DOI: 10.21917/ijsc.2011.0022

KNOWLEDGE ENGINEERING TO AID THE RECRUITMENT PROCESS OF AN INDUSTRY BY IDENTIFYING SUPERIOR SELECTION CRITERIA

N. Sivaram[1] and K. Ramar[2]

[1]*Department of Computer Science and Engineering, National Engineering College, Tamil Nadu, India*
E-mail: Sivaram.Natarajan@airtelmail.in
[2]*Sri Vidhya College of Engineering and Technology, Tamil Nadu, India*
E-mail: kramar_nec@rediffmail.com

Abstract
Recruitment of the most appropriate employees and their retention are the immense challenges for the HR department of most of the industries. Every year IT companies recruit fresh graduates through their campus selection programs. Usually industries examine the skills of the candidate by conducting tests, group discussion and number of interviews. This process requires enormous amount of effort and investment. During each phase of the recruitment process, candidates are filtered based on some performance criteria. The problem domain is complex and the aspects of candidates that impact the recruitment process is not explicit. The intelligence of the recruitment process is spread among the domain experts and extracted through knowledge acquisition techniques. This research focuses on investigating the underlying criteria and tries to capitalize on the existing patterns, to minimize the effort made during the recruitment process. The approach here is to provide the insights through in-depth empirical characterization and evaluation of decision trees for the recruitment problem domain. Experiments were conducted with the data collected from an IT industry to support their hiring decisions. Pruned and unpruned trees were constructed using ID3, C4.5 and CART algorithms. It was observed that the performance of the C4.5 algorithm is high. The recruitment process differs for each industry based on the nature of the projects carried out. Experiments were conducted to determine the attributes that best fits the problem domain. Using the constructed decision trees discussions were made with the domain experts to deduce viable decision rules.

Keywords:
Recruitment Process, Decision Trees, Selection Criteria, Machine Learning

1. INTRODUCTION

Human resource is one of the back bones for industries to maintain their competitive advantages in the knowledge economy. Selecting fresh people with high talent and potential retention is challenging and daunting task faced by any HR department. The demand is increasing year over year and the supply has also been on the raise. But, the problem is with the quality in the supply for its recruitment. Different companies are fighting it out by taking the first slot, opting for dual placements, addressing the gaps by faculty development programs, etc. Campus recruitment is the predominant mode of recruitment for fresh talented graduates. Because of the inconsistency in the quality of the students produced by different universities and the type of skill set they acquire during their program, selecting the right candidate among those who graduate becomes a herculean task. This involves lot of effort by the recruiting team and money spent for the process is phenomenal. One of the mechanisms used by the industries is to conduct tests and group discussions during the filtration process. The selection process uses different criteria which include the average of their semester marks, marks obtained in the aptitude, programming & technical tests conducted by the company, group discussion, technical and HR interviews. These criteria are common for all the students, but the skill level of the students vary since they are from different disciplines and backgrounds. The time taken and expenditure for conducting group discussions and interviews consumes more than 90% of the total effort of the recruitment process. It has been observed that 1 among 120 students who apply get selected and the ratio of number of candidates selected against the number of candidates interviewed after tests is approximately 1:20. Reducing these ratios will immensely help the industries to save the effort. This research focuses on determining a set of selection criteria to be applied to filter candidates based on their background and academic data. The study is made using the recruitment database maintained in the IT industry and addresses various key research issues of the domain as discussed in Section 2.

Data mining is a process that explores huge quantities of data to discover meaningful rules and patterns. A number of machine learning, knowledge engineering, and probabilistic-based methods have been proposed to analyze the data and extract information. The most popular methods include Bayes' theorem, regression analysis, neural network algorithms, clustering, genetic algorithms, decision trees and support vector machines [3], [4]. Data mining tools assist experts in the analysis of observations of behaviour. Such data are vulnerable to co-linearity because of unknown interrelations. It is factual in data mining that the subset of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships that exist across other parts of the domain. To tackle this issue, the analysis could be augmented by Design Of Experiment (DOE)-based or choice modeling methods for human-generated data. In such case, during the experimental design, inherent correlations are either forbidden, or removed altogether. Decision tree is a simple data mining approach used to establish the hidden knowledge in the data for classification and prediction. They have the advantage of easy interpretation and understanding for the decision makers to compare with their domain knowledge for validation and justify their decisions [1].

