Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Placement Prediction through Data Mining

Rakesh Kumar Arora
Dept. of Computer Science,
Krishna Engineering College,
Ghaziabad, UP, India

Dr. Dharmendra Badal
Dept. of Mathematical Science & Computer Applications,
Bundelkhand University,
Jhansi, UP, India

Abstract—Recruitment is one of the most important function for any organization as they seek talented and qualified professionals to fill up their positions. Majority of the companies have been focusing on campus recruitment to fill up their positions. This method is the best way to get the right resources at the right time with minimal cost and within minimum time frame. While the industry get the best talent from different institutes/universities, students too get chance to start their career with some of the best companies in the corporate world at the beginning of their career. The focus of this paper is to identify those set of students that are likely to face difficulty in getting the placements. The result of analysis will assist the academic planners to design the strategy to improve the performance of students that will help them in getting placed at the earliest.

Keywords— Data Mining, Business Intelligence, WEKA, Data Visualization, Decision Tree.

I. INTRODUCTION

A large number of self financing private institutes and universities have opened over the last decade with the objective of providing quality education to students in various fields of engineering and other professions. The factors affecting the quality of education include faculty profile, infrastructure, working environment and vision of the institute and most importantly placements of institute. If a student is spending heavily in paying fees to pursue the course, he also expects good placements from the institute. The good placements affect the quality of intake in the institute which further improves the academic results of the institute and hence subsequent placements. The major objective of campus placements is to identify the set of qualified and talented professionals before they complete their education. Campus placements are beneficial to both industry and to the institute. This process reduces the time for an industry to pick the candidate and train them according to their need while the students get exposed to the corporate environment at the right time and learn how to prepare themselves for the competition.

A large number of foreign universities have also got approval from the ministry of Human Resource and Development to compete with the Indian Universities. After the entry of these foreign universities, the survival of these self financed private educational institutes has become further challenging. Since the motive of most of the self financed institutions is to maximize the profit, hence they are not able to compete with the foreign institutes resulting in the closure of these institutes.[1]

In order to compete with the foreign universities and Government aided Indian Institutes these self financed private institutes and universities should increase their budget on hiring experienced and qualified faculty so as to provide excellent subject knowledge and optimum use of institute resources. This will have major impact upon the performance of the students, resulting in good academic results, placements and increased quality intake in the institutes. Like this these institutes will be able to sustain their existence competing with the good Indian and Foreign institutes.[1]

II. METHODOLOGY

Decision trees are a simple, but powerful form of multiple variable analysis. A decision tree is a special form of tree structure. The tree consists of internal nodes where a logical decision has to be made, and connecting branches that are chosen according to the result of this decision. The nodes and branches that are followed constitute a sequential path through a decision tree that reaches a leaf node (final decision) in the end.[2]

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. [3]

The decision tree algorithm is simple top down greedy algorithm. The major step of algorithm is to continue to divide leaves that are not homogeneous into leaves that are as homogeneous as possible until no further division is possible. The algorithmic steps for decision tree algorithm is as follows:[4]

1. Let the set of training data be S. If some of the attributes are continuous-valued, they should be discretized. Once that is done, put all of S in single tree node.
2. If all the instances in S are in same class, then stop.
3. Split the next node by selecting an attribute A from amongst the independent attributes that best divides or splits the objects in the node into subsets and create decision tree node.
4. Split the node according to the values of A
5. Stop if any of the following conditions are met, otherwise continue with step 3

Fig 1: Steps for Decision Tree Algorithm

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. There are two types of pruning:

1. Post pruning (performed after creation of tree)
2. Online pruning (performed during creation of tree) [5].

The steps to extract classification rules from tree are mentioned below:

1. Represent the knowledge in the form of IF-THEN rules.
2. One rule is created for each path from the root to a leaf.
3. Each attribute-value pair along a path forms a conjunction.
4. The leaf node holds the class prediction

The analysis using decision tree is being done with the help of WEKA tool. WEKA, formally called Waikato Environment for Knowledge Learning supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. [6]

III. ANALYSIS

The placement process in the institute is a yearly affair brimming with activities. It is a process involving the active participation of the placement cell and the final year students of all streams. The study was carried out on the 116 students passed from MCA department of reputed engineering college of Ghaziabad. The attributes considered for analysis of students along with their possible values are reflected in Table 1.

