A Model for Analyzing Employee Turnover in Enterprises Based on Improved XGBoost Algorithm

Linzhi Nan[1], Han Zhang[2*]
School of Economic and Trade, Haojing College of Shaanxi University of Science and Technology, Xi'an, 710000, China[1]
Shaanxi Hantong Consulting Services Co., LTD, Xianyang, 712000, China[2]

***Abstract*—To accurately predict the possibility of employee turnover during enterprise operation and improve the benefits created by talents in the enterprise, research based on the limit gradient enhancement algorithm has received widespread attention. However, with the exponential growth of various types of resignation reasons, this algorithm is not comprehensive enough when dealing with complex character psychology. To solve this problem, this study uses the limit gradient enhancement algorithm to predict employee turnover in the Company dataset, and uses differential automatic regression moving average variable optimization to generate a fusion algorithm. The research first involves stepwise regression processing of the training data, expanding the objective function to a second-order Taylor expansion; Then variance coding is added to the square integrable linear white noise, and the step cooling curve is smoothed by changing the temperature control constant; Then to calculate the root mean square error of Newton's law of cooling, and obtain its derivative loss variable. Linear white noise is the chaotic data produced by the improved extreme gradient lifting algorithm in forecasting the original data of enterprise employees, which will affect the results of data preprocessing in the loss analysis. In order to reduce the operation error of the algorithm, the step cooling curves are drawn according to the cooling law, and then their root mean square errors are calculated. Finally, the fusion algorithm studied was applied to the Company dataset and the prediction accuracy of the particle swarm optimization algorithm was tested and compared with the fusion algorithm. A total of 400 experiments were conducted, and the fusion algorithm achieved a prediction accuracy of 398 times, with an accuracy rate of 99.5%; The accuracy of particle swarm optimization algorithm is close to that of fusion algorithm, at 83.2%. The experimental results indicate that the algorithm model proposed in the study can accurately predict the possibility of employee turnover in enterprises, and the company will also receive timely information to make the next budget step.**

***Keywords*—Data preprocessing; linear white noise; root mean square error; newton's law of cooling; step cooling curve**

I. INTRODUCTION

In the rapidly advancing society of internet technology, as competition between companies intensifies, the number of employees in enterprises is gradually receiving widespread attention, and the requirements for algorithms are also increasing [1-2]. In the operation of the enterprise, talents rely on interpersonal relationships and key technologies to continuously create profits for the enterprise. Such employees will be widely valued within the company, but leaving is also inevitable. The resignation of employees is not conducive to the development of the enterprise, and may even lead to technical gaps that can lead to internal management problems. So, corporate executives are gradually paying attention to the reasons for employee turnover. In this context, algorithms for analyzing employee psychology have received widespread attention. In recent years, the Limit Gradient Lifting Algorithm (XGBoost) has attracted the attention of many scholars due to its powerful self-checking ability [3]. However, XGBoost is only suitable for analyzing differentiable loss variables. For non-differentiable variables, the algorithm marks them as isolated points, which affects the accuracy of the detection data. To address this issue, this study is based on Differential Autoregressive Moving Average Variable (DAMAV) to optimize XGBoost and generate a fusion algorithm (DV-XGBoost) pioneering. This algorithm calculates the root mean square error of non-differentiable variables and can convert non differentiable variables into differentiable loss variables. For the experimental design of the analysis model of employee turnover in enterprises, the research first collects the human resources data of enterprises, and determines the completeness of employee turnover prediction according to the relevant experience of employee turnover in the past; Then clean the collected data to ensure the availability of the data, and extract the characteristics of the data according to the business needs of the enterprise. For the key characteristics of employee turnover, DV-XGBoost algorithm is used to judge it, and the data is divided into training set and test set. The evaluation indexes include accuracy, accuracy and F_1 value. The innovation of this study is mainly reflected in the following two points. First, aiming at the limitations of XGBoost algorithm, this study proposed a fusion algorithm based on DAMAV to optimize XGBoost (DV-XGBoost). This fusion algorithm shows its unique innovation when dealing with underivable variables. The second innovation is that for the non-derivable variable, the traditional XGBoost algorithm will mark it as an isolated point, while the DV-XGBoost algorithm converts it into a derivable loss variable by calculating the root-mean-square error of the non-derivable variable. This method is innovative in solving the problem of non-derivable variables. The contribution of this research is mainly reflected in the following three points. First, DV-XGBoost algorithm improves the accuracy of detection data by converting non-derivable variables into derivable loss variables. It is of great practical significance for enterprises to analyze the reasons of employee dimission and then make corresponding strategies. The second is to broaden the application range of XGBoost, a powerful algorithm that is widely used in all kinds of data contests and real-world

problem solving. However, it is only suitable for analyzing derivable loss variables; DV-XGBoost algorithm further broadens the application range of XGBoost by solving the problem of non-derivable variables. Finally, it can promote the progress of talent management, and enterprises can better understand the reasons for employees' dimission, so as to improve the company's talent management strategy, reduce the employee dimission rate, improve employee satisfaction, and promote the stable development of enterprises.

The research is mainly divided into six sections. Section I mainly analyzes and summarizes the application and effectiveness of the current XGBoost model. Section II introduces the factors that affect the employee turnover model and constructs the DV-XGBoost model. Section III is the experimental study on enterprise characters. Result and discussion is mentioned in Section V and Section VI concluded the paper.

