



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

TALENT KNOWLEDGE ACQUISITION USING C4.5 CLASSIFICATION ALGORITHM

Anitha Mary Florence.T¹ and Ms.Savithri.R²

¹PG Student, ²Assistant Professor,
Department of Information Technology
Rajalakshmi Engineering College, Chennai

Abstract - In HR field, monitoring skills of talented professionals is one of the challenges to be dealt and handled by HR professionals, especially to ensure the right person for the right job at the right time. This task involves a lot of managerial decisions that depends on various factors such as human experience, knowledge, preference and judgment. All these factors can cause inconsistent, inaccurate, inequality and unpredictable decisions. In this case, knowledge discovered from talent knowledge acquisition process can be used by professionals in HR to handle various tasks in talent management. In the proposed system the C4.5 classifier algorithm is used to identify the skill set in order to evaluate the performance of the individual. This technique will be used to construct classification rules to predict the potential talent which helps in determining whether the individual is fit for the appraisal or not.

Keywords— Talent Knowledge Acquisition, Classification, HR Dataset, C4.5 Algorithm

I. Introduction

Knowledge acquisition is a crucial process in knowledge management and knowledge engineering fields. The acquisition process can be implemented by elicitation, collection, analysis, validation of knowledge. The process to identify an existing talent in organization is among the top challenges faced by an organization. Classification is among popular technique in Data mining especially for knowledge discovery and future plan. The decision tree algorithm is among the powerful classification algorithm that has its advantage such that it can produce a model that represent interpretable rules or logic statement that is more suitable for analyzing categorical outcomes. The C4.5 classification algorithm is easy to understand as the derived rules has straightforward interpretation. Due to these reasons this study has aimed to use this classification algorithm to handle issues on Human Talent Acquisition. Talent is considered as the capability of any individual to make a difference in current and future performance of an organization. Besides that Talent Management can be defined as the process to ensure leadership continuity in key positions and encourages individual advancement. This study ensures talent knowledge acquisition by using employee's performance records. This can be done by identifying talent patterns from existing data in HR databases as useful and valuable knowledge.

II. Related Works

Data Mining is the process of discovering new patterns from large data sets [1]. The goal of data mining is to extract knowledge from a data set and present it in a human-understandable structure. Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups [2]. Classification methods make use of mathematical and statistical techniques such as decision trees, linear programming, neural network and support vector machines. In this section, we first review a few of the related work on HR Talent Acquisition using data mining techniques.

Ruskova et al.[3] presents a Decision Support System (DSS) based on fuzzy logic model for Human Resources Appraisal and Selection (HRAS). DSS can be used in web based human resource solutions (job-search, enterprise portal, etc.). This study presents the forming the personnel of an organization or a project team as a Fuzzy decision making process. Huang et al. [4] proposed the appointment of managers in enterprises through fuzzy neural network, to construct a new model for evaluation of managerial talent and accordingly to develop a decision support system in human resource selection. This research method is collecting expert data as a questionnaire survey and the fuzzy

neural network trains the concrete database based on the questionnaire for getting best network model in different training conditions.

Wei-Shen et al[5] in their work applied linguistic variables and fuzzy membership function to integrate heterogeneous data and simplify the computational complication during data mining process. Their work suggest to partition each attribute by fuzzy partition method with triangle membership function in their tool been built. Hamdan et al. [6] in this paper proposed the HR DSS framework and the potential intelligent techniques that can be used to develop the IDSS application in any phases of decision making processes. They described the potential to implement the IDSS in Human Resource Management (HRM) activities. Jantan et al. [7] in their work propose solution for some talent management problems using Data Mining techniques. They identified potential talent of employees by predicting their performance using past experience knowledge. In this study the focus is on identifying the patterns that relate to the talent and finding the potential mining technique for talent forecasting. Their results show that decision tree or neural networks are best suited for talent forecasting.