A decision tree based approach has been proposed in the paper, to identify the relation between the applicant's data and their probability of selection. In this study, demographical details of candidates such as name, gender, date of birth, place of birth, details of schooling, performance in the school public exams, percentages obtained during graduation, stream of graduation, test marks obtained in the recruitment process are used to extract the hidden pattern. Using the knowledge

obtained, a set of selection criteria is generated which could help reducing the recruitment effort of the company.

2. MOTIVATION

This study is intended to analyze the issues involved in the recruitment process of fresh graduates, and find out a way to reduce the time and cost involved. Decision tree is chosen to solve the problem, since it is simple and easy to construct and analyze them. The research questions include

- Which attributes of the candidates used in the selection process impacts the selection of the candidate?
- Whether automatic decision tree generation is a feasible approach for the problem?
- Which among the various decision tree construction methods best fits the problem?
- What is the accuracy and complexity of decision trees when the entire object attributes versus subset of them is used in the construction of trees? (This analysis is made since right set of input attributes influences the performance of most of the classifiers)
- Whether pruning during construction of trees help to give better solution for the problem?
- What is the accuracy and complexity of the constructed tree, when all the instances versus equal number of instances for each class are given for learning? (This analysis is made to determine whether large number of instances in a particular class misleads the learning process).

3. BACKGROUND

This section provides an overview of the system, upon which the analysis is made and an introduction to the various decision tree construction methods.

3.1 SELECTION PROCESS

Selecting the right persons for the right job is the most important challenge in the human resource management. The various selection methods include analysis of application form, self-assessment, telephone screening, tests depending on the requirement of the industry (such as aptitude, technical, programming, personality, interest test, etc.) [6], [7]. Generally industries use a combination of the selection methods, based on their job nature, cost, time, accuracy, culture and acceptability. According to Lewis, there are three aspects of selection criteria. They are organizational criteria, functional/departmental criteria and individual job criteria. Finally, the recruitment committee must consider the adaptation of the job, departmental and organizational characteristics to the applicant's characteristics [2]. Hence the recruitment committee designs each level of the recruitment process to reflect their needs.

The recruitment process in the campus interviews of an IT industry includes filtering based on their semester marks in the graduation, marks obtained in the aptitude, technical and programming tests conducted by the industry, grade obtained during group discussion and interviews. The company prepares a set of questions to test, if the candidate is really capable of applying what he/she learnt in his/her course of study. The questions also map to the expectations and job description for which he is recruited. To check his presentation, communication and behavioral skills a Group Discussion is conducted.

3.2 DECISION TREES

Decision tree is a tree structure, where internal nodes denote a test on an attribute, each branch represents the outcomes of the test and the leaf node represents the class labels. Decision tree induction is the learning of decision trees from class-labeled training tuples. Construction of decision trees is simple and fast, and does not need any domain knowledge and hence appropriate for exploratory knowledge discovery. In general, decision tree classifiers have good accuracy, but successful use of it depends on the data at hand. Decision trees are used for classification and classification rules are easily generated from them. An unknown tuple X can be classified, given its attribute values by testing the attribute values against the decision tree. The general decision tree algorithm takes the training data set, attribute list and attribute selection method as input. The algorithm creates a node, and then applies attribute selection method to determine the best splitting criteria and the created node is named by that attribute. Subset of training tuples is formed using the splitting attribute. The algorithm is called recursively for each subset, till the subset contains tuples of same class. When the subset contains tuples from the same class a leaf is attached with a label of the majority class in the training set from the root. ID3, C4.5, and CART adopt a greedy, non-backtracking approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner [3],[11]. The three methods vary in the splitting criterion used to partition the data. All the three construction algorithms are applied for the problem and are evaluated in this paper.