TABLE I: PARAMETERS USED FOR ANALYSIS

Parameters	**Values**
MCA_Result	A, B, C
Communcation_Skills	Good, Average, Poor
Programming_Skills	Good, Average, Poor
Participate_in_any_activity	Yes, No
Gender	Male, Female
12th_Result	A, B, C
Graduation_Result	A, B, C
Placement	Yes, No

The data file normally used by WEKA is in ARFF (Attribute-Relation File Format) file format, which consist of special tags to indicate different things in the data file. After collecting and cleaning the data, the classification of data is done using J48. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The output generated is displayed in Fig 2.

```
Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ABC
Instances:    116
Attributes:   8
              MCA_Result
              Communcation_Skills
              Programming_Skills
              Participate_in_any_activity
```



```
            Gender
            12th_Result
            Graduation_Result
            Placement
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------
MCA_Result = B
|   Gender = MALE: NO (43.0/18.0)
|   Gender = FEMALE: YES (15.0/5.0)
MCA_Result = C
|   Graduation_Result = B
|   |   Communcation_Skills = AVERAGE
|   |   |   Gender = MALE: NO (7.0/2.0)
|   |   |   Gender = FEMALE: YES (6.0/1.0)
|   |   Communcation_Skills = GOOD: YES (11.0/3.0)
|   |   Communcation_Skills = POOR: NO (1.0)
|   Graduation_Result = C: NO (12.0/3.0)
|   Graduation_Result = A
|   |   Programming_Skills = GOOD: YES (2.0)
|   |   Programming_Skills = POOR: NO (1.0)
|   |   Programming_Skills = AVERAGE: NO (4.0)
MCA_Result = A: YES (14.0/2.0)

Number of Leaves  :         11

Size of the tree :   17

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          82               70.6897 %
Incorrectly Classified Instances        34               29.3103 %
Kappa statistic                          0.4173
Mean absolute error                      0.3829
Root mean squared error                  0.4375
Relative absolute error                 76.6636 %
Root relative squared error             87.5586 %
Total Number of Instances              116

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.804     0.383      0.662      0.804      0.726      0.767    NO
                   0.617     0.196      0.771      0.617      0.685      0.767    YES
Weighted Avg.      0.707     0.287      0.718      0.707      0.705      0.767

=== Confusion Matrix ===
 a  b   <-- classified as
 45 11 |  a = NO
 23 37 |  b = YES
```

Fig 2: Output

The accuracy is around 71%. The kappa statistic measures the agreement of prediction with the true class where value 1.0 signifies complete agreement. The confusion matrix or contingency table in this example has two classes, and therefore a 2x2 confusion matrix is being displayed. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified.

The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. The Precision is the proportion of the examples which truly have class x among all those which were classified as class x. The F-Measure is simply 2 * Precision * Recall / (Precision + Recall), a combined measure for precision and recall.[10]

As per J48 Algorithm, parameters that reflect noise or outliers need to be removed, hence only those targeted node are shown by tree which have some value of precision and recall. The tree generated is represented in Fig 3.