II. Related Works

A very important branch of human psychological prediction technology, namely employee turnover prediction in enterprises, plays a very important role in computer learning and the rational use of human resources in enterprises [4]. Lu et al. designed real driving tasks to extract data and proposed a stress monitoring model based on driving behavior. Driving is described by the acceleration of the vehicle, and the driving environment is quantified using an extended residual network model. According to the distribution range of driver ambiguity, the video image is segmented into sub regions. They constructed an Extreme Gradient Enhancement (XGBoost) model to monitor stress, and compared it with other models, the XGBoost model outperformed mainstream learning algorithms. It can also surpass most traditional models without using psychological data [5]. Deng et al. combined XGBoost and multi-objective optimization genetic algorithm for cancer classification. They first sort genes based on XGBoost ensemble selection, effectively removing unrelated genes and generating the most relevant genome for this class; then, their fusion algorithm searches for the optimal subset based on gene groups. They conducted comprehensive experiments using learning classifiers on publicly available microarray datasets to compare state-of-the-art feature selection methods. Their experiments have shown that the algorithm outperforms existing algorithms in multiple evaluation indicators such as accuracy, precision, and recall [6]. Li et al. developed a reliable prediction method that estimates the sink area of the road surface based on the main feature of road surface temperature. They proposed chaotic particle swarm optimization and segmented regression strategy to optimize the XGBoost model. Compared with the classical learning algorithm, their experimental results show that the root mean square error and absolute error of XGBoost algorithm are increased by 5.80 and 1.59 respectively. This algorithm has obvious advantages in dealing with nonlinear problems, and can also reduce the frequency of deflection without affecting its estimation accuracy, promoting rapid evaluation of road conditions [7]. Tao et al. proposed a robust method for diagnosing turbine blade icing. They extracted features of short-term icing effects based on icing physics to establish stacked XGBoost models for blade icing diagnosis. They evaluated the methods proposed in the wind farm and further compared them with models based on a single algorithm. The results indicated that their mixed features enhance the similarity between different datasets, and their model has higher accuracy and better generalization power compared to models based on a single algorithm [8].

Li et al. proposed an orthopedic classification prediction model based on the XGBoost algorithm. After building the XGBoost model, they also built the same model based on the random forest algorithm, and made a comparative analysis of them. Compared with random forest model, XGBoost algorithm prediction model has higher accuracy and is more suitable for orthopedic clinical data. The XGBoost algorithm can handle diverse medical data and better meet the requirements of diagnostic timeliness and accuracy [9]. Gu et al. proposed a new LC decision model that enables automated vehicles to make human decisions. Their method combines a deep encoder network with the XGBoost algorithm, using time series from multiple sensors to establish a robust multivariate reconstruction model; then, they reconstructed the error using normal data for training data extraction. They adopted the XGBoost algorithm with Bayesian optimization for the multi-parameter problem of autonomous LC decision-making process. This model can accurately identify the LC behavior of vehicles [10]. Li et al. cross matched the bass with the spectral database of the Sloan Digital Telescope to obtain the spectral categories of known samples. Then, the samples were cross matched with the ALLWISE database, and they constructed different classifiers using the XGBoost algorithm based on the optical and infrared information of the samples. Finally, all selected items in the bass directory are classified by these classifiers. When the prediction results of binary classification are the same as those of multi class classification, the prediction results of light sources without infrared information can be used as a reference. Their classification results have great reference value for future research [11]. Osman et al. developed a model for predicting groundwater levels. They tested three machine learning models: Xgboost, artificial neural network, and support vector regression. The experiment shows that if the combination of rainfall data with a delay of 3 days is used as input, the performance of the model is the worst; For all input combinations, their proposed Xgboost model outperforms the other two models. When using groundwater level with a 1-day delay as input, the performance of the Xgboost model is significantly improved. Their research results provide application prospects for the Xgboost algorithm to predict future groundwater levels [12].

The research of multiple scholars mentioned above has found that the predictive ability of Xgboost algorithm is very popular internationally, but there is still little research on fusion algorithms. This study pioneered the introduction of differential autoregressive moving average variables, taking into account the impact of employee personal factors and internal control within the enterprise, and generated a fusion algorithm (DV-XGBoost).

III. Construction of Employee Turnover Information Model based on XGBoost

With the development of the human resources industry, in order to rationalize the distribution of employees in enterprises, research on various information has become increasingly popular, and the probability of employee transfer is the most important part [13]. However, objective factors such as the variety of resignation data and the complexity of employee psychology have increased the workload of predicting employee turnover information devices. This study combines the XGBoost algorithm with DAMAV, first introducing a model built on XGBoost, and then describing the fusion method of the two.

A. Establishment of XGBoost Model for Enterprise Employee Resignation

Before analyzing data on employee turnover in enterprises using models, it is first necessary to preprocess the data. Preprocessing is the most complex part of the experiment, and the results of this part not only occupy time, but also determine the predictive ability of the data. Real data often contains a large amount of data noise and data redundancy [14]. These artificially generated abnormal data are very detrimental to the results, so these data need to be cleaned. The preprocessing work for the data is Fig.1.

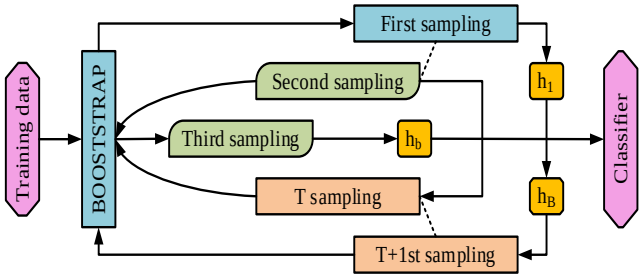


Fig. 1. Diagram of data preprocessing process.

From Fig. 1, the cleaning of training data is first run in the boot program, improving missing values and noise during each sampling process; Then oversampling and under sampling the data, and then transforming the data into a whole smooth curve to get the final output results. The variables selected through Fig. 1 can be used for stepwise regression, and the complex collinear data that meets the requirements is called biased estimation. The data that conforms to the characteristics of continuity can be obtained through a penalty function, as Formula (1).

$$p_\lambda\left(B_j\right)=\begin{cases}\lambda\left(B_j\right) & \left|B_j\right|\le\lambda \\ -\dfrac{\left(\left(B_j\right)^2-2a\lambda\left(B_j\right)+\lambda^2\right)}{2(a-1)} & \lambda\le\left|B_j\right|\le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & a\lambda\le\left|B_j\right|\end{cases} \tag{1}$$

