgWorkHours	8~16
19	MaritalStatus	Yes, No
20	HaveChildren	Yes, No
21	NumberCompaniesWorked	0~3
22	TotalWorkingYears	0~37
23	YearsatCompany	0~36
24	YearsinCurrentRole	0~14
25	YearswithCurrentManager	0~12
26	WinningCount	0~5
27	PerformanceRatingLastYear	Low, Good, Excellent, Outstanding
28	PercentSalaryIncrease	0~30
29	YearsSinceLastPromotion	0~32
30	TrainingTimesLastYear	0~5
31	NativePlace	Local, Nonlocal
32	PhysicalCondition	Healthy, Unhealthy

Table 3: Importance score of 30 features.

No.	Feature	Score	No.	Feature	Score
1	MonthlyIncome	0.2816	16	EducationField	0.0137
2	OverTime	0.2762	17	WinningCount	0.0100
3	Age	0.0665	18	Gender	0.0093
4	DistanceFromHome	0.0431	19	NumberCompaniesWorked	0.0073
5	YearsatCompany	0.0317	20	HaveChildren	0.0069
6	PercentSalaryIncrease	0.0306	21	EnvironmentSatisfaction	0.0065
7	YearsinCurrentRole	0.0289	22	RelationshipSatisfaction	0.0062
8	TrainingTimesLastYear	0.0257	23	JobSatisfaction	0.0060
9	YearsSinceLastPromotion	0.0250	24	EmploymentNature	0.0056
10	YearswithCurrentManager	0.0218	25	MaritalStatus	0.0036
11	AvgWorkHours	0.0211	26	PerformanceRatingLastYear	0.0031
12	TotalWorkingYears	0.0208	27	WorkLifeBalance	0.0015
13	JobLevel	0.0173	28	PhysicalCondition	0.0007
14	Education	0.0141	29	JobRole	0.0007
15	DepartmentType	0.0140	30	NativePlace	0.0004

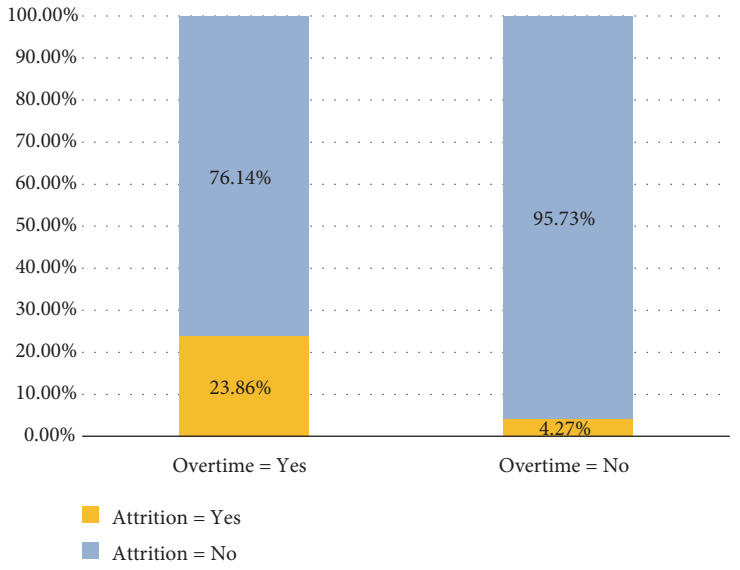

Figure 4: Relationship between overtime and attrition.

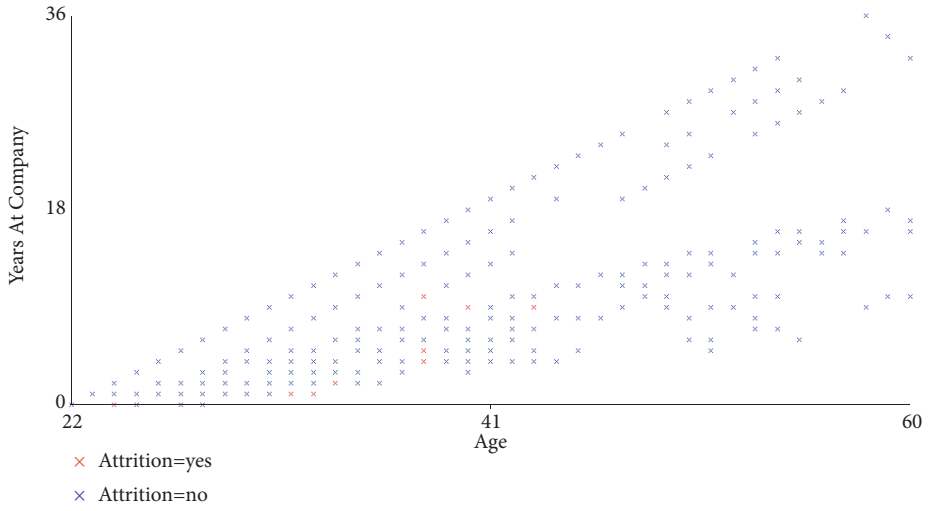

Figure 5: Relationship among age, years at the company, and attrition.

