PERFORMANCE ANALYSIS OF HR DATA USING HYBRID ALGORITHM NAÏVE BAYES AND DECISION TREE IN BIGDATA

S.Chitra[1], Dr.P.Srivaramangai[2]
Research Scholar[1], Department of Computer Science, Marudupandiyar College(Affiliated to Bharathidasan University), Thanjavur - 613 403,Tamilnadu,India.

Associate Professor[2],
Department of Computer Science, Marudupandiyar College(Affiliated to Bharathidasan University), Thanjavur - 613 403,Tamilnadu,India.

Abstract: Human capital is the most effective resource to hiring the highly qualified personnel for improving the world economy and also for developing company's management. Data mining and clustering methods have been utilized in personnel performance evaluation. In this paper, the existing data mining algorithms Random Forest with hybrid Naïve Bayes and Decision Tree data mining approach, which makes the beneficiary of statistical modeling and intelligent of data mining techniques to extract meaningful set of patterns from Employee data repositories. In this paper, the main objective is to find out and analyze the accuracy of single data mining techniques and then it compares with the accuracy of hybrid data mining techniques to diagnose the employee performance. Performance analysis of these classification algorithms, Hybrid algorithm Naïve Bayes and Decision Tree performs better than other algorithms. WEKA software is used for running these algorithms and also utilized in machine learning.

Keywords: HR Data, Naïve Bayes, Decision Treee, Random Forest, Big data, Data mining

I. INTRODUCTION

Data mining technique is one of the best linear unbiased estimator, which means decision tree and neural network were found for developing predictive models in several fields [1]. Data Mining emerges AI technology or Knowledge Discovery in Database (KDD) which has been provided for exploration and analysis in large quantities of data to find out meaningful patterns and rules. In recent times, many researchers involves on solving HRM (Human Resource Management) difficulties that can utilize Data mining approach. In essence, most of the Data Mining researches in HR problems domain focus on Talent Analytics and few apply in other activities like training, planning, administration, personnel selection task, and etc. [2].

In any organization, talent analytics has become a crucial approach in HR functions. It considers the capability of any individual to ensure a significant difference to the current and future performance of the organization [3]. Human resource planning clearly emphasizes that processes for managing people in organization. Instead of talent management ensures the leadership continuity in key talent areas and encourage individual advancement; and decision to control supply, demand and flow of talent using human capital engine [4]. Talent management process is a crucial part and needs of some attention from HR professionals. The talent management process recognizes the key talent areas and identifies the people in the organization who constitute and conduct the key talent and progress activities to retain and engage for the talent pool to move into more vital roles [5]. These processes involve HR activities that need to be integrated into an efficient system.

Data mining techniques is mainly used to extract relevant and interesting knowledge from data patterns using the different classification and prediction models that can be allocated to support Performance analysis in HRM [6]. The powerful data analysis tool is to extract useful knowledge from vast amount of data available in HRM field. In last few decades, Talent analytics is the most important source of IT industry all over the world. During the employee performance analysis involves in different fields, single data mining techniques are showing satisfactory level of accuracy. The deployment of hybrid data mining methods demonstrates provident level of accuracy. In this paper, single decision tree data mining methods such as Random Forest, C4.5, and CART algorithms are analyzed and evaluated with hybrid data mining algorithm to achieve better results in HRM and also to formulate employees training courses, self improvement plan, and so forth.

1.1 Performance Management

Performance management system is an objective goal of ensuring the process directed to maximize the productivity of teams, employees, and ultimately work with the organization. While accomplishing the organizational strategy involves improving and measuring the value of the workforce. Although HR function contributes training and performance management, and appraisal significant role [6]. Even if performance appraisal arises at a specific time, performance management accomplishes ongoing, dynamic, and continuous process. A part of the PM system analyzes

the individual in the organization such as appraisal, rewards and training, is integrated for the use of continuous organizational efficiency. With PM, the effort of each and every worker must be directed toward achieving strategic goals.

Performance appraisal (PA) considers a formal system of review and evaluation of person or team task management. A critical point in the definition is the word formal, because in actuality, managers should be reviewing an individual's performance on a continuing basis [7]. For many organizations, the primary goal of an appraisal system is to raise individual skills and organizational performance. There may be other goals, however. A potential difficulty with PA, and a possible cause of much dissatisfaction, is expecting too much from one appraisal plan.