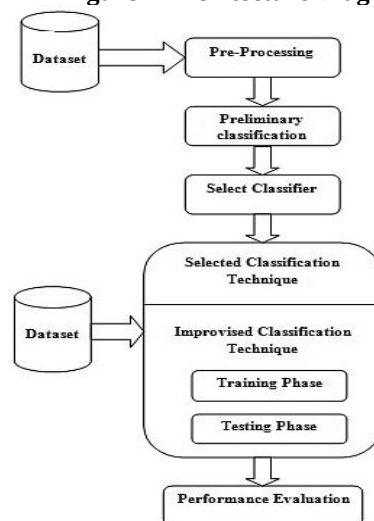
Jantan et al. [8] suggests the potential HR system architecture based on the data mining technique to solve the talent management task by identifying the patterns that relate to the talent by using prediction technique by producing prediction rules to identify the talent. Zhang et al. [9] comprises the view of Yihua et all in explaining performance evaluation concept of HR management using Data mining technology. This study presents the comprehensive evaluation indicators by means of clustering methodology to design and analyse performance evaluation and evaluates their overall level. It assesses the employees' performance from three perspectives professionalism, the work performance and capacity. Jantan et al. [10] In their study solved the problem of talent management by suggesting the classification and prediction techniques. They demonstrated an example to suggest the feasibility of the classification technique for the academicians' talent data. The accuracy of classifier is based on the percentage of test set samples that are correctly classified. The second phase of experiment is to compare the accuracy of classifier for attribute reduction. Here the attempt is to study the accuracy of the classifier for all important attributes. The results for full attributes show the highest accuracy for C4.5(95.1%)..

Othman et al. [11] in their study Jantan et all compares the various classifier algorithms based on decision trees to facilitate talent management. Here they predict the talent performance by means of knowledge acquired from the existing databases. The performance factors for this study were based on the academicians' performance. Their study focused on identifying pattern relating to academicians' performance. Finally their research resulted by stating C4.5 as the suitable classifier for the dataset especially for forecasting talent. This work predicts the error rate of a learning technique. Boolean reasoning technique is been used for the shortest length reduction. But attribute reduction techniques were not preferred.

III. Proposed Work

The processing steps applied to the HR data are given in the following Fig.1

Figure 1 Architecture Diagram



Data Pre-processing is the first step that was carried out on the HR dataset to prepare the data for mining. Pre-processing includes data cleaning and filling-in of missing values in the data. Feature selection is applied to the dataset to identify the key attributes and obtain a reduced relevant subset of key attributes to be used in the classification exercise. Classification exercise was carried out with different data mining algorithms to identify the best and most suitable classification algorithm. The chosen algorithm was then improvised to obtain the best classification rate.

A.PREPROCESSING

Data pre-processing was applied to HR data to prepare the raw data [8]. Pre-processing is an important step that is used to transform the raw data into a format that makes it possible to apply data mining techniques and also to improve the quality of data [9]. It can be noted from the related work, that attribute selection plays an important role in identifying parameters that are important and significant for proper breast cancer diagnosis. It was also found that the prediction quality was retained even with a small number of non-redundant attributes.

B.PRELIMINARY CLASSIFICATION

A preliminary analysis was carried out on the 200 data records with different classification techniques using Promotion Recommendation as target class and above mentioned continuous attributes as input attributes.

The classification techniques were applied with the above mentioned continuous attributes. The comparison of error rate obtained for the various classification techniques is given in Table 1.

Table 1: Comparison of Classification Techniques

S. No	Technique	Error Rate
1	C-RT	0.16
2	CS-MC4	0.14
3	C 4.5	0.02
4	ID3	0.48
5	K-NN	0.02
6	LDA	0.12
7	NAIVE BAYES	0.18
8	PLS-LDA	0.12
9	RND TREE	0.0
10	SVM	0.16

It is seen from the table that RND TREE algorithm has the lowest error rate of 0.0%. The rule set, however was found to be too large and unwieldy and hence becomes difficult to apply to any dataset. Classification Algorithms KNN and C4.5 gives ~98% classification rate.

The classification rules obtained from the training set in the training phase is given below.

Efficiency < 6.5000 then Promotion_Recommendation = N

Efficiency >= 6.5000

C_Language_Skill_Level < 6.5000 then Promotion_Recommendation = N

C_Language_Skill_Level >= 6.5000

Journals < 0.5000 then Promotion_Recommendation = N

Journals >= 0.5000

Java_Skill_Level < 5.5000

Workshop < 1.5000 then Promotion_Recommendation = N

Workshop >= 1.5000 then Promotion_Recommendation = Y

Java_Skill_Level >= 5.5000 then Promotion_Recommendation = Y

C.CLASSIFICATION USING C4.5

C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the issues not dealt adequately by ID3 [19]. These include avoidance of over fitting the data; reduced error pruning, rule post-pruning, handling continuous attributes and handling data with missing attribute values [20]. C4.5 classification technique uses entropy and information gain for tree splitting. In testing phase we used training data with known result and. C4.5 algorithm was applied to obtain the rule set. In the testing phase, the classification rules obtained were applied to the whole pre-processed data. The results obtained are analysed.