3.2.1 ID3 algorithm

ID3 is an iterative algorithm that uses information gain as splitting criterion to construct the tree. For each attribute A, the method calculates the information gain as the difference between the information required to classify the data set based on just the proportion and the information required to classify after partitioning on A. The expected information needed to classify a tuple in the training set D is given by (1) [3], [9]:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i, and is estimated as the ratio of number of instances in class C_i in D to the total number of instances in D. The amount of information still required to classify D, after splitting them using A with v possible values is calculated using (2).

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2)$$

Information gain obtained by branching the training set on the attribute A is given as in (3).

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

The algorithm is recursively applied for the subsets till all the members of the set belongs to the same class.

3.2.2 C4.5 Algorithm

C4.5 algorithm is a successor of ID3 that uses gain ratio as splitting criterion to partition the data set. The algorithm applies a kind of normalization to information gain using a "split information" value. Split information for an attribute A with v values is defined as in (4) [3]:

$$split\inf_A(D) = -\sum_{i=1}^{v}\frac{|D_i|}{|D|}\times\log_2\left(\frac{|D_i|}{|D|}\right) \quad (4)$$

where $|D_i|$ is the number of instances in the training set D with i^{th} value for the attribute A and |D| is the total number of instances in the training set. Gain ratio is defined as in (5) and the attribute with maximum gain ratio is selected as the splitting attribute [3].

$$Gainratio(A) = \frac{Gain(A)}{Split\inf(A)} \quad (5)$$

3.2.3 CART Algorithm

CART is a recursive partitioning method that builds classification and regression trees for predicting continuous dependent variables and categorical predictor variables. The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are purer than the data in the parent subset [10]. Gini index is used to measure the impurity of D, the set of training tuples as given in (6).

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (6)$$

where p_i is the probability that an instance in D belongs to class C_i and is estimated using (7)

$$p_i = \frac{|C_{i,D}|}{|D|} \quad (7)$$

$|C_{i,D}|$ is the number of instances in D that belong to category C_i and |D| is the total number of instances in the training set. Gini index uses binary split for each attribute, for an discrete attribute A with v known distinct values, $P = \{a_1,a_2,a_3,\ldots,a_v\}$, best binary split is determined by examining all possible subsets of P. For each subset S of P, a binary test of attribute A of the form $A \in S$ is performed, given an instance I, this test is satisfied if the value of A for I is in S. There are 2^v-2 possible ways, to form two partitions of the data, D, based on a binary split on A, after eliminating the empty set and the set P. For each binary split, the weighted sum of the impurity of each resulting partition is calculated using (8). The gini index of a binary split on A that partitions the training set D into D_1 and D_2 is

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \quad (8)$$

For a discrete valued attribute, the split that gives the minimum gini index is selected as its splitting attribute [3].

For a continuous valued attribute, the point giving minimum gini index is chosen as the split point of the attribute. The set of possible split points are, determined by sorting the values and then by taking midpoint of the adjacent values. Using (8) gini index is calculated for the attribute, where D_1 is the set of instances with value of A less than or equal to split point and D_2 is the set of instances with value of A greater than split point. The reduction in impurity incurred by a binary split on a discrete or continuous valued attribute A is given as in (9)

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad (9)$$

The attribute that maximizes the reduction in impurity is selected as the splitting attribute.

3.2.4 Tree Pruning

When a decision tree is built certain branches may reflect anomalies in the training data due to noise which is removed by the tree pruning techniques. The tree pruning techniques uses statistical measures to remove the least reliable branches. Postpruning and prepruing are the two common approaches. In the prepruning approach the tree is pruned by deciding not to further split the subset of training tuples at a given node. Postpruning techniques removes subtrees from a fully grown tree, by replacing a subtree with a leaf labeled as the most frequent class in it.