Uses of Performance Appraisal

- **Human Resource Planning:** In assessing a firm's human resources, data must be available to identify those who have the potential to be promoted or for any area of internal employee relations. Through performance appraisal it may be discovered that there is an insufficient number of workers who are prepared to enter management.
- **Recruitment and Selection:** Performance evaluation ratings may be helpful in predicting the performance of job applicants. For example, it may be determined that a firm's successful employees (identified through performance evaluations) exhibit certain behaviors when performing key tasks [7]. These data may then provide benchmarks for evaluating applicant responses obtained through behavioral interviews.
- **Training and Development:** Performance appraisal should point out an employee's specific needs for training and development. By identifying deficiencies that adversely affect performance, T&D programs can be developed that permit individuals to build on their strengths and minimize their deficiencies. An appraisal system does not guarantee properly trained and developed employees. However, determining T&D needs is more precise when appraisal data are available.
- **Career Planning and Development:** career development is a formal approach used by the organization to ensure that people with the proper qualifications and experiences are available when needed. Performance appraisal data is essential in assessing an employee's strengths and weaknesses and in determining the person's potential.
- **Compensation Programs:** Performance appraisal results provide a basis for rational decisions regarding pay adjustments [8]. Most managers believe that you should reward outstanding job performance tangibly with pay increases. They believe that the behaviors you reward are the behaviors you get. Rewarding behaviors necessary for accomplishing organizational objectives is at the heart of a firm's strategic plan. To encourage good performance, a firm should design and implement a reliable performance appraisal system and then reward the most productive workers and teams accordingly.
- **Internal Employee Relations:** Performance appraisal data are also used for decisions in several areas of internal employee relations, including promotion, demotion, termination, layoff, and transfer. For example, an employee's performance in one job may be useful in determining his or her ability to perform another job on the same level, as is required in the consideration of transfers. When the performance level is unacceptable, demotion or even termination may be appropriate.
- **Assessment of Employee:** Potential some organizations attempt to assess an employee's potential as they appraise his or her job performance [8]. Although past behaviors may be a good predictor of future behaviors in some jobs, an employee's past performance may not accurately indicate future performance in other jobs.

II. DATA MINING TECHNIQUES

The matching of data mining problems and talent management needs is very crucial Therefore, it is very important to determine the suitable data mining techniques. In HRM, there are some interests on solving HRM problems using data mining approach [9]. Several techniques that are used for data classification are decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic [10, 11]. Recently, Hybrid data mining approaches have gained much popularity; however, a few studies have been proposed to examine the performance of hybrid data mining techniques for response modeling. A hybrid approach is built by combining two or more data mining techniques. A hybrid approach is commonly used to maximize the accuracy of a classifier. In this section explain the Random Forest and Hybrid Naïve Bayes – Decision Tree algorithms and compare their performance. A hybrid model can improve the performance of basic classifier.

A. RANDOM FOREST

Random forest approach is a combination classification method proposed by Breiman in 2001. Using bagging method, random forest method will draw multiple training sample sets that are different from each other. Every single sample set builds a decision tree with randomly selected attributes [12]. Random forest uses CART algorithm for building trees. Considering the large number of built trees, random forest method is characterized with good ability to resist noise and outstanding performance in the classification capability.

Many variants of RF which are characterized [13] by 1) the way each individual tree is constructed, 2) the procedure used to generate the modified data sets on which each individual tree is constructed, 3) the way the predictions of each individual tree are aggregated to produce a unique

consensus prediction. An important feature of RF is its out-of-bag (OOB) error. Each observation is an OOB observation for some of the trees, i.e. it was not used to construct them and can thus be considered as an internal validation data set for these trees. The OOB error of the RF is simply the average error frequency obtained when the observations from the data set are predicted using the trees for which they are OOB. Through this internal validation, the error estimation is less optimistic and usually considered as a good estimator of the error expected for independent data. Although this is by far the most widely applied version, this standard RF method has an important pitfall.