In Formula (1), a is a constant with a value of, and the adjusted parameter is denoted as λ, with a range of non negative values. B_j is called effective data set, which is often based on natural logarithm [15]. At this point, the unit of random entropy is recorded as bits and is only related to B_j. As the value of B_j increases, the range of random entropy variables becomes larger. The relationship between them is Formula (2).

$$H\left(B_j\right)=-\sum_{i=1}^{n}p_i\log p_i \tag{2}$$

In Formula (2) above, p_i is a random variable with a value range over $[0,+\infty]$. The improvement of XGBoost algorithm relative to decision trees lies in gradient correlation, which is a running model that enhances decision-making ability. The range of regularization terms determines the complexity of the algorithm, and the objective function determines its optimal solution, as shown in formula (3).

$$\underline{y}=\sum_{t=1}^{T}f_t\left(x_i\right),\ f_t\in F \tag{3}$$

In Formula (3), the number of XGBoost decision trees is denoted as T, the decision forest is represented as F, and the specific XGBoost decision tree is denoted as f_t. If the previous prediction for round t is denoted as $\underline{y}$, then the objective function can be expanded using second-order Taylor expansion. In order to measure the quality of the decision tree, the scoring function of XGBoost is introduced, as listed in Formula (4).

$$Mark=0.5\left[\frac{G_L^2}{\alpha_L+\lambda}+\frac{G_R^2}{\alpha_R+\lambda}-\frac{\left(G_L+G_R\right)}{\alpha_L+\alpha_R+\lambda}\right]-\gamma \tag{4}$$

In Formula (4), G_L, G_R represent the segmentation points on the left and right sides of the mean, respectively; The mean values of multiple segmentation points are denoted as α_L, α_R; The difference in the parameters of the decision tree is called γ. The higher the value of $Mark$, the better the quality of the XGBoost decision model. The decision tree is continuously segmented according to this method, and the existing nodes in the time series can be predicted according to the past segmentation points and linear white noise, as expressed in Formula (5).

$$Y_t=\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+\ldots+\varphi_p y_{t-p}+u_t \tag{5}$$

In Formula (5), the predicted value of the existing nodes is Y_t, and the autoregressive coefficients between the segmentation points are represented by φ_p, with a range of positive integers of $(1,p)$. Linear white noise is recorded as u_t, and white noise sequence in regional time can be calculated by Formula (6) [16].

$$W_f\left(\phi,\chi\right)=\chi^{-0.5}\int_{-\infty}^{+\infty}\delta(t)\eta^*\left(\frac{t-\phi}{\chi}\right)dt=\left\langle\delta,\kappa_{(a,b)}(t)\right\rangle \tag{6}$$

In Formula (6), the shift factor and displacement factor are denoted as ϕ, χ, and their value ranges are non negative.

$\delta(t)$ represents a square integrable signal, complex conjugation is denoted as $*$, and $\kappa_{(a,b)}(t)$ is used to represent the cluster function [17]. When XGBoost calculates the importance of features, it studies the method of selecting and calculating the obtained values. By traversing all intermediate nodes, it maps the number of feature times of the branch. The trained tree model in XGBoost regression tree is exhibited in Fig. 2.

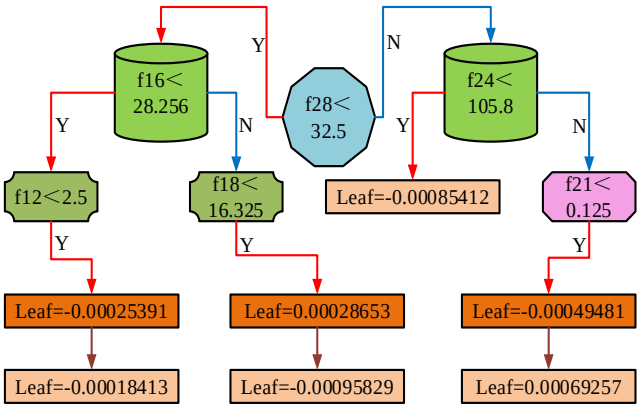


Fig. 2. XGBoost regression tree model trained in.

There are seven leaf nodes in Fig. 2, represented by rectangles, with the remaining shapes serving as attribute partitioning conditions. From Fig. 2, the complete XGBoost regression tree model has a total of four layers, including 13 nodes. The importance of customer data in the features can be intuitively seen in the figure, which provides guidance for the selection of later features. Dimension explosion is prone to occur during encoding extraction, leading to an increase in the working time and isolated points of the model. To avoid this issue, the study used variance encoding to change qualitative features by labeling samples, and the formula for calculating sample labels is Formula (7).

$$P\left(y=Y\middle|k=K\right)=\frac{N_Y^K}{N^K} \tag{7}$$

In Formula (7), $P\left(y=Y\middle|k=K\right)$ is called the posterior probability of sample labels, y is the selected number of target classes, the total number is recorded as Y, k is the value of qualitative characteristics, and the number of samples under this condition is recorded as E. When the feature value is K, the sample size is converted to N^K. In order to balance the mean effect between prior probability and posterior probability, conditional parameters with multiple repetitions and variability are introduced. The calculation method is Formula (8).

$$Q\left(y=Y\middle|k=K\right)=\phi*\bar{O}\left(y=Y\right)+\left(1-\phi\right)*P\left(y=Y\middle|k=K\right) \tag{8}$$

In Formula (8), the probability of mean encoding is denoted as $Q\left(y=Y\middle|k=K\right)$, and $\bar{O}\left(y=Y\right)$ can control the slope of the conditional parameter, that is, as the value of $\bar{O}\left(y=Y\right)$ increases, the rate of change of ϕ with y,k becomes slower.

B. Fusion Algorithm based on Differential Automatic Regression Moving Average Variables

DAMAV is suitable for predicting time series and is widely used in linear regression, as expressed in Formula (9).

$$h_t=\mu_0+\mu_1 h_{t-1}-\nu_t-\nu_1\varpi_{t-1} \tag{9}$$

In Formula (9), the order of autoregression and moving average is recorded as h_t,μ_0, ν_t represents the error in stochastic process, and the number of differences made in regional time is expressed as ϖ_{t-1} [18]. On a practical basis, importing the formed dataset into the constructed model can complete autonomous training and calculate the results. The newly created dataset can also keep the model fresh, update its automation capabilities, and effectively reduce human resources, as demonstrated in Fig. 3.

