International Journal of Research in Advent Technology, Vol.2, No.9, September 2014
E-ISSN: 2321-9637

An Efficient Classification Algorithms for Employee Performance Prediction

V.Kalaivani[1], Mr.M.Elamparithi[2]
M.Phil. Scholar, Department of Computer Science,
Assistant Professor, Department of PG Computer Applications, STC, Tamilnadu
Email: vanivelusamy@rediffmail.com[1], Email: parithi_1980@yahoo.com[2]

Abstract-In any organization's talent management is becoming an increasingly crucial method of approaching HR functions. Talent management can be defined as an outcome to ensure the right person is in the right job. Human talent prediction is the objective of this study. Due to that reason, classification and prediction in data mining which is commonly used in many areas can also be implemented in this study. There are various classification techniques in data mining such as Decision tree, Neural networks, Genetic algorithms, Support vector machines, Rough set theory, Fuzzy set approach. This research has been made by applying decision tree classification algorithms to the employee's performance prediction. Decision tree is among the popular classification technique which generates a tree and a set of rules, representing the model of different classes, from a given data set. Some of the decision tree algorithms are ID3, C5.0, Bagging, Random Forest, Rotation forest, CART and CHAID. In this study, C4.5, Bagging and Rotation Forest algorithms are used. Experiments were conducted with the data collected from an institution which is implemented in WEKA tool.

Index Terms **-**Human talent, Data Mining, Classification, Decision tree, C4.5, Bagging, Rotation Forest.

1. INTRODUCTION

Human resource has become one of the main concerns of managers in almost all types of businesses which include private concerns, educational institutions and governmental organizations [6]. Talent is considered as any individual who has the capability to make a significant difference to the current and future performance of the organization. In fact, managing talent involves human resource planning that regards processes for managing people in organization.

Besides that, talent management can be defined as an outcome to ensure the right person is in the right job; process to ensure leadership continuity in key positions and encourage individual advancement; and decision to manage supply, demand and flow of talent through human capital engine. In HRM, talent management is very important and need some attentions from HR professional [1].

Data mining is a step in the KDD process concerned with the extraction of patterns from the data. Nowadays, there are some researchers on solving HRM problems that uses Data mining approach. Basically, most of the Data Mining researches in HR problems domain focus on personnel selection task and few apply in other activities such as planning, training, talent administration and etc. [1].

Recently, with the new demands and the increased visibility of HR management, thus, HRM seeks a strategic role by revolving to data Mining methods. This can be done by identifying the patterns that relate to the talent. The patterns can be generated by using some of the major Data Mining techniques. The matching of Data mining problems and talent management needs are very important, in a way to define the suitable Data Mining techniques.

2. DATA MINING AND ITS TECHNIQUES

Data mining is a collection of techniques for efficient automated discovery of patterns in large databases. That must be actionable so that they may be used in an enterprise's decision making process. Data mining techniques provides a way to use various data mining tasks such as classification, regression, time series analysis, clustering, summarization, association rules and sequence discovery, to find solutions for a specified problem.

2.1 Classification

Classification involves finding rules that partition the data into separate groups. The input for the classification is the training data set, whose class labels are previously known. Classification explores the training data set and constructs a model based on the class label, and intentions to allocate a class label to the future unlabeled records. Since the class field is well-known, this type of classification is known as supervised learning. There are several classification models such as Decision Tree, Genetic algorithms, statistical models and so on.

2.2 Association rules

Association rule is the descriptive model of data mining this enables us to establish association and relationship between large and classified data items

based on certain attributes and characteristics. The result of Association rules can help prevent failures by some appropriate measures.

2.3 Clustering

Clustering is a method of grouping data into different groups, so that the data in each cluster share similar trends and patterns. Clustering creates a major class of data mining algorithms. The algorithm tries to automatically partition the data space into a set of regions or clusters, to which the instances in the table are assigned, either deterministically or probability-wise. The objective of the process is to identify all sets of similar instances in the data, in some best manner.

2.4 Neural Networks

Neural Networks are a new paradigm in computing, which involves developing mathematical structures with the ability to learn. The methods are the result of academic attempts to model the nervous system learning. Neural networks have the significant capability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

2.5 Web Mining

Web mining is a specialized application of data mining. Web mining is a technique to process data available on Web and search for useful data. Web mining enables us to determine web pages, text documents, multimedia files, images and other kinds of resources from web. Pattern extraction is a web mining process to monitor the original or uploaded web pages, extract information from them and generate matches of a specific pattern with necessary information specified by a user. The pattern extraction process enables us to efficiently surf and access data available on the web.

