Advances in Social Science, Education and Humanities Research, volume 543
Proceedings of the 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021)

IT Industry Recruiting Knowledge System

Tien Yu Huang*

[1] *Associate professor, Economic and Management college, ZhaoQing University,ZhaoQing City, Guangdong,China*
**Corresponding author. Email: 3069565060@qq.com*

ABSTRACT
Professional workers are essential assets for enterprises. With the development of big data applications, mobile computing and electronic commerce, the needs of IT industry professionals are increasing. Enterprises cannot be competitive with others without recruiting, training and keeping them. Because of the need of human resources, training high-quality talents are the key success factors for IT industry. We constructed a data model framework and implement data mining techniques such as CART, cluster analysis, and MARS to explore the relationship between employee resume of basic data and their working performance for recruiting new employees by human resources department as a reference. It could help them to filter resumes and recognize the relationships between information in the resume and future performance of employees and help decision makers of human resource management to find out and choose potential highly performance employees for the enterprises.

Keywords: *IT industry, Data mining, CART, Cluster analysis, MARS, Human Resources Management*

1. INTRODUCTION

Big data is a huge and complicated data set. Gartner [1] defined “Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation”. It can be applied to many industry and research fields, including statistic data in government department, transaction data of bank and medical research development. Internet, intelligent mobile device, internet of Things, cloud computing contributes to the development of big data. No matter whether you agree or not, you provide your data to the net when searching on the net and press “yes” in social APP or using ATM or paying through credit card. These data were transmitted to databases. These data were analysed to find out some truth which was hidden behind. The type of big data is different from the traditional database architecture. IBM [2] stated big data as “V” characteristics, “Volume”,” Variety”,” Velocity” and “Veracity”. “Volume” stands for the amount of data; “Velocity” stands for the types of kinds of data; “Velocity” stands for the accuracy of data. Since the development of big data applications, the urgent job vacancies are increasing. The development of IT technology created a lot of job opportunities for IT professionals in big data applications such as web search service and new business operation model such as electronic commerce, business process re-engineering, etc.

When enterprises are recruiting employees, they are mostly subjective whether the applicants get the chance to be employed or not based on the decision makers’ past experience, the company needs for professional skills, and the resumes of the applicants. Choosing suitable employees and forecast their future performance are becoming more and more important to organizations. Cole et al.[3] proposed a framework of recruiter assessment of applicants’ resume content to predict applicant mental ability and big five personality dimensions and give a clear suggestions for selecting appropriate candidates for their enterprises. In the other hand, we will focus on the basic information from the candidates ‘resumes. Employers want to choose the right people in the right position. It is necessary for companies to build a recruit system to filter good and suitable talents in a comparative subjective way.

Moreover, compensation management is the important role of human resources department of enterprises, it is not only essential to the cost of human resources but also cause to the resignation of employees whose salaries did not meet their expectation. The development of IT industry deeply relies on talents as human resources assets but is still having problems to find suitable candidates. The professional skills and knowledge are not easily to be identified by the employers, and more serious problem is the brain drain. Due to the above situations stated above, we tried to implement data mining method to build a recruit knowledge system for human resources department and give suggestions. It could help employers to make right decision in recruiting professionals and prevent unnecessary waste of enterprise resources. In this paper, we would like to build a talent filtering system using data mining skills in IT industry.

2. RELATED WORK

The concept of big data is widely used in the analysis of data of enterprises inside itself such as data warehouse, data mining, business intelligence, and statistics application in the past decade. Data mining is the technique of using one or more than one computing techniques to analyse and extract data and knowledge. We aim to dig hidden trend and deeper knowledge from the data model built or from the summaries through the decision process. We will discuss three data mining techniques to the knowledge system such as CART, cluster analysis and MARS. The results can support human resource in clear establishing clear strategies for the acquisition and the development of the right skills needed to leverage big data.

2.1. CART

In CART, the following equation represents the sum of the multiplication of the proportions of each two different categories in the node, which can be interpreted as p(i | node) is the purity of the i category in the node, and p(j | node) is the purity of the j category in the node where i is the category of one category , and j is another category.

$$i(node) = p(\, j \mid node)$$

Decision tree is frequently used in data mining, generally be used in classify binary and many classifications. Input values can be continuous or discreet, and output is tree model describing decision process. Decision trees produce remarks of categories or probabilities and could transform the results into rules of association rules and displayed graphically.