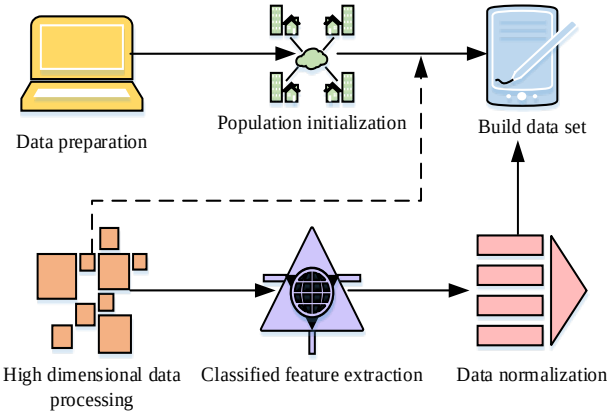


Fig. 3. XGBoost regression tree autonomous learning flow chart.

Fig. 3 shows the self-learning process of XGBoost regression tree. Firstly, it is necessary to collect a complete and comprehensive dataset, which can be formed on the network or experimentally measured; Then, set the data format for the algorithm, which is the target variable or feature value; Next, the outliers in the data set are screened, such as outlier and noise; Finally, the formatted data is run in the algorithm to extract the corrected values of the parameters [19]. Among them, to determine the indicators of predictive performance, this study selected Root Mean Square Error (RMSE) to consider it, as displayed in Formula (10).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(\theta^i-\overline{\overline{\theta}}^{=i}\right)^2}{n}} \tag{10}$$

In Formula (10), the true value is denoted as θ, and the predicted value is represented by $\overline{\overline{\theta}}$. In employee turnover in enterprises, temperature values are used instead of turnover values. The lower the temperature, the more severe the employee turnover in the enterprise is. In the iterative process of the algorithm, the training results will make the weight larger, so that the model can be steadily improved on the basis of the mean value, which can be described by the Newton's

Cooling Law (CL). The formula for CL is Eq. (11).

$$T'(t) = -\vartheta\left(T(t) - H\right) \tag{11}$$

In Formula (11) above, the temperature of the object is denoted as T, t represents the operating time, and the cooling rate is the derivative of T changing with t, called $T'(t)$. Room temperature is represented by H, where ϑ is a constant with a value range of $[0.1,1.0]$ to control the cooling rate. When the value of ϑ is fixed, the larger the temperature difference in the environment, the faster the cooling rate. In the employee turnover model, when compared to the average salary of all enterprises, the lower the salary, the faster the employee turnover rate. To describe the degree to which employees attach importance to the company, sample weights were introduced to simulate the degree to which employees pay attention to nearby companies, and Formula (12) was established.

$$w_k = w_0 e^{-\sigma\varsigma_k} \tag{12}$$

In Formula (12), the popularity of company k in the training set is denoted as w_k. The average treatment of all companies is represented by w_0. The heat loss coefficient of the model is denoted as σ. The company's decay rate over time is represented by ς_k. The XGBoost model is a foundational learner suitable for various types, which can transform linear problems into regression problems and is suitable for most work environments. XGBoost's loss function is composed of regularization terms, and its expansion is Formula (13).

$$\Omega = \omega\xi + 0.5\psi\sum_{j=1}^{\psi}\zeta_j^2 \tag{13}$$

In Formula (13), ξ represents the richness of the decision tree, and ω is the weight of the decision tree. The work done by the leaves in unit time is recorded as ζ, and the output regularization weight is expressed as ψ. When the XGBoost model predicts employee turnover, missing values in features can be ignored and assigned to leaves, thereby improving the overall training speed. In the working process of the decision tree, out of pocket errors are obtained from out of pocket data. The performance used to calculate the decision tree is Formula (14).

$$\text{Im } por\tan ce\left(FeatureX\right) = \left(\sum_{i=1}^{N} errOOB_2 - errOOB_1\right)/N \tag{14}$$

In Formula (14), the importance of the feature is denoted as $\text{Im } por\tan ce\left(FeatureX\right)$, and the bag contains a total of N data, where $errOOB$ represents the out of bag error value [20]. To train all samples in the dataset, a random variable is randomly extracted from the type features, and then converted into a numerical value based on the label of the previous sample. The weight coefficient of the priority is adjusted as Formula (15).

$$E_k^i = \frac{\left(\sum_{j=1}^{p-1}\left(E_{\Delta j,k} = E_{op,k}\right)Y_{\sigma j}\right) + ap}{\left(\sum_{j=1}^{p-1}E_{\Delta j,k} = E_{op,k}\right) + a} \tag{15}$$

In Formula (15), $E_{\Delta j,k}$ represents the training set in the input algorithm model, and the differentiable loss variable is denoted as $E_{op,k}$. a, p is a constant, and its value size reflects the priority of the sample. This study is based on the improved XGBoost regression tree algorithm model of DAMAV (DV-XGBoost), as listed in Fig. 4.

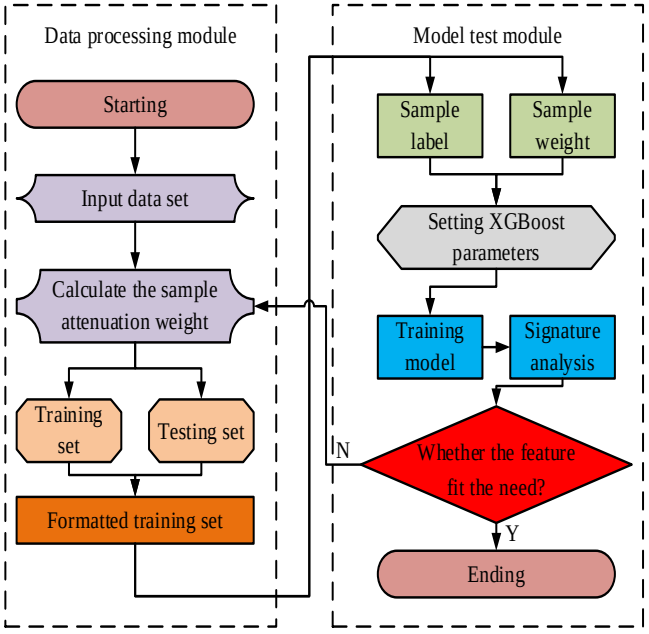


Fig. 4. Flow chart of improved DV-XGBoost regression tree algorithm model.