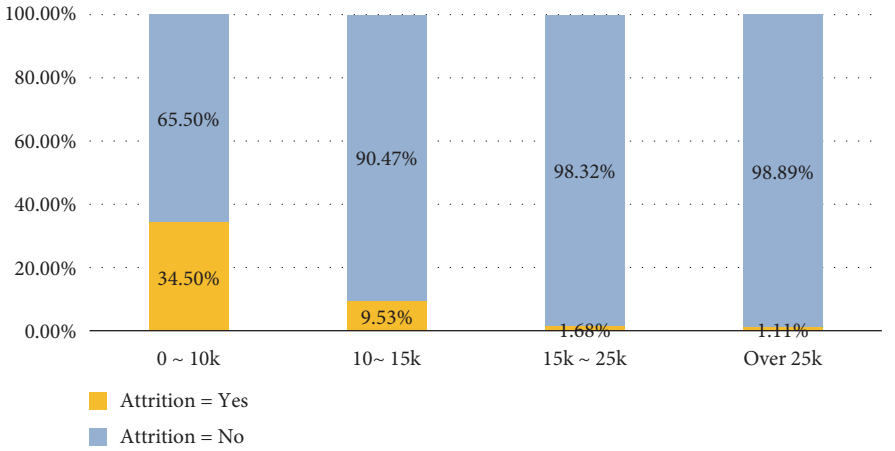

Figure 6: Relationship between monthly income and attrition.

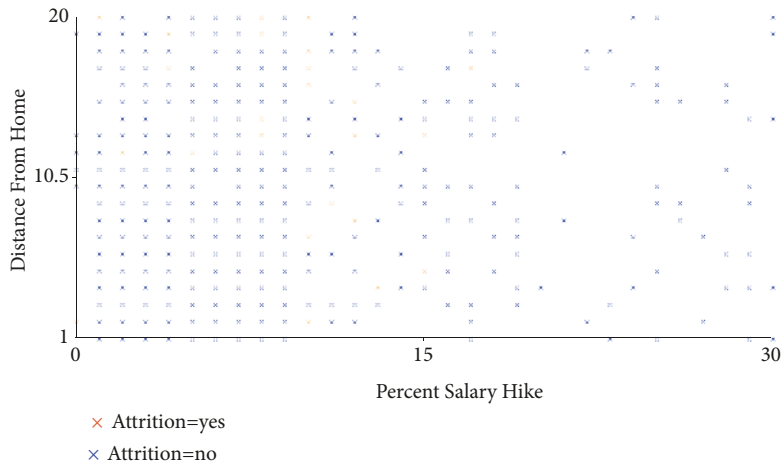

Figure 7: Relationship among percent of salary increase, distance from home, and attrition.

Table 4: Experimental results of different algorithms.

Algorithms	Recall	F-measure	ROC Area	ACC (%)
RF	0.627	0.698	0.850	92.65
C4.5	0.561	0.564	0.773	91.05
Logistic	0.469	0.259	0.807	90.20
BP	0.502	0.560	0.781	89.30
WQRF	0.653	0.711	0.881	92.80

Table 5: Prediction results for 10-fold cross-validation with employee dataset.

Algorithms	a	b	Classified as
RF	1683	46	a=No
	101	170	b=Yes
C4.5	1669	60	a=No
	119	152	b=Yes
Logistic	1677	52	a=No
	144	127	b=Yes
BP	1650	79	a=No
	135	136	b=Yes
WQRF	1679	50	a=No
	94	177	b=Yes

algorithm predicts 170 people correctly, the C4.5 algorithm predicts 152 people correctly, the Logistic algorithm predicts 127 people correctly, and the BP predicts 136 people. However, our WQRF algorithm predicts 177 people correctly, which is the best out of the five algorithms.

As can be seen in Figure 8, the ROC curve of the WQRF algorithm is closer to the upper left corner and the C4.5 is closer to the right. WQRF has a larger AUC value than the other algorithms, which implies better predictive ability. In sum, for the real employee dataset, the experiment proves that WQRF has a better ability to predict employee turnover than RF, C4.5, Logistic, and BP.