CART uses cost complexity pruning algorithm, a postpruning approach which assumes that the bias in the resubstitution error of a tree increases linearly with the number of leaf nodes. The pruning technique starts from the bottom of the tree. For each internal node, N, it computes the cost complexity of the subtree at N, and the cost complexity of the subtree at N if it were to be pruned, the two values are compared. If pruning the subtree at node N would result in a smaller cost complexity, then the subtree is pruned. This techniques uses a pruning set of class-labeled tuples is used to estimate cost complexity. This set is independent of the training set used to build the unpruned tree and of any test set used for accuracy estimation. The algorithm generates a set of progressively pruned trees. In general, the smallest decision tree that minimizes the cost complexity is preferred [3], [12].

C4.5 uses pessimistic pruning, similar to the cost complexity method uses error rate estimates to make decisions regarding subtree pruning. However, the method does not use a prune set, instead estimates error rates using the training set [3].

4. PROPOSED METHOD

This research aims at mining the recruitment data to explore the relationships between the historical data of people and the possibility of being recruited in the company. Through the proposed methodology, the hidden knowledge about the criteria to be fixed during the selection process could be extracted from the data of previous years. Fig 1 shows the steps involved in the mining process.

The mining process begins with the step to gather knowledge from the domain experts. Knowledge acquisition is a process that includes elicitation, collection, analysis, modeling and validation of knowledge for knowledge engineering. Some of the important issues involved in knowledge acquisition are the knowledge is hidden within the domain experts and is not with a single expert. Interviews were conducted with the domain experts to understand the problem and the knowledge required to solve the problem. The knowledge acquired is used along with the recruitment database maintained in the industry to form the dataset for experimentation.