The DV-XGBoost model in Fig. 4 can be divided into two modules, totaling four parts. Firstly, to calculate the attenuation weight of the input dataset and divide it into a test set and a training set; Then, format the training set in the data preparation module and input it into the model test set; Next is to iterate the labels and weights of the samples to analyze the parameters of the DV-XGBoost model; Finally, the feature is judged. If it meets the requirements, the final value is output. Otherwise, the data preparation process is returned and the attenuation weight is recalculated.

IV. Experimental Study on Enterprise Character Loss Information Based on XGBoost

To verify the effectiveness of the DV-XGBoost algorithm in practical applications, iteration and accuracy verification were conducted. Finally, the DV-XGBoost model was applied to the Company dataset for simulation experiments.

A. DV-XGBoost System Development Environment and Model Parameter Determination

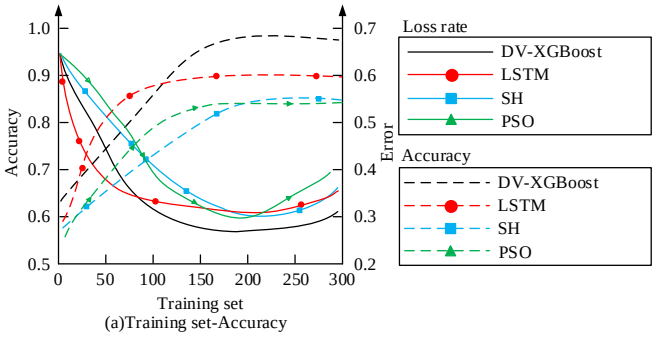
This study selected a self-collected Company dataset, including four types of companies: fine organic, electronic

engineering, aerospace, and automation, with a total of 6523 enterprises. Considering the limited types of data, the dataset will be divided into a training set and a testing set in a ratio of 2:3. The specific equipment and software used in the experiment are Table I. In the experiment of DV-XGBoost, the research first sets the necessary experimental conditions. There are some key parameters to be set in DV-XGBoost algorithm, including iteration times and acceleration factor because the problem of employee turnover in enterprises is difficult to predict and the computing resources are limited. Therefore, the optimization objective function is studied to solve the optimal solution of DV-XGBoost. The objective function has a clear optimization goal, and there are constraints on the range of the number of brain drain for each iteration of DV-XGBoost algorithm, the position, velocity and fitness data of each particle are collected and used to calculate the convergence performance of DV-XGBoost algorithm.

Experimental Parameters

Data set	Development language	Number of cores	Internal storage
Company	Python 8.5	8	1024 G
Operating system	Display card	Database	Processor
127Ubnutu 22.01.21	37.0 GHz	Mysqppl 5.30.2023	Intel Core i9
Web development framework	Language	Operator	Model
Django2.22.3	Easy Chinese	Electric, orangic...	F2.9LII-US M

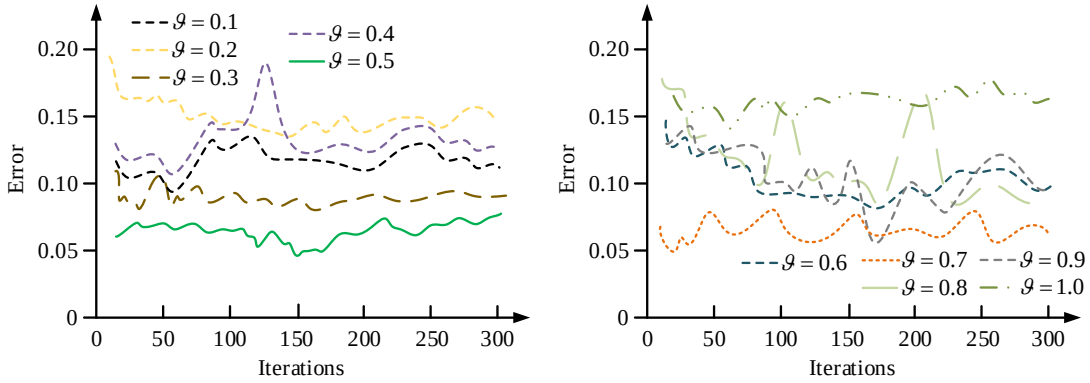
The collected dataset needs further processing to enable the studied algorithm to learn. The dataset was processed using DV-XGBoost for iterative optimization. To verify its accuracy, traditional Spotted Hyena Algorithm (SH), Long Short Term Neural Network (LSTM), and Particle Swarm Optimization Algorithm (PSO) will be compared with it. The accuracy and error rate results in the training set are Fig. 5.



(a)Training set-Accuracy

Fig. 5. Comparison of accuracy-training set image and error rate-training set image.

From Fig. 5, before 100 training sessions, the accuracy of DV-XGBoost algorithm is slightly lower than that of SH and PSO, and the error rate is higher. However, when the iterations reaches 100 or more, the accuracy of DV-XGBoost is higher than both algorithms, and tends to stabilize at 190 iterations, which is higher than the other three algorithms. Although increasing the iterations may reduce the operational efficiency of the model, after comprehensive consideration, the accuracy weight of the model is higher. Therefore, the DV-XGBoost algorithm proposed in the study has better performance. After the learning of the DV-XGBoost algorithm is completed, it is also necessary to consider the parameter determination during testing, as demonstrated in Fig. 6.



