Calculating Decision Rules for Choosing Appropriate Candidates of a Job Using C 4.5 Algorithm

Khin Thuzar Tun[1], Ei Chan Lin[2]

[1, 2] *Department Of Information Technology, Technological University (Mawlamyine)*

***Abstract*- Hiring the right employee for a job profile is one of the most essential business processes which affect the quality of human capital within any company. It is highly essential for the companies to ensure the recruitment of candidate for a particular job, which in turn provide qualified workforce for the organization. CV analysis can be one of the most time consuming parts of a recruiter's job, requiring to decide considerable skills accurately as well as quickly. The proposed system aims to make choosing appropriate candidates based on job experience and other key skills using data mining technique with algorithm C4.5. As a result, this system could produce the decision rules selecting suitable candidates for the desired company or job.**

***Keywords*- Data Mining, Decision Tree, C4.5, Recruitment Candidates, Decision Rules**

I. INTRODUCTION

Hiring the right person for the right job is a common challenge faced by all companies. Searching the right candidate from a large number of applicants can feel like looking for a needle in the haystack. In these situations, traditional methods of recruitment can be too expensive and time-consuming to be a viable option. This means the employee selection, (i.e. choice of right candidates for a position), is the fundamental part of many industry areas and represents an important part of each sophisticated system of the human resource management. In general, the selection of employee for the job is the process not only high organizational and intellectual complexity but also accurate and fast response.

Selecting from the group of candidates involves two main processes: (a) listing of suitable candidates; (b) making decision to choose the candidates who dispose convincingly and truthfully to be served the most efficient and successful work for job offered. Decision on the suitable candidate is performed as a sequence of selection procedures and consists in the comparison of selection criteria with the predictors, i.e. expressed and confirmed abilities, skills, knowledge, competences, and overall creative and capable potential of the candidates.

Several researches have already been carried out in area of supported-by-software choice of suitable candidates to specific job position. The proposed system aims to use C4.5 decision tree algorithm to make decision and calculate talent of listed candidates. Also presents the forecast of job candidates' talent by decision trees technique: C4.5 algorithm. This system deals with the concept of well-known decision trees to choose suitable candidates for recruiting, selecting and assigning steps to be continually conducted. Summing up, the calculation of decision rules based on C4.5 algorithm is mainly presented for the selection of proper applicants in this paper.

II. RELATED WORKS

Human resource has become one of the main concerns of managers in almost all types of businesses which include private companies, educational institutions and governmental organizations. Chein and Chen have used several attributes to predict the employee performance. The specification of age, gender, marital status, experience, education, major subjects and school tires as potential factors might affect the performance. As a result of study, Chein and Chen found that employee performance is highly affected by education degree, the school tire, and the job experience [1].

Several studies used data mining for extracting rules and predicting certain behaviors in several areas of science, information technology, human resources, education, biology and medicine. For example, Beikzadeh and Delavari [2], data mining techniques has used for suggesting enhancements on higher educational systems. Al-Radaideh et al., [3] data mining techniques also used to predict university students' performance. Many medical researchers used data mining techniques for clinical extraction units using the enormous patient data files and histories, Lavarc was one of such researchers [4]. Mullins et al. used to extract disease association rules using unsupervised methods based on patients' data [5].

The researcher Kahya [6] reviewed previous studies, describing the effect of experience, salary, education, working conditions and job satisfaction on the performance. As a result of the research, it has been found that several factors affected the employee's performance. The position or grade of the employee in the company was of high positive effect on performance. Working conditions and environment, on the other hand, had shown both positive and negative relationship on performance. Highly educated and qualified employees showed dissatisfaction of bad working conditions and thus affected the performance negatively. In addition, experience showed positive relationship in most cases, while education did not yield clear relationship with the performance.

In the study of Salleh et al., [7], the authors have 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.

The same authors, Jantan et al., [8] have used decision tree C4.5 classification algorithm to predict human talent in HRM, by generating classification rules for the historical HR records, and testing rules on unseen data to calculate accuracy. The authors intend to use these rules in creating a DSS system that can be used by managements to predict employees' performance and potential promotions.

III.CLASSIFICATION TECHNIQUES IN DATA MINING

Classification is a data mining technique that assigns items in a collection to target categories. The objective of classification is to accurately predict the category which is unknown for each case in the data. For example, a classification model could be used to identify student results as pass, good, very good or excellent. Classification algorithms can be applied to many applications such as biomedical and drug response modeling, customer segmentation, business modeling, marketing and credit analysis, etc. Accuracy of the model refers to the percentage of correctly classified instance made by the model when compared with the actual classification in the test data.

There are different types of classification models. They are as follow:

- Decision Tree Method
- Rule-based Method
- Nearest Neighbor
- Neural Network
- Deep Learning
- Naïve Bayes
- Support Vector Machines

A. Decision Tree Method

Decision tree method is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal node, and leaf node.

The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model.