Random Forest Algorithm

- For b= 1 to B
- Draw a bootstrap sample(new training sets by random sampling) Z of size N from the training data
- Grow a random-forest tree Tb to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree until the minimum node size n_{min} is reached.
    i. Select m variables at random from the p variables.
    ii. Pick the best variable split- point among the m.
    iii. Split the node into two daughter nodes.
- Output the ensemble of trees $\{\ \}_1^B$.

B. Hybrid Naïve Bayes – Decision Tree Algorithm

Naïve Bayes and Decision Trees are two of the most important classification algorithms for prediction purposes, due to their accuracy, easiness and effectiveness. There prediction accuracies can be increased further by combining the advantages of both the algorithms mentioned above, using a Hybrid Naïve Bayes – Decision Tree algorithm. This algorithm gives high prediction accuracy as compared to Naïve Bayes and Decision Tree used individually, but the time complexity does not increase by a great extent [14]. The implementation of the algorithm is divided into two parts. In the first part, Naïve Bayesian and Decision tree models are created and assessed individually on the training data. In the second part, the class probabilities obtained on every instance of the test set, are weightily averaged based on the classification accuracies obtained on training data [14]. Finally the result of the Hybrid Naïve Bayes – Decision Tree Algorithm, is compared with the results of Random Forest Algorithm calculated individually.

Algorithm:

- **Input:** The Models built in the first phase i.e. C4.5, NB, Their respective accuracies ACCC4.5, ACCNB, Test Data instance denoted by‘s’

STEPS:

- For every class label (in this case 2 class labels) ‘c’ of test instance‘s’. Here ‘c’ is either a ‘promoted’ or ‘not promoted’
- Calculate $P(c|s)_{C4.5}$ by using the decision tree model (C4.5). The formula is as follows:

$$P(c|s)_{c4.5} = \frac{\sum_{i=1}^{k} \delta(c_i, c)}{k}$$

Where ‘k’ is the number of training instances in that particular leaf node where‘s’ falls,
‘c_i ‘ is the class of the test instance ‘s’ [14]. $\delta(\cdot)is\ a\ binary\ function$, The value of the function is equal to ‘0’ if both the parameters are not equal and ‘1’ if they are equal [14]. For this dataset, Calculate $P(\text{not promoted}|s)_{C4.5}$ by using the decision tree model (C4.5) and the above formula Calculate $P(\text{promoted}|s)_{C4.5}$ by using the decision tree model (C4.5) and the above formula

- Calculate $P(c|s)_{NB}$ by using the Naïve Bayesian classifier model (NB). The formula is as follows:

$$P(s|c)_{NB} = P(c)\prod_{j=1}^{m} P(a_j|c)$$

Here, ‘m’ denotes the total number of attributes, ‘ a_j ’ is the value of jth attribute of the test instance ‘s’ and ‘c’ as mentioned earlier, is the value of the class attribute [14]. In this case, the Prior probability i.e. P(c) [14] is calculated by using the formula:

$$P(c) = \frac{\sum_{i=1}^{n} \delta(c_i, c) + 1}{n + n_c}$$

Here ‘c_i’ is the is the value of the class attribute of the ith training row, ‘n’ is the total number of rows in training set, ‘n_c’ is the total number of classes (in this case: 2) [14] and $\delta(\cdot)is\ a\ binary\ function$, The value of the function is equal to ‘one’ if both the parameters are equal and ‘zero’ if they are not equal [14]. The Conditional Probability $P(a_j|c)$ [14] is calculated by using the formula:

$$P(a_j|c) = \frac{\sum_{i=1}^{n} \delta(a_{ij}, a_j)\delta(c_i, c) + 1}{\sum_{i=1}^{n} \delta(c_i, c) + n_j}$$

Here ‘ a_j ’ is the value of jth attribute of the test instance ‘s’, ‘ a_{ij} ’ is the value of jth attribute of the training set row or instance ‘ i ‘, ‘c_i’ is the is the value of the class attribute of the ith training row and ‘n_j’ is the total number of values that the jth attribute can have [14]. For this dataset,
Calculate $P(\text{promoted}|s)_{NB}$ by using the Naïve Bayesian model (NB)
Calculate $P(\text{not promoted}|s)_{NB}$ by using the Naïve Bayesian model (NB)