(a) The rate control constant of the step cooling curve is 0.1-0.5

(b) The rate control constant of the step cooling curve is 0.6-1.0

Fig. 6. Error-training times image of regularization term.

The parameter of this study is the rate control constant $\vartheta \in [0.1,1.0]$ of the cooling curve between the enterprise and employees. From Fig. 6, when the rate control constant of the step cooling curve is 0.5 and the number of iterations is 150, the error rate is the lowest, which is 0.042. Therefore, the final number of iterations was determined to be 150 and the rate control constant was taken as 0.5.

B. Experimental Verification of Employee Turnover Prediction in Enterprises based on DV-XGBoost

To verify the accuracy of the DV-XGBoost model in predicting employee turnover in enterprises, simulation experiments were conducted. It evaluates the practicality of the DV-XGBoost algorithm by observing whether an employee has resigned. First, the DV-XGBoost algorithm is initialized, and then the employee enterprise information flow is entered in the data preparation module. Finally, the rate control constant of the step cooling curve is set to 0.5, and the

employee dynamics and resigned employee information within 60 days are collected. The image drawn after calculating the error is Fig. 7.

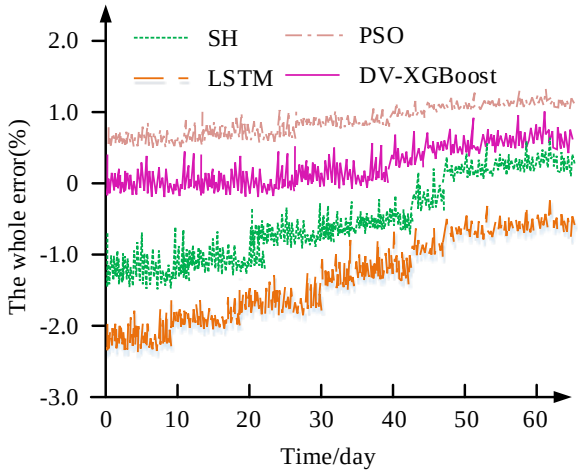


Fig. 7. Total error-time image of four algorithms.

Fig. 7 shows the error comparison of four algorithms in the experiment. In Fig. 7, after 38 days, the error of DV-XGBoost in determining employee turnover has approached zero, while the other three algorithms have a wide range of error fluctuations. Especially for the PSO algorithm, on the third day, the highest error value of the four algorithms was -2.94%. The total error range of DV-XGBoost, SH, LSTM, and PSO is significantly different, making it easy to compare algorithm performance. However, relying solely on the analysis of total error is not objective enough, so the study analyzed the four model analysis errors caused by individuals or enterprises, as shown in Fig. 8.

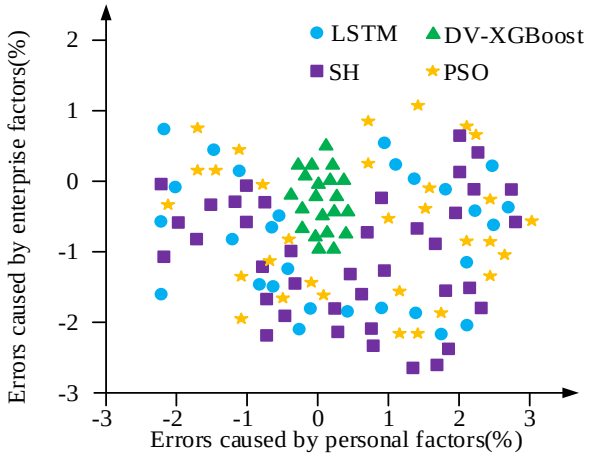


Fig. 8. The types and genres errors of the four algorithms.

From Fig. 8, it can be clearly seen that the experimental results of the DV-XGBoost model are concentrated in the range of total error of 0. The error range caused by personal factors is [-0.6%, 0.8%], while the error range caused by corporate factors is [-1.0%, 1.1%]. The error distribution of the remaining three algorithms is wide, and the distribution of larger errors is sparse. To more intuitively distinguish the ability of the four algorithms to correct errors, 400 experimental data records were conducted and the images displayed in Fig. 9 were plotted.

From Fig. 9, in 400 error testing experiments, LSTM has the largest range of error variation, with the highest frequency of errors recorded as [-0.75%, 0.62%]; Next is the SH algorithm, which is between [-0.18%, 0.23%]. The error variation range of PSO is close to DV-XGBoost, with values above [-0.12%, -0.04%]; the error curve of DV-XGBoost fluctuates between -0.03% and 0.02%, with the smallest fluctuation range. Excluding the LSTM and SH algorithms with the highest error ranking, only comparing the experimental results of DV-XGBoost algorithm and PSO for correct prediction, and drawing the error matrix. The resulting image is Fig. 10.

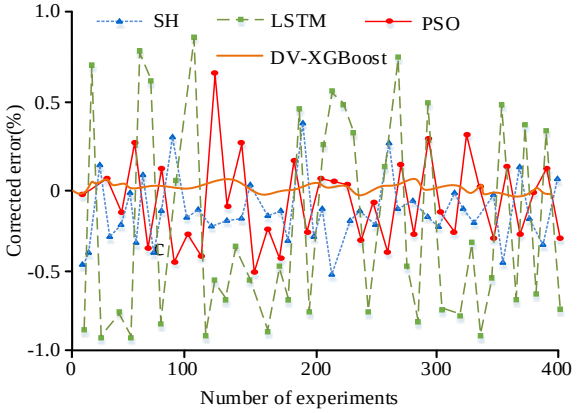


Fig. 9. Error changes of four algorithms in four hundred calibration experiments.