5. Conclusion and Future Work

In this study, an improved RF algorithm, the WQRF based on the weighted F-measure, is proposed. The main idea is to follow two steps. First, the random forest algorithm is used to order feature importance and reduce dimensions. Second, the selected features are used with the random forest algorithm and the F-measure values are calculated for each decision tree as weights to build the prediction model. In the area of employee turnover forecasting, compared with the RF, C4.5, Logistic, and BP, the WQRF algorithm shows significant improvement in various performance indicators, especially recall and F-measure. In the employee dataset of a branch of a communications company in China, the key influencing factors for employee turnover were found using the WQRF algorithm; these were monthly income, overtime, age, distance from home, years at the company, and percent of salary increase, with the monthly income as the most important factor. Therefore, increasing employee income is undoubtedly one of the most effective ways to retain employees.

The algorithm proposed here has also been applied in the prediction of employee turnover in industries such as education, medical, finance, and other fields. Different enterprise employees will have some similar characteristics but also some special attributes, which does not affect the use of this algorithm.

If an organization can predict in advance which employees may turnover, it can develop a plan and take measures to reduce this possibility, address the need to hire replacements quickly, and make other adjustments to retain employees in key positions. Using the WQRF approach, HR administrators can predict better employee turnover and take timely action.

This algorithm can be applied as well to the problem of unbalanced data classification in other fields, for example, in predictions of customer churn, cancer screening, and abnormal testing. These problems have the same characteristics, such as a focus on a minority class, and the requirement that it is "better to mischeck than to misdetect." Compared with other algorithms, generally, the WQRF algorithm has a

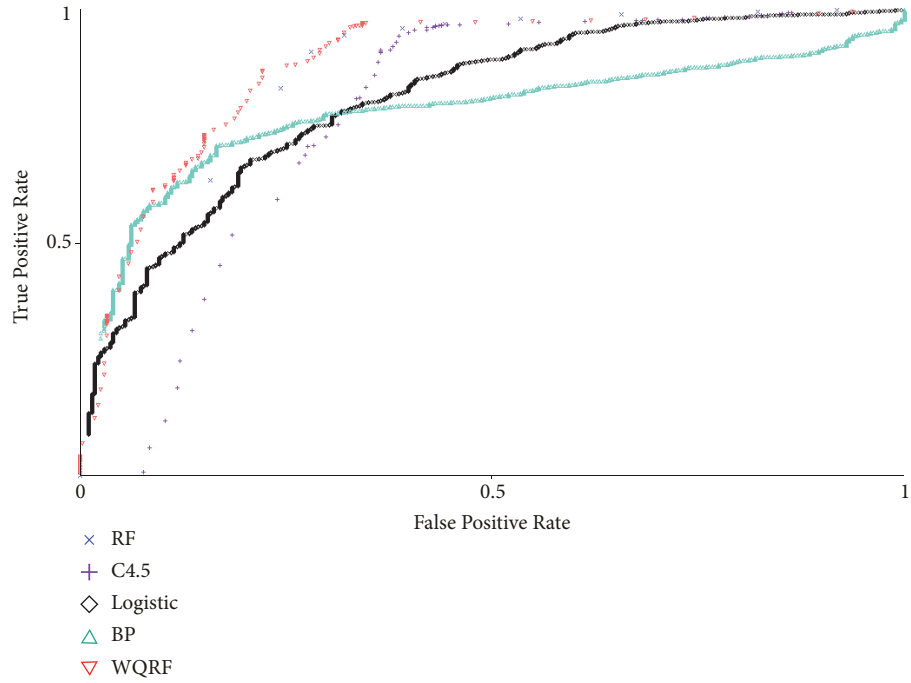

Figure 8: ROC curves of five different algorithms.

higher recall rate for a minority class. In the future, we will theoretically analyze the data characteristics in these fields and further verify and optimize the WQRF.

In terms of study limitations, adding the feature sorting and weight calculation processes does mean that the modeling has a higher cost in terms of time than other algorithms. In future research, improvement in operational efficiency and overall prediction accuracy could be analyzed. In addition, this algorithm is designed for unbalanced data and not suitable for industries with high turnover rate. How to increase the universality of the algorithm still needs to be studied further.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China (Grant no. 61672117).

References

[1] A. Denvir and F. McMahon, "Labour turnover in London hotels and the cost effectiveness of preventative measures," *International Journal of Hospitality Management*, vol. 11, no. 2, pp. 143–154, 1992.