Lu stated [4] data mining is the process to classify data by the features of them, apply historical data to build forecast model, and classify data set into clusters to understand the difference inside the clusters and also search for the association rules to discuss the influences between the variables. CART has enhanced features and capabilities that address the shortcomings of CART giving rise to a modern decision tree classifier with high classification and prediction accuracy.[5] Mauro et al. [6] also discussed the job opportunities related to big data which human resources department wanted to know what kinds of job requirement listed for recruiting talent people in big data fields, and generated classification of job rules and skill sets. Yung-Tsan Jou [7] found out key assessment indicators of training and performance evaluation using data mining neural network and decision tree C5.0 within 19 performance indicators of Taiwan training quality system to provide better training quality system to enhance the quality of human resource training.

2.2. Cluster Analysis

In statistics, Ward's method is a criterion applied in hierarchical cluster analysis[8]. Ward's minimum variance method is a special case of the objective function approach originally presented by Joe H. Ward, Jr.[9] .Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function[9].

Hitka et al. [10]used clustering analysis as a strategic advantage of human resource management to classify workers into three groups. There are differences between three groups in motivation of working. In this way, employer can use different motivation and management skills to train and motivate them for better working performance. Punj et al. [11] discussed about cluster analysis is the statistical method for classification, unlike other statistical methods for classification, such as discriminant analysis and automatic interaction detection, it makes no prior assumptions about important differences within a population.

2.3. MARS

Multivariate adaptive regression splines (MARS) is a form of regression analysis introduced by Jerome H. Friedman in 1991 in statistics. The MARS model building procedure automatically selects which variables to use (some variables are important, others not), the positions of the kinks in the hinge functions, and how the hinge functions are combined. MARS builds models as the form

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x).$$

The model is a weighted sum of basic functions $B_i(x)$. Each c_i is a constant coefficient. For example, each line in the formula for performance above is one basis function multiplied by its coefficient[12]. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables. It can automatically create correct models to predict continuous and binary dependent variables and is good at finding the best complex variable conversion and interaction of data structures hidden in high-dimensional data. The basic concept is to use several paragraphs of explanation equations (Spline Basis Function, BF) to solve multiple and complex data problems with a combination of more flexible forecasting models. In the model, the number of interpretation equations is determined by the interaction between the parameters of the data itself. In addition, after evaluating the loss of fit (LOF) of each interpretation of equation, the number of influencing variables is determined. MARS can predict discontinuous or continuous dependent variables based on different data attributes and retain important information in the data [13]. TIRYAKI et al. did an experimental study was performed, and the mechanical behaviour of impregnated wood was determined as a result of the experimental process. Multiple adaptive regression splines (MARS), teaching–learning based optimization (TLBO) algorithms and conventional regression analysis (CRA) were applied to different regression functions by using experimentally obtained data. The experiment results showed that MARS is the optimal method compared with the other two methods.

3. RESEARCH METHODS

The set of data with more than 1000 records of employee data together with the performance data of each of employee. Performance grade is the dependent variable while the rest of the variables are independent variables. The variables are showed in Table 1 as showed.

Table 1. Employee resume item data list

Item	category	Description
1	Position	Position description
2	Birthday	Employee birthday
3	Sex	Male, Female
4	Work start date	Employee Work Start date
5	Marital status	Marital status
6	Education level	Top education degree
7	Height	Employee height (in cm)
8	Weight	Employee Weight(in kg)
9	Nationality	China.etc.
10	Address	City,Road,Section,Lane, Alley, Floor, No.
11	Salary	Monthly payment
12	Load	Family-related cost
13	Senior Years	Number of years worked
14	Performance grade	Pass or improvement-needed

We tried to implement CART (classification and regression tree) as analysis tool for analyzing continuous and categories variable data. The decision tree will be constructed and prune to a tree by testing data sets. We decide the results of pruning based on the performance of pruning phases. CART is good when dealing with missing values. Gini method is frequented used in calculated impurity in CART algorithm. If the value is high, it means there are many categories in the classification. We implemented 10-folds cross validation method to prevent sampling errors, divided samples to ten sets, and random sampled by the scale of good or bad samples in the sets. We chose nine sets as training sets in every round of folding using three methods such as clustering analysis and multi-linear regression and CART algorithm to construct classification model.