3. RELATED WORKS

Lipsa Sadath (2013) used Data Mining (DM) techniques for automated intelligent decisions from rich employee database for predictions of employee performance implementing the finest KM strategies, thus achieving stable HR system and brilliant business.

Qasem et al. (2012) discussed data mining techniques which were utilized to build a classification model to predict the performance of employees. They used Decision Tree for build the classification model, where various classification rules were generated. To validate the generated model, more than a few experiments were conducted using real data collected from several concerns.

Salleh et al. (2011) tested the influence of motivation on job performance for state government employees in Malaysia. The study showed a positive relationship between affiliation motivation and job performance. As people with higher affiliation motivation and strong interpersonal relationships with colleagues and managers tend to perform much better in their jobs.

Jantan et al. (2010) present the study on how the potential human talent can be predicted using a decision tree classifier. By using this technique, the pattern of talent performance can be identified through the classification process. In that case, the hidden and valuable knowledge discovered in the related databases will summarized in the decision tree structure. In this study, they use decision tree C4.5 classification algorithm to generate the classification rules for human talent performance records. Finally, the generated rules are evaluated using the unseen data in order to estimate the accuracy of the predication result.

Jantan et al. (2010) also propose the potential data mining techniques for talent forecasting. Data mining technique is the best balanced estimator, decision tree and neural network and is found useful in developing predictive models in many fields. In this study, they attempts to use classifier algorithm C4.5 and Random Forest for decision tree; and Multilayer Perceptron (MLP) and Radial Basic Function Network for neural network. They focus on the accuracy of the techniques to find the suitable classifier for HR data. The data are for management and professional employees from higher education institution.

Same authors Jantan et al. (2011) discussed Human Resources (HR) system architecture to forecast an applicant's talent based on information filled in the HR application and past experience, using Data Mining techniques. The goal of the paper was to find a way to talent prediction in Malaysian higher institutions. So, they have specified certain factors to be considered as attributes of their system, such as, professional qualification, training and social obligation. Then, several data mining techniques (hybrid) where applied to find the prediction rules. ANN, Decision Tree and Rough Set Theory are examples of the selected techniques.

Chein and Chen (2006) have worked on the improvement of employee selection, by building a model, to predict the performance of newly applicants. Depending on attributes selected from their CVs, job applications and interviews. Their performance could be predicted to be a base for decision makers to take their decisions about either employing these applicants or not. As a result for their study, they found that employee performance is highly affected by education degree, the school tire, and the job experience.

Engel et al. (Elsevier-2014) exploits an Approach to improve the performance of Weka, a popular data mining tool, through parallelization on

GPU-accelerated machines. As a result, they observed speedup levels of at least 49%; drastically decreasing the time, the algorithm consumes to handle a dataset.

4. METHODOLOGY

This study has been made by applying decision tree classification algorithms to the employee performance prediction. A decision tree is a classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given data set. Experiments were conducted with the data collected from an institution. C4.5, Bagging and Rotation Forest algorithms are used in this system.

4.1 C4.5 Algorithm

The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. In building a decision tree, we can compact with training sets that have record with unknown attribute values by estimating the gain, or the gain ratio, for an attribute by in view of only those records where those attribute values are available. The gain and gain ratio is calculated as follows:

$$\text{Gain}(X, T) = \text{Info}(T) - \text{Info}(X, T) \quad \text{---------Eq. (1)}$$

$$\text{Gain Ratio}(X, T) = \frac{\text{Gain}(X, T)}{\text{Split Info}(X, T)} \quad \text{----Eq. (2)}$$

We can classify records that have unknown attribute values by estimating the probability of the various possible results. Unlike CART, which generates a binary decision tree, C4.5 produces trees with variable branches per node. When a discrete variable is selected as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

Algorithm

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1, s_2, \cdots$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, \ldots, x_{p,i})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then persists on the smaller sub lists.

4.2 Bagging Algorithm

Bagging decision trees, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction. However, there are strong empirical indications that bagging and random subspace methods are much more robust than boosting in noisy settings [21].

Algorithm: Bagging

The bagging algorithm-create an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equality-weighted prediction.

Input:

= D, a set of d training tuples;
= k, the number of models in the ensemble;
= a learning scheme (e.g., decision tree algorithm, back propagation, etc.)

Output: A composite model, M*.

