The Application of The Decision Tree Algorithm Based on K-means in Employee Turnover Prediction

Zhang Yunmeng, Zhang Chengyi
School of Science, Xi'an Polytechnic University, Xi'an 710600, China
E-mail: 125295694@qq.com

Abstract. The employee's voluntary turnover is one of the most difficult problems for the company. The research uses a human resource data as an example. K-means is used to classify employees, and then each type of the decision tree algorithm is used to conduct prediction and analysis of turnover. After research and analysis according to the experimental results, the two types of employees with high turnover rate are analyzed, and the management is provided with measures to prevent the occurrence of some acts of turnover.

1. Introduction

Employee turnover refers to the voluntary outflow of employees who are unwilling by enterprises but willing by individuals [1]. Because the voluntary turnover of employees will lead to additional expenses for the company and thus affecting the normal operation of the whole company, the issue of employee turnover has been a hot research topic at home and abroad. At present, research on employee turnover prediction has always been a hot issue. Chinese and foreign researchers used different methods to predict employee turnover. In 2008, Xia Guoen and Jin Weidong used SVM under structural risk minimization criteria to predict customer churn and compared it with other algorithms [2]-[3]. In 2011, Li Gaoguo used naive bayes to classify and established a probability model to predict the retention of employees [4]. In 2014, Qu Xiaoli used the decision tree C4.5 algorithm to predict the employee turnover of H group retail chain stores [5]. In 2018, Zhang Shichao studied the human resource analysis and prediction method based on random forest with a score of 0.98 [6]. In 2019, Li Qiang and Yan Liang built an LRA prediction model by stacking integrated learning algorithm combined with adaboost and random forest basic algorithms to achieve an enterprise employee turnover forecast and the forecast accuracy rate is 89.09% [7].

However, the above methods basically solved the problem of employee turnover prediction with a single algorithm, and did not consider the classification of employees' different characteristics. Therefore, this paper uses K-means algorithm to classify employees, and then uses the decision tree model to classify each data. In this way, the accuracy of prediction can be improved by distinguishing different types of employees, so as to better solve the problem that companies worry about employee turnover.

2. Research method

2.1. Employee clustering and k-value determination

2.1.1. Employee clustering
The advantage of the K-means algorithm is that it is simple and time-consuming to calculate [8]. Therefore, this study uses the K-means method to classify employees. The specific steps are:

1. First enter the cluster value k, that is, we want to divide the data into k categories;
2. Randomly forming k data points as the clustering center of the data set;
3. Calculate the distance from each data point in the data set to each k, and if a certain data point is closest to a cluster center, it is classified as such;
4. Calculate the average of the distances in each class, and select a new cluster center;
5. If the distance between the new cluster center and the old cluster center is less than a certain range, then we think that clustering can end the algorithm;
6. If the distance between the new cluster center and the old cluster center varies greatly, you need to continue iterating (3)~(5) steps.

2.1.2. Determination of k value
In the steps of K-means method, the elbow method and the contour coefficient method are used in this study to confirm the k value.

1. The elbow method. The important measure of the elbow method is SSE (sum of the squared errors),

\[ SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \]

Where, \( C_i \) is the i th cluster, \( p \) is the sample point in \( C_i \), \( m_i \) is the center of mass of \( C_i \) (the mean of all samples in \( C_i \)). The relationship between SSE and k value is like an elbow, and the k value corresponding to the elbow is the real clustering number of data.

2. The contour coefficient method. The important measure of the contour coefficient method is the contour coefficient. The contour coefficient of a data point \( X_i \) is defined as follows:

\[ S = \frac{b - a}{\max(a, b)} \]

Where \( a \) is the average distance from other data points of the same type \( X_i \), called the degree of cohesion, and \( b \) is the average distance from all data points in the nearest class of \( X_i \), called the degree of separation. The most recent class is defined as follows:

\[ C_j = \arg \min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2 \]

Where \( p \) is a sample in a cluster \( C_k \). The maximum k value of the average contour coefficient is the optimal cluster number.

2.2. Resignation forecast
The decision tree is a tree structure similar to the flow chart [10] as shown in Figure 1:

![Decision Tree Diagram](image)

Figure 1. Each internal node in the figure represents a test on an attribute, each branch represents a test output, and each leaf node represents a category.

3. Research process

3.1. Data introduction
The experimental data in this study are derived from Kaggel's human resources analysis data, a total of 14999. The data includes variables: satisfaction_level; last_evaluation; number_project; average_monthly_hours; time_spend_company; Work_accident; left; promotion_last_5years; sales; salary.

3.2. Data pre-processing

3.2.1. Data conversion

2,999 pieces of the original data were extracted set as the experimental data set, which is called “Track”. The data set Track has included 10 variables, of which the content of sales and salary is text. The text should be converted into numerical value for subsequent analysis.

3.2.2. Data analysis

First, after checking the data set Track, it was found that there were no missing values or vacancies, and the data was clean and effective. Then, we found that left is significantly related to satisfaction_level, correlation coefficient $r = 0.39$, as shown in Figure 2.

![Figure 2. Linear correlation graph of each variable. The data in the square in the graph is the correlation coefficient between the two variables.](image-url)

3.3. Experiment analysis

3.3.1. Determine k value

Use the elbow method and the contour coefficient method on the data set Track to determine the value of the classification number $k$. 
In Figure 3, it is obvious that SSE starts to stabilize when k=2, and elbow appears at k=2, so it is judged according to the elbow method that the optimal value of k is 2. In Figure 4, the contour coefficient. The largest point appears at k=2, so the best value of k is also determined according to the contour coefficient method. Therefore, combining the two methods, the optimal value of k is 2.