CART model can also find relatively important variables in the process of model construction. Empirical results of CART model are described as follows. There are four variables retained after being strained, namely "Education level", "Marital status", "Position", and "years of worked". The CART model uses a binary recursive method to classify data by taking "Education Level" as the classification standard, when the education level of the sample data belongs to university and master's degree, the model judges it as good performance, and vice versa. In this model, education level is used as the only segmentation node. The reason is that the importance of learning is 98% as the important variables selected in this model, and the rest of other variables are not high. The average overall identification rate is 57%. We also implemented Cluster Analysis with the same set of employee data. This research took the advantages of the hierarchical classification and non-hierarchical classification. Firstly, we tried to find the initial suitable number of groups of data using hierarchical classification and continue with the second phase of grouping with non-hierarchical classification. When we construct classification model of clustering analysis, firstly we use wards method of hierarchical analysis to do data classification. The result of the clustering analysis showed that there are three groups in total of 14 observed variables. They are “Education”, “job title” and “age”. The average total degree of discrimination is 62.5 %.

The MARS model constructed in this research compares the importance of 14 variables to explore the important selection variables that affect the outcome of education and training courses. Because the MARS model uses the cumulative sum of several interpretation equations to find out the characteristics of the data, researchers can not only see the important variables from the MARS model, but also know from the interpretation equations that the important variables are in practice. The construction of its model is mainly based on several basic equations to form a complete model, which is then filtered by the LOF criterion to retain significant and important variables. The average overall identification rate is 64.1%.

Comparing the three models, it can be seen that the overall identification rate of multi-adaptive cloud shape regression is 64.1%, which is the highest. On the whole, the identification result of CART is generally low. Through the analysis of the three analysis tools, it is known that the best correct discrimination rate can reach 64.1%.

4. CONCLUSION

Since the empirical study of all the above research methods, we could see the best methods of separating employees with better performance. The results showed that the features of highly potential employees are “education level” and” position”. It mainly indicated “University or college graduated” in education level. The education level needed for high level management and technical level is “master degree”, and the requirement is increasing. These highly technically and management professionals are urgently required especially in the new technology fields such as big data, cloud or mobile computing fields, etc.Our research provides a knowledge recruiting system with selection criteria for IT industry, it might contribute more precise resume screening for interview, and prevent unnecessary waste for time, labour cost and resources. Recruiting appropriate "talents" is the foundation for companies to increase production capacity and increase profits.

REFERENCES

[1] M. A. Beyer and D. Laney. The importance of big data: A definition. Stamford, CT: Gartner, 2012

[2] Sam B. Siewert. Big data in the cloud, https://www.ibm.com/developerworks/library/bd-bigdatacloud/index.html,July 2013

[3] Cole, Michael S.; FEILD, Hubert S.; GILES, William F. Using recruiter assessments of applicants’ resume content to predict applicant mental ability and big five personality dimensions. International Journal of Selection and Assessment, 2003, 11.1: 78-88.

[4] Chi-Jie Lu,Tian-Shyug Lee,shu-Han Hsiao,Bo-Rong Hsu, A Study on Applying Data Mining Classification Technologies for Recruitment。Journal of Data Analysis,7(2) ,1 – 27,2012

[5] Priyam, Anuja, et al. Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 2013, 3.2: 334-337.

[6] De Mauro, Andrea, et al. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management, 2018, 54.5: 807-817.

[7] Yung-Tsan Jou, Yih-Chuan Wu, Wen-Tsann Lin. Applying Decision Tree and Neural Network to Raise the Performance of Human Training Quality. Journal of Quality, 22(5),383-403,2015

[8] https://en.wikipedia.org/wiki/Ward%27s_method

[9] Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236–244.

[10] Hitka, Miloš, et al. Cluster analysis used as the strategic advantage of human resource management in small and medium-sized enterprises in the wood-processing industry. BioResources, 2017, 12.4: 7884-7897.

[11] Punj, Girish; Stewart, David W. Cluster analysis in marketing research: Review and suggestions for application. Journal of marketing research, 1983, 20.2: 134-148.

[12] Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines". The Annals of Statistics. 19 (1): 1–67. CiteSeerX 10.1.1.382.970. doi:10.1214/aos/1176347963. JSTOR 2241837. MR 1091842. Zbl 0765.62064

[13] Horng, Shih-Cheng; LIN, Shieh-Shing. Merging crow search into ordinal optimization for solving equality constrained simulation optimization problems. Journal of computational science, 2017, 23: 44-57.

[14] Tiryaki, Sebahattin, et al. Performance evaluation of multiple adaptive regression splines, teaching–learning based optimization and conventional regression techniques in predicting mechanical properties of impregnated wood. European Journal of Wood and Wood Products, 2019, 77.4: 645-659.