Method:

(1) For i=1 to k do//create k models:
(2) Create bootstrap sample,D*i*,by sampling D with replacement;
(3) Use Di to derive a model,M*i*;
(4) End for

To use the composite model on a tuple, X:

(1) If classification then
(2) Let each of the k models classify X and return the majority vote;
(3) If prediction then
(4) Let each of the k models predict a value for X and return the average predicted value;

4.3 Rotation Forest Algorithms

Rotation Forest is a recently proposed method for building classifier ensembles using independently trained decision trees. It was found to be more accurate than bagging, AdaBoost and Random Forest ensembles. Rotation Forest [19] draws upon the Random Forest idea. The base classifiers are also independently built decision trees, but in Rotation Forest each tree is trained on the whole data set in a rotated feature space. As the tree learning algorithm builds the classification regions using hyper planes parallel to the feature axes, a slight rotation of the axes may lead to a very different tree. The effect of rotating the axes is that classification regions of high accuracy can be constructed with fewer trees than in bagging and AdaBoost.

Rotation Forest Algorithm:

1. The feature set is randomly split into K subsets.
2. PCA is applied to each subset.
3. All principal components are retained.
4. Arrange the PCA coefficients in a matrix (rotation matrix).

5. Apply the rotation matrix to the data features.
6. Build each decision tree on the rotated training data.

4.4 Datasets

Data is collected from the educational institutions. In order to collect the required data, a questionnaire was prepared and distributed to employees working in an institution. The questionnaire was filled by 217 employees. After the questionnaires were collected, the process of preparing the data was accomplished. First, the information in the questionnaires has been transferred to Excel sheets. Then, the types of data has been reviewed and modified. These files are prepared and converted to (ARFF) format to be compatible with the WEKA data mining toolkit. The database consists of 53 attributes and the input dataset is shown in Fig 1.

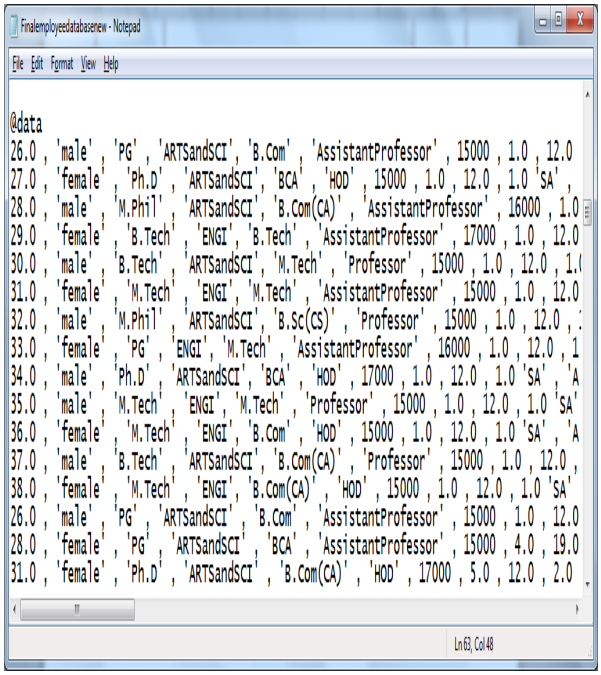


Fig.1. Input Dataset in Text Editor

4.5 Experimental Results and Discussion

Decision tree algorithms were applied with WEKA and the accuracy of the classification techniques with cross-validation test is depicted in table 1.It is observed that C4.5, Bagging algorithms have poor accuracy compare with Rotation Forest algorithm and not suitable for this problem domain due to the nature of the data. This table shows an accuracy percentage for C4.5, Bagging and Rotation forest algorithm with cross-validation of 10 folds test option.

Table 1 Results of Decision Tree Algorithms with cross-validation test

Algorithm used	**Cross-validation 10-folds % accuracy**
C4.5	41.47%
Bagging	45.62%
Rotation forest	51.46%

The accuracy of the classification techniques with Training set test is depicted in table 2.It is observed that C4.5, Bagging algorithms have poor accuracy compare with Rotation Forest algorithm and not suitable for this problem domain due to the nature of the data. This table shows an accuracy percentage for C4.5, Bagging and Rotation forest algorithm with Training set test option.

Table 2 Results of Decision Tree Algorithms with Training set test

Algorithm used	**Training set % accuracy**
C4.5	84.79%
Bagging	75.57%
Rotation forest	100%

4.6 Analysis of Algorithm with Accuracy

The Fig 2 indicates that above specified three algorithms has a different accuracy value. An accuracy percentage for C4.5 algorithm with cross-validation of 10 folds test option is 41.47% and the training set test option is 84.79%.