- Calculate $P(c|s)_{C4.5}$ - NB by using the formula given below:

$$P(c|s)_{c4.5-NB} = \frac{ACC_{c4.5} \times P(c|s)_{c4.5} + ACC_{NB} \times P(c|s)_{NB}}{ACC_{c4.5} + ACC_{NB}}$$

For this dataset, Calculate $P(\text{promoted}|s)_{C4.5\text{-}NB}$ Calculate P(not promoted | s) C4.5-NB

- Find the maximum value of $P(c|s)_{C4.5\ -\ NB}$ which is obtained in the previous step. For this dataset, If ($P(\text{promoted}|s)_{C4.5\text{-}NB}$ > $P(\text{not promoted}|s)_{C4.5\text{-}NB}$) {

Then the Class Label of the instance is 'promoted ' according to the hybrid algorithm } else if (P(not defaulter | s)$_{C4.5\text{-}NB}$ > P(defaulter | s) $_{C4.5\text{-}NB}$){ Then the Class Label of the instance is 'not promoted' according to the hybrid algorithm }

- **Output:** The Class Label (defaulter or not defaulter) for a Test Data instance denoted by 's'

IV. EXPREMENTAL RESULT

In the experimental phase, we attempt to impart employee's performance patterns in the existing HR databases using selected common classification techniques.

In the model construction phase, the main classification techniques are Random forest and Hybrid Naïve Bayes and Decision Tree algorithm.

The dataset contains 1000 records from 10 years (2001-2010) performance evaluation marks. Each record holds evaluation marks for selected factors and the total mark for each of the year. The dataset is organized into 10 fold cross validation training and test dataset. The experimental tool used was WEKA. WEKA (Waikato Environment for Knowledge Analysis) is used for classifying data in this work. Weka is one of the popular suites of machine learning software developed at the University of Waikato. It is open source software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. The stages in the process include the following:

- **Data gathering:** To collect the information about the employees, the dataset was obtained from database records of IT industry employees' database kept at the IT Tech Software Solutions, Chennai, and Tamilnadu.
- **Pre-processing:** This phase involve that dataset preparation before applying data mining techniques. While using traditional pre-processing methods like data cleaning, transformation of variables and data partitioning were applied.
- **Data Mining:** In this stage, data mining algorithms are applied in order to predict students' performance. The classification algorithms Hybrid Naïve Bayes – Decision Tree and Random Forest algorithms are employed and compared.
- **Interpretation:** At this stage, the obtained models are analyzed to determine the employees' performance. This section is done by observing the factors that appeared (in the rules and decision trees) and how they are related for consideration and interpretation.

The training data sample involved a range of 12 attributes about which 12 attributes are input values and the Promotion Recommendation attribute serving as the target class. The target class attribute is discrete in nature with Yes/No as the values. The table 1 summarizes the list of attributes used for the talent acquisition process.

Table 1: Employee Performance Dataset

TALENT ATTRIBUTES	DESCRIPTIONS
ID	Unique identity of the employee
Category	P-Professional, S-Support Staff
Qualification	Doctorate, Master, Bachelor, Diploma and Certificate
Efficiency	Work Outcome
Experience	Work Experience
Technical Qualification	Knowledge and skill
Marks	Evaluation Marks
Programming Language skill	Knowledge and skill
International Conference	Activities and Contribution
Journals	Activities and Contribution
Workshop	Activities and Contribution
Promotion Recommendation	Activities and Contribution

Using above attributes for training dataset depends upon the related factors for employee performance. Depending on these values the learning algorithm will predict the employee performance. To compare the predicted value analyze with the actual value in the database. For evaluating the experimental results, 'Confusion Matrix' is used which is a common evaluation criterion for any classification model. Using this, the parameters like Accuracy, Precision, Recall and F-Measures are used and the corresponding values obtained through experiment is displayed in Table 2 with respect to different learning techniques.