3.3.2. Data clustering

Import the dataset Track into SPSS22, named Track1. Change the measurement of the variable according to the nature of the variable to achieve accurate analysis of the result. Then cluster the dataset Track1, the result of 3.3.1, so gather Track1 to 2 classes can be obtained from Table 1.

Table 1. Number of cases in each cluster.

| Number of clusters | 1   | 1468 |
|--------------------|-----|------|
| Effective          | 2   | 1531 |
| Lost               |     | 0    |

As can be seen from Table 1, there are 1468 data in the first category, which are separately exported and stored as Track2-1; there are 1531 data in the second category, which are separately exported and stored as Track2-2; 2999 data are effectively divided into two categories, and stored separately.

3.3.3. Data decision

The decision tree operations are performed on the data sets Track1, Track2-1, and Track2-2, respectively, and the three sets of data sets are analyzed by using the four algorithms included in the SPSS22 tree: CHAID, exhaustive CHAID, CRT, and QUEST. 80% of the data is selected as the training data, and the remaining 20% is the test data. The results of the analysis are summarized in Table 2.

Table 2. The output of the decision tree after two types of K-means.

| Growth method | Track1 Mean value after classification | Track2-1 Mean value after classification | Track2-2 Mean value after classification | Increase amount | Mean of increase |
|---------------|--------------------------------------|----------------------------------------|----------------------------------------|-----------------|-----------------|
| CHAID         | 90.4%                                 | 91.5%                                  | 92.3%                                  | 91.9%           | 1.5%            |
| Detected value| 92.0%                                 | 91.3%                                  | 91.4%                                  | 91.4%           | 0.7%            |
| Exhaustive CHAID | 90.7%                            | 90.8%                                  | 92.1%                                  | 91.5%           | 0.7%            |
| Detected value| 91.0%                                 | 94.2%                                  | 90.0%                                  | 92.1%           | 1.1%            |
| CRT           | 91.7%                                 | 93.7%                                  | 91.4%                                  | 92.6%           | 0.8%            |
| Detected value| 90.4%                                 | 93.6%                                  | 93.8%                                  | 93.7%           | 3.3%            |
| QUEST         | 90.1%                                 | 91.2%                                  | 91.0%                                  | 91.1%           | 1.0%            |
It can be seen from Table 2 that the combination of K-means and CRT method is 2.1%, which is obviously larger than the increase of K-means combines with other methods. Among them, the accuracy rate in Track2-1 is 93.7%, and the accuracy of detection data is accurate. It is also as high as 93.6%. Therefore, after using the CRT method for the data sets Track2-1 and Track2-2, there are Figure 5 and Figure 6.

Figure 5. Track2-1 the decision tree of whether to lose employees.

Figure 6. Track2-2 the decision tree of whether to lose employees.

4. Research results and analysis

From the analysis of the human resources data, we found that the combination of K-means and CRT algorithm improves the prediction accuracy of whether to leave or not. The highest accuracy reached 93.7%, the testing data of the accuracy is as high as 93.6% through the decision tree algorithm to get the two types of the characteristics of the high turnover of staff as follows: (1) employee satisfaction to the company <=0.465, and participate in the project number <=2.5; (2) when the working life is >3.5, and the performance evaluation is >0.765.

When low employee satisfaction to the company, the fewer number of project, the lower salary, the likelihood of leaving may be higher. So for this kind of leadership should take the initiative to understand why employees on the company's degree of satisfaction is not high, if you have some measures to improve the employee satisfaction, the class will greatly reduce the employee turnover rate. When employees have a long working life and high performance evaluation, they are more likely to quit. This type of employees tends to belong to the advanced talents, leaders of the company should improve the promotion mechanism of the company and strive to retain top talents so as to achieve better development of the company.

In general, process and results of this research can provide new ideas for relevant workers and provide new application directions for academic researchers. It is hoped that further exploration can be carried out by using machine learning, regression analysis, neural network and other technologies to further explore, so as to achieve a higher accuracy rate, solve more practical difficulties for enterprise companies, and help enterprises to develop better and faster.
References

[1] Chen, X.M. (2002) New Human Resource Management. Central Compilation and Translation Press, Beijing.

[2] Xia, G.E., Jin, W.D. (2008) Customer Churn Prediction Model Based on Support Vector Machine. Systems Engineering Theory and Practice, 28(1): 71-77.

[3] Vladimir, N.V., Xu, J.H., et al. (2004) Statistical Learning Theory. Electronic Industry Press, Beijing.

[4] Li, G.G. (2011) Analysis and Prediction of Staff Turnover Based on Data Mining Technology. Human Resource Management, (1): 75-79.

[5] Qu, X.L. (2014) Data Mining Technology in Staff Turnover Application—A Case Study of the Data of Staffs form H Group. Technology & Innovation Management, 35(05): 452-456.

[6] Zhang, S.C. (2018) Research on Human Resource Data Analysis and Prediction Methods Based on Random Forest. Communications World, (08): 313-314.

[7] Li, Q., Zhai, L. (2019) Analysis and Research of Employee Turnover Prediction Based on Stacking Algorithm. Journal of Chongqing Technology and Business University (Natural Science Edition), 36(01): 117-123.

[8] Zhang, J., Duan, F. (2013) Improved K-means Algorithm for Optimizing Initial Clustering Center. Computer Engineering and Design, 34(5): 1691-1694+1699.

[9] Xu, J.R., Zhan, Y.Z. (2018) Improved K-means Fast Clustering Algorithm Based on Spark. Journal of Jiangsu University (Natural Science Edition), 39(03): 73-80.

[10] Li, R.P. (2010) Research on Decision Tree Classification Algorithm in Data Mining. Journal of East China Institute of Technology (Natural Science Edition), 33(2): 192-196.