1) Training Phase:

The training data sample involved a range of 15 attributes about which 14 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 2 summarizes the list of attributes used for the talent acquisition process.

Table 2: Talent Dataset

TALENT ATTRIBUTES	DESCRIPTION
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
C Language Skill Level	Knowledge and Skill
Java Skill Level	Knowledge and Skill
.Net Skill Level	Knowledge and Skill
ITES Skill Level	Knowledge and Skill
International Conferences	Activities and Contribution
Journals	Activities and Contribution
Workshop	Activities and Contribution
Promotion Recommendation	Activities and Contribution

2) Testing Phase and Performance Analysis

In testing phase we applied the C4.5 classification rules obtained from random Sample to the complete 200 records. The actual and predicted values obtained in the classification exercise are shown in the confusion matrix given in Table 3. The tabular representation of the predicted values for the sensitivity and specificity are given in Table 4.

Table 3: Confusion Matrix for Talent Acquisition

0.0200			
Confusion Matrix			
	Y	N	Sum
Y	99	1	100
N	3	97	100
Sum	102	98	200

Table 4: Error Prediction for Talent Data

Error rate	
Values Prediction	
Value	Recall
Y	0.9705
N	0.9897

It can be observed from the confusion matrix, that 4 of 200 records are classified ambiguously. 3 of the YES cases have been classified as NO (false negatives). 1 of the NO case have been classified as YES (false positives). Some of the performance measures are now evaluated. Accuracy is the percentage of records correctly classified out of the total records.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{(99+97)}{(99+97+1+3)} = 0.98$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} = \frac{99}{(99+3)} = 0.97$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} = \frac{97}{(97+1)} = 0.99$$

IV. CONCLUSION

This system attempts to classify HR Talent Acquisition data into the groups of Promotion Recommendation using C4.5 classification algorithm. We used a training set of a random sample of 200 records and then applied the classification rule set obtained to the Human Resource dataset. The obtained dataset here is 98% in the training phase. We have compared the performance of C4.5 algorithm with other classification techniques. Future enhancement of this work includes improvisation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy.

References

- [1] SoumenChakrabarti, Earl Cox, Eibe Frank, Ralf HartmutGüting, Jaiwei Han, Xia Jiang, MichelineKamber, Sam S. Lightstone, Thomas P. NadeauRichard E. Neapolitan, Dorian Pyle, MamdouhRefaat, Markus Schneider, Toby J. Teorey, Ian H. Witten, "Data Mining-Know it all", Morgan Kaufmann Publishers, 2009.
- [2] J. Han and M. Kamber, - "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, 2000.
- [3] Nina At. Ruskova, "Decision Support System for Human Resources Appraisal and Selection, First International IEEE Symposium "Intelligent Systems" September 2002.
- [4] Huang, L.C., et al, Applying fuzzy neural network in human resource selection system. inProceeding NAFIPS '04, IEEE Annual Meeting of the Fuzzy information 2004. 2004.
- [5] Tai, W.S. and C.C. Hsu (2005), A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method.
- [6] Jantan, H.; Hamdan, A.R.; Othman, Z.A, Potential Intelligent Techniques in Human Resource Decision Support System (HR DSS),IEEE 2008.
- [7] Hamidah Jantan, Abdul Razak Hamdan, Zulaiha Ali Othman, Towards applying Data Mining Techniques for Talent Management,2009 International Conference on Computer Engineering and Applications.
- [8] 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.
- [9] Zhang Yihua Wang Yuan*,Research and Application of the Data Mining Technology on the modern enterprise performance evaluation system, 2009, International Conference on Information Management, Innovation Management and Industrial Engineering.
- [10] Jantan, H.; Hamdan, A.R.; Othman, Z.A, Classification for Talent Management Using Decision Tree Induction Techniques2nd Conference on Data Mining and Optimization27-28 October 2009, Selangor, Malaysia.
- [11] Hamidah Jantan, Abdul Razak Hamdan and Zulaiha Ali Othman (2011). Talent Knowledge Acquisition using Data Mining Classification Techniques. 3rd Conference on Data Mining and Optimization (DMO).
- [12] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers,1993.
- [13] http://en.wikipedia.org/wiki/C4.5_algorithm - C4.5 Algorithm Description