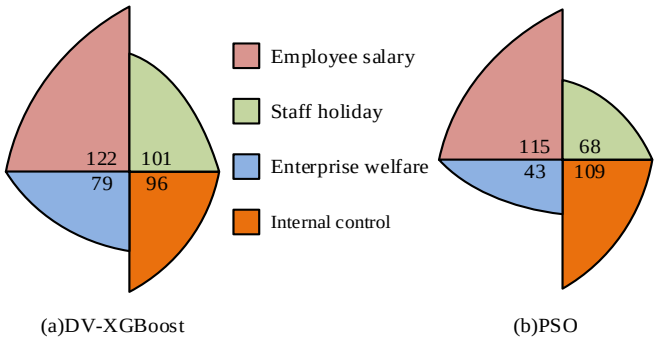


Fig. 10. Error matrix of DV-XGBoost algorithm and PSO algorithm.

Fig. 10 predicts four types of employee turnover in companies based on four conditions: employee salary, employee leave, and enterprise characteristics, as well as internal control, based on employee differences. The experimental results accurately predicted by DV-XGBoost and PSO are presented. The prediction accuracy of DV-XGBoost reached 398 times, with an accuracy rate of 99.5%, and the accuracy rate of PSO was 83.2%. To observe the experimental results of DV-XGBoost and PSO more intuitively, a linear fitting graph based on matrix drawing of two algorithms and Golden Sine algorithm (GS) was studied. The predicted values of the two were compared with the true values, as expressed in Fig. 11.

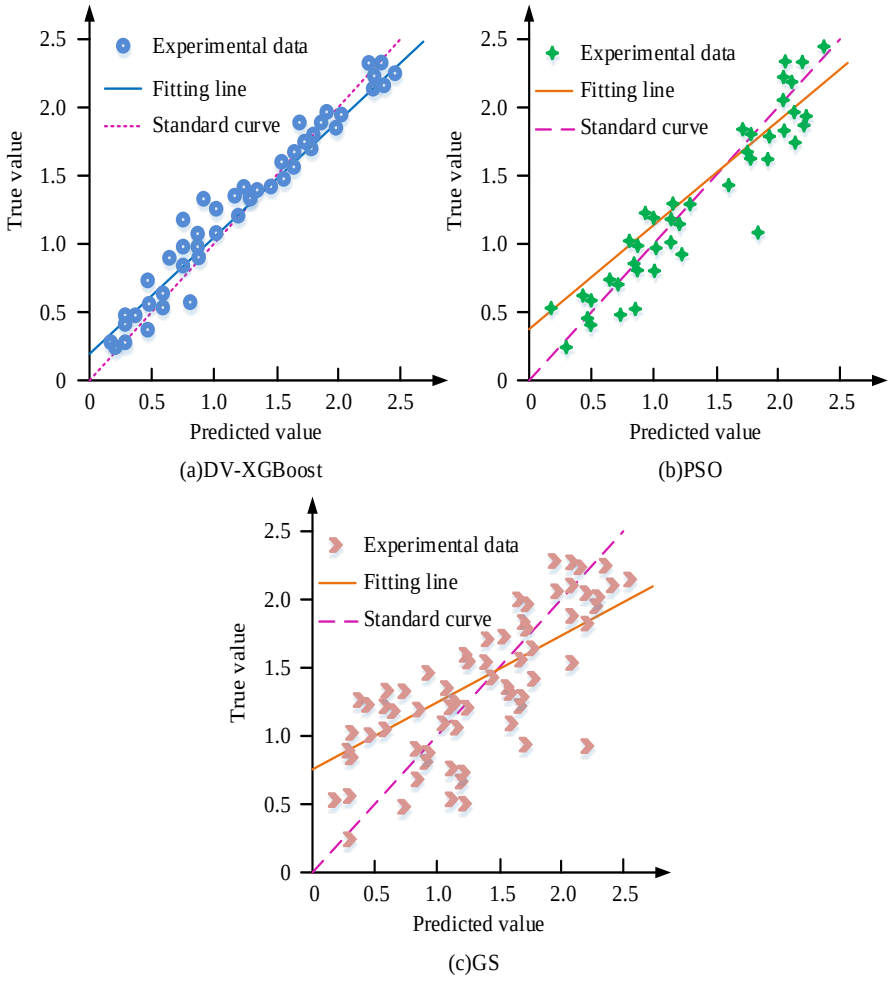


Fig. 11. Linear fitting diagram of DV-XGBoost PSO and GS.

Fig. 11 shows the comparison of three algorithms on predicted and true values. From Fig. 11, the linear fit (R^2) of the DV-XGBoost algorithm is 0.9914, the R^2 of PSO is 0.9547, and the R^2 of GS is 0.8825, indicating that there is no underfitting in the model. In summary, it can be concluded that the DV-XGBoost algorithm model can accurately predict the situation of employee turnover in enterprises, provide internal control warnings, or notify the human resources department (HR) of the company to release recruitment information.

V. Results And Discussion

With the increase of employee turnover rate in enterprises, the utilization rate of human resources is in a downward trend. In order to predict this kind of problem, and then make relative measures as early as possible, this study uses the improved limit gradient lifting algorithm to draw a step cooling curve for the collected employee's psychological data according to the cooling law, so as to predict their turnover possibility. In order to reduce the chaotic data during the running of the algorithm, the experiment was carried out in the Company data set. In order to analyze the proposed DV-XGBoost algorithm extensively, the experimental results are compared with PSO, SH and GS, and finally the number of iterations is 150, and the rate constant is 0.5. In the error analysis experiment, the error curve of DV-XGBoost fluctuates between-0.02% and 0.04%, with the smallest fluctuation range. The prediction accuracy of DV-XGBoost is 99.5%, and that of PSO is 83.2%. The linear fitting of DV-XGBoost is 0.9914, the linear fitting is excellent, and the linear fitting of PSO is 0.9547. The experimental results show that the DV-XGBoost model proposed in this study has strong robustness in predicting employees' psychology and is suitable for improving the utilization rate of human resources for the company. However, the algorithm is only applicable to companies, and the research on the employment factors of school employees is still insufficient, which will be gradually improved in future research.