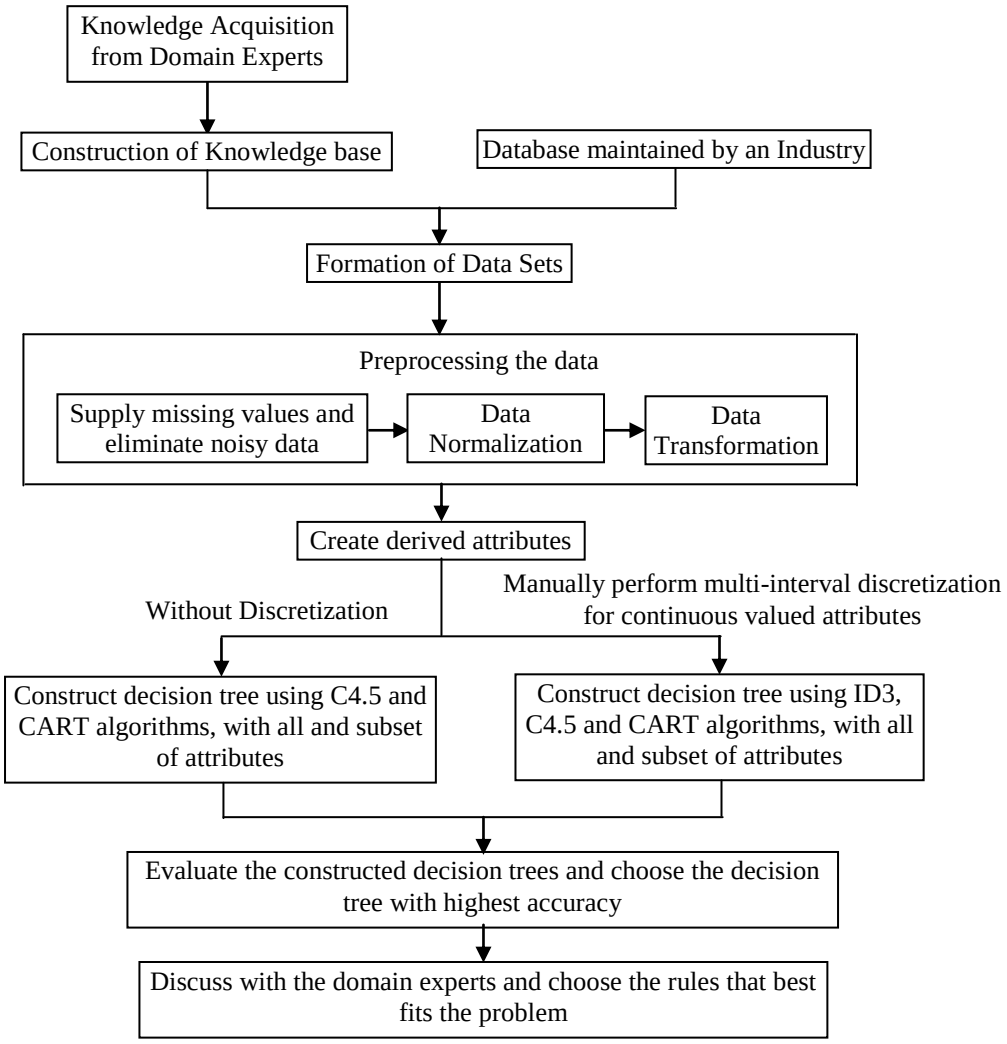

Fig.1. A Data mining framework for Recruitment Mining

The data collected from the industry is complex and have noisy, missing and inconsistent data. The data is preprocessed to improve the quality of data and make it fit for the data mining task. The data used are transformed into appropriate formats to support meaningful analysis. Some more attributes are derived using the acquired knowledge to support the mining process.

Decision trees were constructed with different construction methods such as ID3, C4.5 and CART. The data collected contains both nominal and continuous data; hence ID3 algorithm could be applied only after discretization. Numerical data were discretized to multiple intervals manually.

The constructed models are reviewed and evaluated before it is used for decision support. The models were evaluated using accuracy as the criteria to assess the performance of the classification method and constructive rules were extracted from it.

5. DATA SETS

Data set comprises recruitment details of two years in an IT industry. Data set of one year was used for training and the other was used for evaluating the constructed model. Attributes in Table.1 was formed from the collected information.

Table.1. Information Extracted from the Recruitment Data

Attribute	Type
First Name	Nominal
Last Name	Nominal
Gender	Nominal
Native State	Nominal
Student ID	Nominal
Year of Birth	Nominal
Place of Birth	Nominal
Name of High school	Nominal
Percentage of High school Marks	Numerical
Name of Higher secondary school	Nominal
Percentage of Higher Secondary Marks	Numerical
Stream of Diploma (If Applicable)	Nominal

Percentage of Diploma Marks (If applicable)	Numerical
Name of the college	Nominal
Stream of Under graduation	Nominal
Percentage of marks in each semesters	Numerical
Consolidate Percentage in Under graduation	Numerical
Skill set in core industrial competencies rated by themselves in a scale of 1 to 10	Nominal
Marks secured in technical test	Numerical
Marks secured in aptitude test	Numerical
Marks secured in programming test	Numerical

The information collected consists of general details such as name, gender, year of birth, native state, place of birth and ID given for the candidate. It also includes the educational details such as the name of the school, percentage of marks obtained in high schools, details regarding his diploma if applicable, name of the college, percentage of marks obtained in every semester of his graduation and total percentage in under graduation, etc. Marks secured by the candidate in the tests conducted by the industry are also included. The status of selection of the candidates, which is the target variable includes different values as shown in Table.2. The recommended status indicates that the candidate has been selected, on hold status indicate that the candidate will be selected if there is future requirement and the status rejected indicate that the candidate is not selected.

Table.2. Selection Status of Candidates

Status ID	Status Description
1	**Recommended**
2	**On Hold**
3	**Rejected**

Negative marks were given for wrong answers in the tests conducted by the industry; hence the dataset has negative marks too. The standard deviation of the marks of the candidates in both the set is tabulated in Table.3.