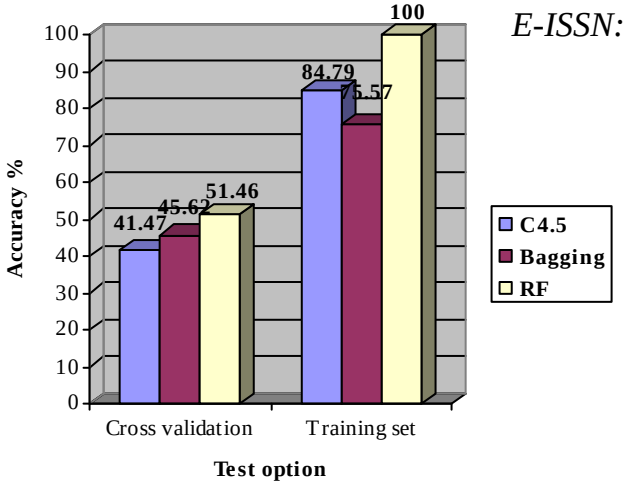
An accuracy percentage for Bagging algorithm with cross-validation of 10 folds test option is 45.62% and the training set test option is 75.57%.

An accuracy percentage for Rotation forest algorithm with cross-validation of 10 folds test option is 51.46% and the training set test option is 100%. In above results, the Rotation forest algorithm has the highest Accuracy value.

Fig.2. Analysis of algorithm with accuracy

4.7 Analysis of Algorithm with Time

The Fig 3 indicates that above specified three algorithms has a different time. Time taken for C4.5 algorithm with cross-validation of 10 folds test option



is 0.02 seconds and the training set test option is 0 seconds.

Time taken for Bagging algorithm with cross-validation of 10 folds test option is 0.11 seconds and the training set test option is 0.08 seconds.

Time taken for Rotation forest algorithm with cross-validation of 10 folds test option is 1.87 seconds and the training set test option is 1.81 seconds. Consider the above four algorithm C4.5 has the minimum time value for build the model.

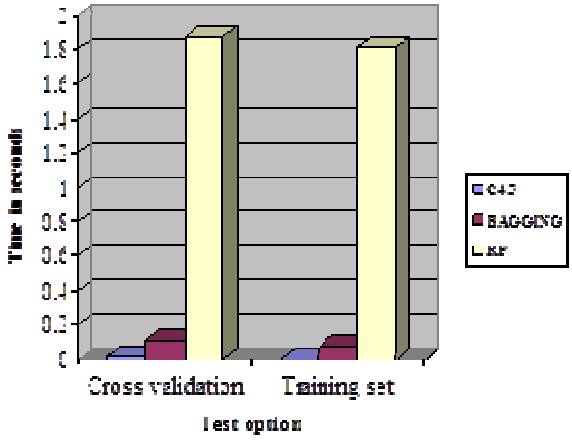


Fig.3. Analysis of algorithms with Time

5. CONCLUSION

Classification is the one of the hottest topics in the area of data mining. The research activities on this topic is reviewed hence, the survey guides the researches to get an idea about the recent advancements with Classification.

On working on performance, many attributes have been tested, and some of them are found effective on the performance prediction. For companies managements and human resources departments, this model can be used in predicting the newly applicant personnel performance. Several activities can be taken in this case to avoid any risk related to hiring poorly performed employee.

In this paper, the parameter time value is from 0 seconds to 1.87 seconds among three algorithms. The accuracy value is from 41.47% to 100% for both the cross-validation and training set test options among three algorithms.

From the performance analysis, each algorithm has performed well. Consider the above results Rotation forest algorithm has better performance than other three algorithms because it has maximum accuracy value 51.46% for cross-validation and 100% for training set test option. Therefore, the Rotation forest Algorithm is more efficient algorithm for employee's performance prediction when compared to other two algorithms.

6. FUTURE WORK

This research has described the significance of the study on the use of data mining classification techniques for employee's performance prediction. The performance and efficiency of this research can be improved through some future enhancements. Some of the future enhancements that can be involved in this research are:

- Other Decision Tree Classification algorithms can be used to obtain better performance.
- Collect more proper data from several institutions or companies.

REFERENCES

[1] Jantan et al. "Towards applying Data Mining Techniques for Talent Management", IPCSIT vol.2, IACSIT Press, Singapore, 2011.

[2] Jantan, H., Hamdan, A.R. and Othman, Z.A. (2010b). "Human Talent Prediction in HRM using C4.5 Classification Algorithm", International Journal on Computer Science and Engineering, 2(08-2010), pp. 2526-2534.