Decision trees are built of nodes, branches and leaves that indicate the variables, conditions, and outcomes, respectively. The most predictive variable is placed at the top of the tree. One of the most significant advantages of decision trees is the fact that knowledge can be extracted and represented in the form of classification (if-then) rules. Each rule represents a unique path from the root to each leaf. A decision tree is used to identify the strategy most likely to reach a goal [9].

There are various decision tree algorithms: ID3 (Iterative Dichotomiser), C4.5, CART (Classification and Regression Tree), MARS, CHAID (CHi- squared Automatic Interaction Detector). Among them, the most popular algorithms are ID3, C4.5, CART.

1) C4.5 Algorithm: C4.5 is a successor to the ID3 algorithm, which has much better performance when operating on numeric continuous attributes in training dataset. To overcome prior limitations with continuous attributes, the algorithm dynamically defines a discrete attribute that splits continuous numeric values into two branches that represent the attributes, partitioning the continuous attributes value into a discrete set of intervals. C4.5 algorithm transforms the tree model from ID3 into sets of it-then guidelines for classification. The advantages and disadvantages are as follows:

There are four advantages.

- C4.5 is easy to implement.
- C4.5 builds model that can be easily interpreted.
- It can handle both categorical and continuous values.
- It can deal with noise and deal with missing value attributes.

There are two disadvantages.

- A small variation in data can lead to different decision trees when using C4.5.
- For a small training set, C4.5 does not work very well [10].

1) Implementation of Decision Tree C4.5 Algorithm: The following steps are performed.

- Prepare the data training
- Determine the roots of the tree.
- Calculate the gain vale.
- Repeat step 2 until all tuples are partitioned.
- The decision tree partition process will stop when all the tuples in the N node get the same class and or no attribute in the tuple being partitioned again and or no tuples in the empty branch.

Calculation steps are conducted by the following equations.

First, calculate overall information gain,

Information Gain = $\sum_{i=0}^{n} -P_i \, log_2(P_i)$

Then, compute information gain for each attribute and Entropy for each.

$$\text{Entropy} = -\sum_{i=1}^{n} \frac{P_{1(i)} + P_{2(i)} + P_{3(i)}}{P_1 + P_2 + P_3} \times IG(P_1, P_2, P_3)$$

Next, Gain is obtained the subtraction of entropy of each attribute from overall information gain. Gain is also needed to compute for each attribute.

Gain = Information Gain – Entropy

The largest gain value among them is the initial split level and the root node of decision tree. Continue the calculation to find the internal nodes until the reach of the leaf nodes in this way.

IV. FLOW CHART OF THE PROPOSED SYSTEM

Fig. 1 shows the flow chart of the proposed system. Firstly, define the worker information (desired skills) which wants to need for the job or company into the database. The proposed system computes the decision rules according to C4.5 decision tree steps in order to predict the appropriate candidates continually. In this paper, the decision rules as the result are produced in the form of decision tree from the root node to leaf nodes.

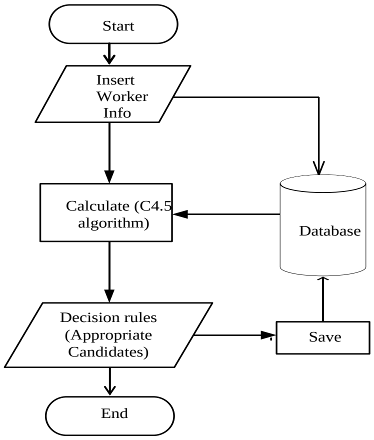
Start
Insert Worker Info
Calculate (C4.5 algorithm)
Database
Decision rules (Appropriate Candidates)
Save
End

Fig.1Flow chart of the proposed system

A. Example Data Set Description

The proposed system was tested in a real recruitment scenario, to get its effectiveness in selecting job applicants. The system's performance evaluation is based on how effective it is in assigning consistent relevance scores to the candidates, compared to the ones assigned by human recruiters.

In the recruitment scenario used in test, a job position announced by an IT company with job requirements is expressed in the following:

Job position - Programmer

Job requirement - B.E. (IT) or B.C.Sc

- minimum 2 years experienced

- must have programming knowledge and English skills.

The training job attributes (gender, education, experience, other certification) are assigned with the possible values initially. A decision tree is implemented based on these training data using C4.5 algorithm.

Table I POSSIBLE VALUE OF EXAMPLE DATASET

Attribute	Description	Possible Value
gender	Gender	Male Female
Edu	Education	BE(IT) B.C.Sc Other
Exp	Experience	Exp < 2("a") Exp >= 2("b")
Certi	Other certification	Certi = Match("M") Certi ≠ Don't match("DM")

The combination of various attributes determines whether the candidate is recruited or not. The possible values and detailed description of these attributes are mentioned in Table I and II.