Table.3. Standard Deviation of Numeric Attributes

Attribute	σ of dataset1	σ of dataset2
Marks secured in technical test	5.257	5.888
Marks secured in programming test	3.639	3.984
Marks secured in aptitude test	2.615	3.357
Percentage of High school Marks	7.676	7.35
Percentage of Higher Secondary Marks	5.873	6.745
Percentage of Marks Secured in First Semester	5.164	9.137
Percentage of Marks Secured in Second Semester	6.721	10.012
Percentage of Marks Secured in Third Semester	6.37	10.368
Percentage of Marks Secured in Fourth Semester	6.227	10.426
Percentage of Marks Secured in Fifth Semester	5.747	10.716
Percentage of Marks Secured in Sixth Semester	5.728	10.473
Percentage of Marks Secured in Seventh Semester	5.491	10.838
Percentage of Marks Secured in Eighth Semester	6.145	10.341
Total Percentage of Marks Secured in Under Graduation	5.046	10.151

Dataset1 consists of 812 records and dataset2 consists of 2192 records. The final status of the candidates after the recruitment process is tabulated in Table.4.

Table.4. Final Status of Candidates

Final Status	Percentage of Records in Dataset1	Percentage of Records in Dataset2
1	4.3	2.91
2	1.01	0.96
3	94.68	96.13

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

Decision trees were built with the datasets using the data mining tool Weka. Trees were built with one dataset and tested with the other dataset to determine the most appropriate one. From Table.4, it may be observed, that the dataset consists of more than 95% of records in the rejected category; hence the constructed trees were very excellent in recognizing the rejected data, however they were not able to identify selected records to a large extent. Therefore the dataset was premeditated and decision trees were constructed with almost equal number of records in both the categories. When the records were chosen for the learning process, the distribution of the status in the original data was maintained. The constructed decision trees were used to classify all the records in the other dataset.

Accuracy of the classifier during testing is used as the metric for deciding the best suited model. First, the input attributes were used as it is in the given dataset (without discretization),

and thereafter by manually discretizing. Manual discretization was performed similar to the grading of marks. Each mark interval of 10 was given a suitable grade. Id3 algorithm cannot be applied for numeric attributes.

The accuracy of the classifiers is as shown in Table.5 and Table.6. The classifiers were able to identify records of both the categories.

Table.5. Accuracy of DTs When Tested with Dataset2

	Without Discretization	Manual Discretization
Id3	NA	45.12%
C4.5	75.23%	77.29%
C4.5 Unpruned	75.14%	76.73%
Cart	70.86%	72.12%
Cart Unpruned	69.91%	72.75%

Pruned and unpruned trees were constructed using the algorithms. It was observed that the accuracy of pruned trees was better than unpruned trees. The constructed unpruned trees were used to study the impact of the input attributes.

Table.6. Accuracy of DTs when Tested with Dataset1

	Without Discretization	Manual Discretization
Id3	NA	50.27%
C4.5	78.23%	79.12%
C4.5 Unpruned	76.04%	78.73%
Cart	74.86%	77.29%
Cart Unpruned	72.37%	76.57%

Fig.2. shows the accuracy of the decision tree classifier models constructed with the different algorithms using dataset1 and dataset2. It may be observed that the tree constructed with the C4.5 algorithm using dataset2, was best in classifying the test data.

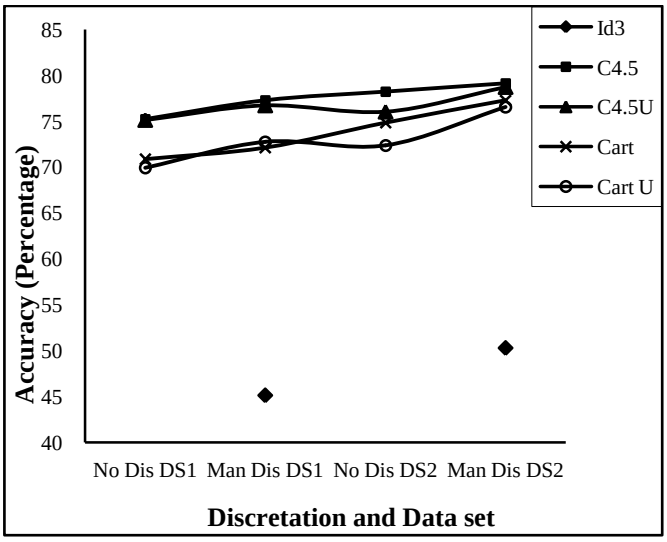

Fig.2. Comparison of the Decision Tree Construction Algorithms for Both the Datasets