[2] T. H. Feeley, J. Hwang, and G. A. Barnett, "Predicting employee turnover from friendship networks," *Journal of Applied Communication Research*, vol. 36, no. 1, pp. 56–73, 2008.

[3] P. A. Gloor, A. Fronzetti Colladon, F. Grippa, and G. Giacomelli, "Forecasting managerial turnover through e-mail based social network analysis," *Computers in Human Behavior*, vol. 71, pp. 343–352, 2017.

[4] W. C. Hong, P.-F. Pai, -Y. Huang, and -L. Yang, "Application of support vector machines in predicting employee turnover based on job performance," in *Proceedings of the International Conference on Advances in Natural Computation*, pp. 668–674, Springer, Berlin, Germany, 2005.

[5] H.-W. Kao, S.-W. Lin, and S.-Y. Wan, "Applying decision tree to predict nursing turnover-a case study in a public hospital," *The Journal of Taiwan Association for Medical Informatics*, vol. 21, no. 4, pp. 15–29, 2012.

[6] H.-J. Chien, *Application of the two-stage cluster analysis on employee voluntary turnover intention [M.S. thesis]*, Yuan Ze University, 2007.

[7] C.-T. Wu, *Using the decision tree approach to forecast contractor turn over trend – a case study of wafer foundry [M.S. thesis]*, National Cheng Kung University, 2008.

[8] T. R. Mitchell, B. C. Holtom, T. W. Lee, C. J. Sablynski, and M. Erez, "Why people stay: using job embeddedness to predict voluntary turnover," *Academy of Management Journal*, vol. 44, no. 6, pp. 1102–1121, 2001.

[9] B. C. Holtom, D. R. Smith, D. R. Lindsay, and J. P. Burton, "The relative strength of job attitudes and job embeddedness

in predicting turnover in a U.S. military academy," *Military Psychology*, vol. 26, no. 5-6, pp. 397–408, 2017.

[10] H. S. Jung and H. H. Yoon, "Error management culture and turnover intent among food and beverage employees in deluxe hotels: the mediating effect of job satisfaction," *Service Business*, vol. 11, no. 4, pp. 785–802, 2017.

[11] L. J. Labrague, D. Gloe, D. M. McEnroe, K. Konstantinos, and P. Colet, "Factors influencing turnover intention among registered nurses in Samar Philippines," *Applied Nursing Research*, vol. 39, pp. 200–206, 2018.

[12] H. Tran, "The impact of pay satisfaction and school achievement on high school principals' turnover intentions," *Educational Management Administration and Leadership*, vol. 45, no. 4, pp. 279–290, 2016.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] S. Malek, R. Gunalan, S. Kedija et al., "Random forest and Self Organizing Maps application for analysis of pediatric fracture healing time of the lower limb," *Neurocomputing*, vol. 272, pp. 55–62, 2018.

[15] S. Wang, X. Liu, T. Yang, and X. Wu, "Panoramic crack detection for steel beam based on structured random forests," *IEEE Access*, vol. 6, pp. 16432–16444, 2018.

[16] H. T. Malazi and M. Davari, "Combining emerging patterns with random forest for complex activity recognition in smart homes," *Applied Intelligence*, vol. 48, no. 2, pp. 315–330, 2018.

[17] F. B. de Santana, A. M. de Souza, and R. J. Poppi, "Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 191, pp. 454–462, 2018.

[18] R. Anitha and S. R. D. Siva, "Development of computer-aided approach for brain tumor detection using random forest classifier," *International Journal of Imaging Systems and Technology*, vol. 28, no. 1, pp. 48–53, 2018.

[19] C. Hu, Y. Chen, L. Hu, and X. Peng, "A novel random forests based class incremental learning method for activity recognition," *Pattern Recognition*, vol. 78, pp. 277–290, 2018.

[20] J. Abellán, C. J. Mantas, J. G. Castellano, and S. Moral-García, "Increasing diversity in random forest learning algorithm via imprecise probabilities," *Expert Systems with Applications*, vol. 97, pp. 228–243, 2018.

[21] H. M. Gomes, A. Bifet, J. Read et al., "Adaptive random forests for evolving data stream classification," *Machine Learning*, vol. 106, no. 9-10, pp. 1469–1495, 2017.

[22] R. Genuer, J. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, "Random forests for big data," *Big Data Research*, vol. 9, pp. 28–46, 2017.

[23] M. Zhu, J. Xia, X. Jin et al., "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[24] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994.