VI. Conclusion

With the growth of internet technology, predicting the turnover psychology of enterprise employees is becoming increasingly important, such as increasing work efficiency for the administrative department of the company and providing early warning for the talent gap period of the enterprise. This study is based on XGBoost and DAMAV to generate the fusion algorithm DV-XGBoost. The experiment took into account both personal and corporate factors during resignation,

and conducted simulation experiments on the Company dataset, comparing with three algorithms such as LSTM. 40% of the Company dataset was extracted and trained on the DV-XGBoost model. Through controlling the rate constant of the cooling curve, the final iteration number was determined to be 150 times, with a rate constant value of 0.5. In the error analysis experiment, a total of 400 experiments were conducted, and the errors of the SH and LSTM algorithms fluctuated within the range of [-2.3%,1.4%] and [-2.2%,1.9%], respectively. The error curve of DV-XGBoost fluctuates between -0.02% and 0.04%, with the smallest fluctuation range; The variation range of PSO is close to DV-XGBoost, between [-0.13%, -0.03%]. Draw an error matrix for the experimental results of DV-XGBoost algorithm and PSO. In 400 experiments, the prediction accuracy of DV-XGBoost is 99.5%, and the accuracy of PSO is 83.2%. This study drew linear fitting graphs for two algorithms based on matrices. The R^2 of DV-XGBoost was 0.9914, indicating excellent linear fitting, while the R^2 of PSO was 0.9547. In summary, the DV-XGBoost model can accurately predict employee turnover in enterprises, improve the efficiency of human resources departments, and enable enterprises to cope with talent shortages. However, the DV-XGBoost model is only suitable for analyzing companies with employees from the same location. For comprehensive companies from multiple regions, it is difficult to analyze the emotional changes caused by local customs among employees, and the model will label them as noise. This is because human psychology is complex and diverse, and the dataset for research and analysis contains fewer types. With the increase of volunteers, it is believed that future research can be improved.

ACKNOWLEDGMENT

The research is supported by Shaanxi Federation of Social Sciences, Research on the Characteristic Development and Innovation Path of County Economy under the Background of Transcendence in Shaanxi New Era, (No.2021ND0035); Rural Revitalization Bureau of Xingping City, Shaanxi Province, Rural Revitalization Plan of Xingping City, (No. GH [2021]00026); Development and Reform Bureau of Xingping City, The High-tech Industrial Park's Industrial Development Plan of Xingping City, (No. HX2022001).

REFERENCES

[1] Guo Y, Mustafaoglu Z, & Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2022, 2(1), 5–9.

[2] Chen J, Zhao F, Sun Y, Y Yin. Improved XGBoost model based on genetic algorithm. International Journal of Computer Applications in Technology, 2020, 62(3): 240-245.

[3] Calanna P, Lauriola M, Saggino A, M Tommasi, S Furlan. Using a supervised machine learning algorithm for detecting faking good in a personality self-report. International Journal of Selection and Assessment, 2020, 28(2): 176-185.

[4] Zhao W P, Li J, Zhao J, D Zhao, J Lu, X Wang. Xgb model: research on evaporation duct height prediction based on xgboost algorithm. Radioengineering, 2020, 29(1): 81-93.

[5] Lu Y, Fu X, Guo E, Tang. XGBoost algorithm-based monitoring model for urban driving stress: Combining driving behaviour, driving environment, and route familiarity. IEEE Access, 2021, 9: 21921-21938.

[6] Deng X, Li M, Deng S, L Wang. Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. Medical & Biological Engineering & Computing, 2022, 60(3): 663-681.

[7] Li Z X, Shi X L, Cao J D, XD Wang,W Huang. CPSO-XGBoost segmented regression model for asphalt pavement deflection basin area prediction. Science China Technological Sciences, 2022, 65(7): 1470-1481.

[8] Tao T, Liu Y, Qiao Y, LG B, JL A, CZ A, WA Yu. Wind turbine blade icing diagnosis using hybrid features and Stacked-XGBoost algorithm. Renewable Energy, 2021, 180(Dec.): 1004-1013.

[9] Li S, Zhang X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. Neural Computing and Applications, 2020, 32: 1971-1979.

[10] Gu X, Han Y, Yu J. A novel lane-changing decision model for autonomous vehicles based on deep autoencoder network and XGBoost. IEEE Access, 2020, 8(99): 9846-9863.

[11] Li C, Zhang Y, Cui C, D Fan, Y Zhao, XB Wu, B He, Y Xu, S Li, J Han. Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost. Monthly Notices of the Royal Astronomical Society, 2021, 506(2): 1651-1664.

[12] Osman A I A, Ahmed A N, Chow M F, YF Huang, A El-Shafie. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 2021, 12(2): 1545-1556.

[13] Song P, Liu Y. An XGBoost algorithm for predicting purchasing behaviour on E-commerce platforms. Tehnički vjesnik, 2020, 27(5): 1467-1471.

[14] Gao S, Li S. Bloody Mahjong playing strategy based on the integration of deep learning and XGBoost. CAAI Transactions on Intelligence Technology, 2022, 7(1): 95-106.

[15] Ünver M, Olgun M, Türkarslan E. Cosine and cotangent similarity measures based on Choquet integral for Spherical fuzzy sets and applications to pattern recognition. Journal of Computational and Cognitive Engineering, 2022, 1(1): 21-31.

[16] Wang H, Yue W, Wen S, X Xu, HD Haasis, M Su, P Liu, S Zhang, P Du. An improved bearing fault detection strategy based on artificial bee colony algorithm. CAAI Transactions on Intelligence Technology, 2022, 7(4): 570-581.

[17] Shahbazi Z, Byun Y. Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. International Journal of Advanced Science and Technology, 2020, 29(4):6979-6988.

[18] Osman A I A, Ahmed A N, Chow M F, YF Huang, A El-Shafie. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 2021, 12(2): 1545-1556.

[19] Oslund S, Washington C, So A, Chen, T, &Ji, H. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. Journal of Computational and Cognitive Engineering, 2022, 1(4): 152-158.

[20] Li C, Zhang Y, Cui C, D Fan, Y Zhao, XB Wu, B He, Y Xu, S Li, J Han. Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost. Monthly Notices of the Royal Astronomical Society, 2021, 506(2): 1651-1664.