[3] Sivaram, Ramar 2010."Applicability of Clustering and Classification Algorithms for Recruitment Data Mining," International Journal of Computer Applications, vol 4, pp.23-28, 2010

[4] Qasem A.Al-Radaideh, Eman AI Nagi "Using Data mining techniques to build a Classification model for predicting employees performance," IJACSA,vol 3,pp.144-151,2012.

[5] Al-Radaideh, Q. A., Al-Shawakfa, E.M., Al-Najjar, M.I. (2006). "Mining Student Data Using Decision Trees", International Arab Conference on Information Technology (ACIT 2006), Dec 2006, Jordan.

[6] Chein, C., Chen, L. (2006) "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry", Expert Systems with Applications, In Press.

[7] Kayha, E. (2007). "The Effects of Job Characteristics and Working Conditions on Job Performance", International Journal of Industrial Ergonomics, In Press.

[8] Salleh, F., Dzulkifli, Z., Abdullah, W.A. and Yaakob, N. (2011). "The Effect of Motivation on Job Performance of State Government Employees in Malaysia", International Journal of Humanities and Social Science, 1(4), pp. 147-154.

[9] Jantan, H., Hamdan, A.R. and Othman, Z.A. (2010a) "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application", International Journal of Humanities and Social Science, 5(11), pp. 694-702.

[10] Senol zafer erdogan, mehpare timor "A Data Mining Application in a Student Database". Journal of aeronautics and space technologies, vol 2, July 2005 pp.53-57

[11] Ruxandra PETRE "Data Mining for the Business Environment," Database Systems Journal, vol. 4, 2013.

[12] Lipsa Sadath "Data Mining:A Tool for Knowledge Management in Human Resource," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol-2, April 2013.

[13] Brijesh Kumar Bhardwaj, Saurabh Pal," Data Mining: A prediction for performance improvement using classification," International Journal of Computer Science and Information Security, Vol. 9, April 2011.

[14] Vaneet Kumar,Dr. Vinod Sharma," Student's Examination Result Mining: APredictive Approach,"International Journal of Scientific & Engineering Research, Volume 3, November-2012 .

[15] Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, Elvi Wani, "Prediction Of Student Academic Performance By An Application Of Data Mining Techniques," International Conference on Management and Artificial Intelligence IPEDR vol.6,2011, IACSIT Press, Bali, Indonesia.

[16] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method," World Journal of Computer Application and Technology 2(2): 43-47, 2014 DOI: 10.13189/wjcat.2014.020203

[17] Yang Yanga, Suzanne S. Faridb, Nina F. Thornhilla"Data mining for rapid prediction of facility fit and debottlenecking ofbiomanufacturing facilities", Journal of Biotechnology 179 (2014) 17–25(Elsevier)

[18] Tiago Augusto Engela, Andrea Schwertner Charaoa,, Manuele Kirsch-Pinheirob, Luiz-Angelo Steffenelc,"Performance improvement of data mining in Weka through GPU acceleration,"Procedia Computer Science 2014 ,pp.93 – 100,Elsevier

[19] J. J. Rodr´ıguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1619–1630, Oct 2006.

[20] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, pp. 1761-1768, 2007.

[21] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publisher, 2006.

[22] J. Ranjan, "Data Mining Techniques for better decisions in Human Resource Management Systems," International Journal of Business Information Systems, vol. 3, pp. 464-481, 2008.

[23] Huang, M.J., Y.L. Tsou, and S.C. Lee, "Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge". Knowledge-Based Systems, 2006. 19(6): p. 396-403.

[24] Tung, K.Y., et al., "Mining the Generation Xer's job attitudes by artificial neural network and decision tree – empirical evidence in Taiwan". Expert Systems and Applications, 2005. 29(4): p. 783-794.

[25] Chen, K.K., et al., "Constructing a Web-based Employee Training Expert System with Data Mining Approach", in Paper in The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007). 2007.

[26] Chien, C.F. and L.F. Chen, "Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing". IEEE Transactions on Semiconductor Manufacturing, 2007. 20(4): p. 528-541.

[27] Tai, W.S. and C.C. Hsu (2005) "A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method". http://www.atlantis-press.com/php/download_papaer?id=46.

[28] K.P.Soman, Shyam Diwakar, V.Vijay, Insight into Data Mining theory and practice, Prentice-Hall of India Private Limited, 2006.

[29] Arun K Pujari, Data Mining techniques, Universities Press (India) Private Limited, 2001.

[30] G.K.Gupta, Introduction to Data Mining with Case Studies, PHI Learning Private Limited, 2006.