Table.7. Sample of Rules Inferred from Decision Trees

If Marks_In_Programming > 5 AND Percentage_In_Higher_Sec_School > 90 AND Marks_In_Technical > 35 AND Percentage_In_BE > 70 then Selected
If College = X OR College = Y AND Percentage_In_BE > 70 AND Percentage_In_BE < 80 AND Percentage_In_Higher_Sec_School > 90 then Selected
If College = not(X) OR College = not(Y) AND Percentage_In_BE > 70 AND Percentage_In_BE < 80 then Rejected
If College = X OR College = Y AND Percentage_In_High_School > 90 AND Percentage_In_Higher_Sec_School > 80 AND Percentage_In_Higher_Sec_School < 90 AND Percentage_In_BE>70 AND Percentage_In_BE<80 then Selected
If College = not(X) OR College = not(Y) AND Percentage_In_High_School > 90 AND Percentage_In_Higher_Sec_School > 90 AND Percentage_In_BE>90 then Rejected

Analyses were made with the decision trees and 20 rules were deduced. The deduced rules were checked for viability with the domain experts and used in the recruitment process. Table.7 lists few such rules. Experiments were also carried out to find out the right set of input attributes for the classification problem. The accuracy of the classifiers increased considerably when only academic background and test marks of the candidates were given as input.

7. CONCLUSIONS

The recruitment mining problem has been identified and the domain has been well studied by interacting with the domain experts. The three popular decision tree construction algorithms, Id3, C4.5 and Cart have been applied for the problem and decision rules have been deduced. Analysis has been made by giving different set of inputs for the classifier by discretizing the continuous attribute and traditional methods. The set of input attributes are not same for all the industries and usually varies every year. Therefore, further analysis could be made to identify the best of attributes that will realize better selection criteria.

REFERENCES

[1] Chen-Fu Chien, Li-Fei Chen, 2008, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems with Application, Elsevier, Vol. 34, pp. 280-290.

[2] M. Saidi Mehrabad, M. Fathian Brojeny, 2007, "The development of an expert system for effective selection and appointment of the jobs applicants in human resource

management", Computers and Industrial Engineering, Elsevier, Vol. 53, pp. 306-312.

[3] Jiawei Han, Micheline Kamber, 2006, "Data Mining Concepts and Techniques", Second Edition Morgan Kaufmann Publishers, San Francisco.

[4] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant, 1996, "The Quest Data Mining System", in Proc. 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, pp. 244-249.

[5] Richard W. Selby, Adam A Porter, 1988, "Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis", IEEE Transactions on Software Engineering, Vol. 14, No. 12, pp. 1743-1757.

[6] Flippo, E. B, 1984, "Personnel management", McGraw-Hill Inc, U.S.

[7] Scarpello, V. G., Ledvinca, J, 1988, "Personnel human resources management", South Western publishing company, U.S..

[8] J.R. Quinlan, 1986, "Induction of Decision Trees", Kluwer Academic Publishers, Netherlands.

[9] Lewis, C.D, 1987, "Employee selection", Nelson Thrones Ltd., London, UK.

[10] Leo Brieman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, 1984, "Classification and Regression Trees", Chapman and Hall/CRC, New York.

[11] Usama M. Fayyad, Keki B. Irani, 1993, "Multi-interval discretization of continuousvalued attributes for classification learning", Thirteenth International Joint Conference on Articial Intelligence, France, pp. 1022-1027.

[12] Ron Kohavi, Ross Quinlan, 1999, "Decision Tree Discovery", in Proc. International Conference on Data Mining and Knowledge Discovery, 6, CA, USA.