The following set of data is considered as the above job position dataset for the proposed system. The dataset comprises of different measures of 12 candidates. The attributes such as Gender,

Education, Experience and Other Certification have been taken into consideration. On the basis of the training set, the entropy and the information gain is computed to resolve the splitting aspect for establishing a decision tree.

The example job positions with possible values are required to assign according to the requirement of each job. For example, Accountant (B.Com/minimum 2 years experienced/LCCI I, II, III), Site Engineer (B.E. (Civil minimum 4 years experienced/English), Japanese Translator (B.A. (Japanese)/minimum 2 years experienced/N2), etc. Decision tree: C4.5 algorithm is used as a support tool in process of candidate selection worker's training data in Table II and are already grouped in specific classes.

TABLE II: EXAMPLE TRAINING CANDIDATE'S INFORMATION

No	Gender	Edu	Exp	Certi	Outcome
1	Male	BE(IT)	a	M	No
2	Male	BE(IT)	b	M	Yes
3	Male	B.C.Sc	b	M	Yes
4	Male	BE(IT)	a	DM	No
5	Male	Other	b	M	No
6	Female	B.C.Sc	b	M	Yes
7	Female	B.C.Sc	a	DM	No
8	Female	BE(IT)	b	M	Yes
9	Female	Other	b	M	No
10	Female	BE(IT)	b	DM	No
11	Female	BE(IT)	b	M	Yes
12	Female	B.C.Sc	a	M	No

In this study, two groups (Yes, No) are assigned in the outcome attribute. The root of a tree is firstly searched by calculating the highest gain among the calculated gain of each attribute. Before this gain, calculates the overall information gain, information gain and entropy of the including every attribute. Repeat these steps until all of the attributes divided. This process will stop when there are no attributes in the system that partitioned completely. The resulting decision tree for example calculation of this study is shown in Fig 2.

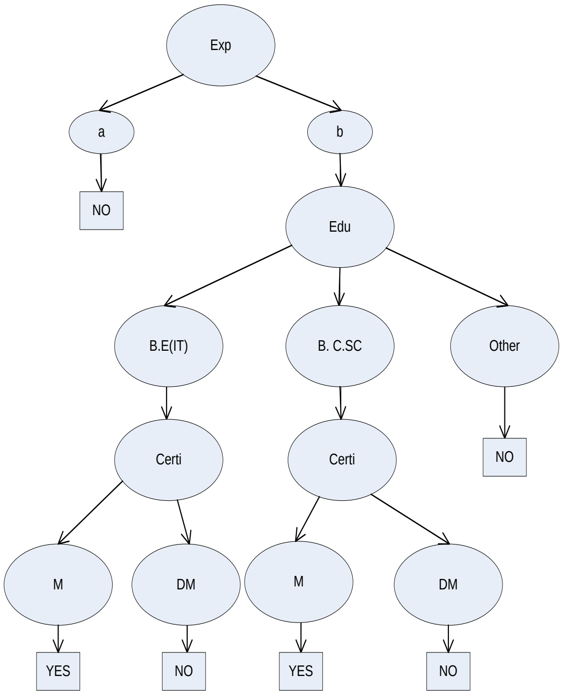

Fig. 2 Final decision tree of the proposed system

V. CONCLUSIONS

In this paper, the classification rules are developed based on the sets of attributes used in IT companies. This tree has been inducted according to the algorithm C4.5. The result of this study provides to select suitable employees for the specific job in an organization efficiently and within a short period of time. The system will help the employers to find appropriate candidates and post the respective job positions. Interactivelly, this study is also useful for job seekers to get the deserved jobs and ranks of them more easily.

VI. FURTHER EXTENSION

The proposed system mainly presents the calculation of decision tree with example dataset. Therefore, it is more suitable if the large public dataset: Human Resource Analysis is applied. Besides that, other data mining technique such as Fuzzy decision tree can be used to perform in further implementation for classification accuracy.

REFERENCES

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

[2]. Delavari, N., PHON-AMNUAISUK S., (2008). Data Mining Application in Higher Learning.

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

[4]. Lavrac, N, (1999). Artificial Intelligence in Medicine :Selected Techniques for Data Mining in Medicine, pp. 3-23.

[5]. Mullins, I., Siadaty, M., Lyman, J., Scully, K., Garrett, C.,

[6]. Millar, W., Mullar, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., Knaus, W, (2006). Computers in Biology and Medicine: Data Mining and Clinical Data Repositories: Insights from 667,000 Patient Data Set, pp. 1351-1377.

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

[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, (2010). Human Talent Prediction in HRM using C4.5 Classification Algorithm, International Journal on Computer Science and Engineering, 2(08-2010), pp. 25262534.

[10]. https://em.m.wikipedia.org/wiki/Decision-tree.

[11]. Bhumika, AdityaRawat, Akshay Jain, ArpitArora and NareshDhami, April, (2017). Analysis of Various Decision Tree Algorithm for classification in Data Mining, International Journal of Computer Applications (0975-8887), Volume 163, No.8.