Attribute/Classifier	Random Forest	Hybrid NB/DT
Accuracy	0.9797	0.9957
Precision Sensitivity	0.9881	0.9981
True Positive Rate	0.9786	0.9975
F-Measure Specificity	0.9733	0.9859
True Negative Rate	0.9712	0.9898
False Positive Rate	0.0064	0.0023
False Negative Rate	0.0114	0.0088

Table 2: Result of Different Classifiers

In table 2, the comparison results were obtained using the various classifiers used for this experiment is presented. Hybrid algorithm performance shows that the best result with a 0.9957 accuracy, followed by the Random Forest

classifier. The Random Forest provided the least accuracy of 0.9797.

Figure 1 illustrates the accuracy of the RF and Hybrid classifiers using 10 – folds cross validation. Hybrid used the shortest time (0.02 seconds) for classification while Random Forest classifier. Therefore, Hybrid also has the least execution time. The two models are considered as the execution time and classification accuracy, it is found that Hybrid takes very short computational time and outperforms all other classifiers. Figure 1 is the bar chart showing comparison of the two classifiers in terms of classification accuracy.

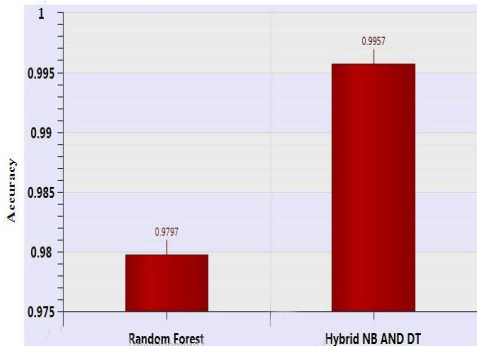

Figure 1: Comparison of accuracy

V. CONCLUSION

In this paper, we discussed a crucial review of pre-existing techniques and their statistical-based data mining algorithms like Hybrid Naïve Bayes – Decision Tree and Random Forest. By using these techniques effortlessly refer to order HR Talent Acquisition data into a set of Promotion Recommendation. So the performance of Hybrid Naïve Bayes – Decision Tree algorithm and other classification techniques were compared and analyzed. These approaches have implemented and tested using IT Tech Software from database administrator. As a result of the hybrid approach achieved over 99% classification accuracy on a test set to compare with other approaches. Future work ensures more proper data from several companies. When the appropriate model is generated, these algorithms could be developed for predicting performance of employees in any type of organization.

VI. REFERENCES

[1]. Tso, G.K.F. and K.K.W. Yau, "Predicting electricity energy comsumption : A comparison of regression analysis, decision tree and nerural networks". Energy, 2007. 32: p. 1761 - 1768.

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

[3]. hein, C.F., Chen, L.F., “Data Mining to improve personnel selection and enhance human capital: A case study in high technology industry, Expert Systems with Applications, 34(1), pp 280–290, 2008.

[4]. Valle, M.A., Varas, S., Ruz, G.A., “Job performance prediction in a call center using a Naive Bayes classifier, Expert Systems with Applications, 39(11), pp 9939–9945. 2012.

[5]. Han and Kamber, “Data Mining: Concepts and Techniques”, Second Morgan Kaufman Publisher, 2006.

[6]. David F. Giannetto, “Get Your Money’s Worth from Incentives,” Business Performance Management 7 (June 2009):

[7]. Stephen Garcia, “Forced Rankings of Employees Bad for Business,” Machine Design 79 (September 13, 2007): 4–5.

[8]. Tom Krattenmaker, “Appraising Employee Performance in a Downsized Organization,” Harvard Management Update 14 (May 2009): 3–5.

[9]. Han, J., Kamber, M., Jian P, “Data Mining Concepts and Techniques”, San Francisco, CA: Morgan Kaufmann Publishers, 2011.

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

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

[12]. Zhang H P, Wang M H, “Search for the smallest random forest”, Stat. Interface l2 381-8, 2009

[13]. Robnik-Sikonja M, “Improving Random Forests”, Proceedings of the 15th European Conference on Machine Learning 359-70, 2004.

[14]. Jiang, Liangxiao, and Chaoqun Li. "Scaling Up the Accuracy of Decision-Tree Classifiers: A Naive- Bayes Combination." Journal of Computers 6.7 (2011): 1325-1331.