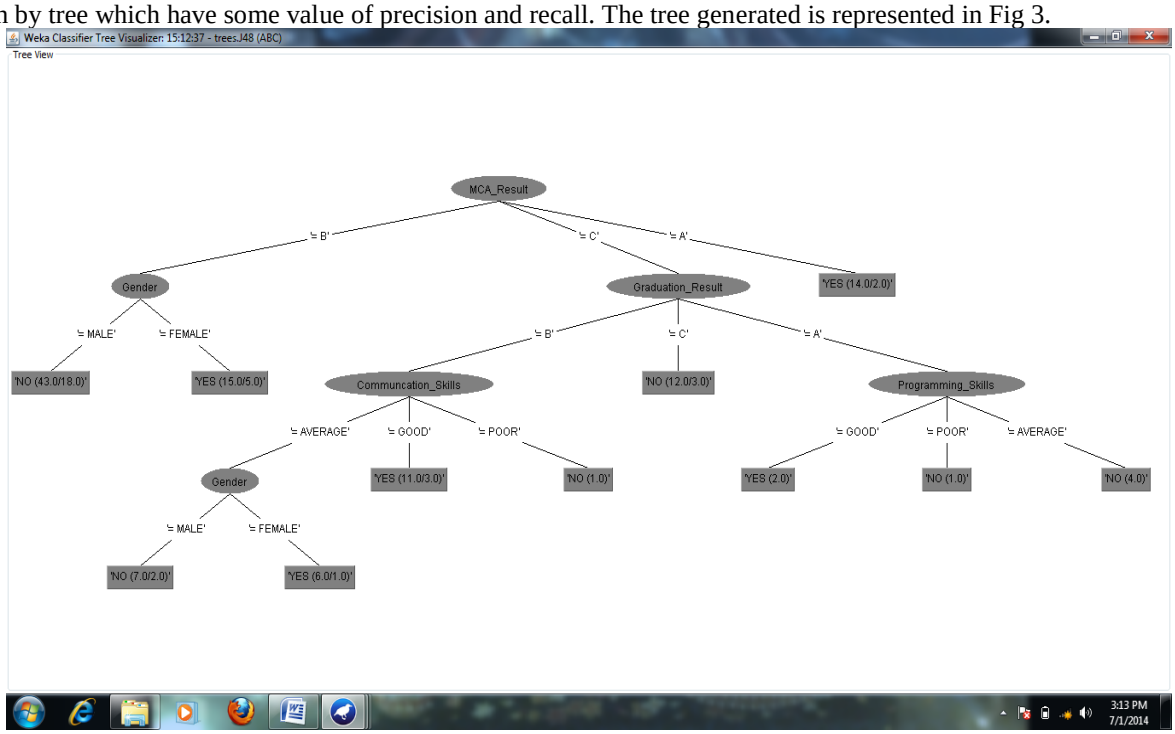

Fig 3: Decision Tree

The classification rules extracted from tree are:

1. If MCA_result = 'A' then student is placed.
2. If MCA_result = 'B' and Gender = 'Female' then student is placed.
3. If MCA_result = 'B' and Gender = 'Male' then student is not placed.
4. If MCA_result = 'C', Graduation_Result = 'A' and Programming_skills = 'Good' then student is placed.
5. If MCA_result = 'C', Graduation_Result = 'A' and Programming_skills = 'Average' then student is not placed.
6. If MCA_result = 'C', Graduation_Result = 'A' and Programming_skills = 'Poor' then student is not placed.
7. If MCA_result = 'C', Graduation_Result = 'B' and Communication_skills = 'Good' then student is placed.
8. If MCA_result = 'C', Graduation_Result = 'B' , Communication_skills = 'Poor' and Gender = 'Female' then student is placed.
9. If MCA_result = 'C', Graduation_Result = 'B' , Communication_skills = 'Poor' and Gender = 'Male' then student is not placed.
10. If MCA_result = 'C', Graduation_Result = 'B' and Communication_skills = 'Poor' then student is not placed.

The placement incharge/ Head of departments can easily identify the set of students from above classification rules that are likely to face problem during campus placement. In order to tackle the problem the Head of Departments can conduct special lectures/ classes for students to improve their academic result. In addition guest lectures from some resource person can be conducted to improve the programming skills of students. Personality Development classes can be conducted to improve the overall personality and communication skills of the students.

IV. CONCLUSION

In this paper, a simple methodology based on decision tree algorithm is being used to analyze the placement details of the students of MCA department of reputed engineering college of Ghaziabad. This methodology will assist the Placement Incharge and Head of Departments in identifying set of students that are likely to face problem during final placements. This identification will help the Head of Departments and Placement Incharge to design the strategies to improve the academic result, programming skills and communication skills of students. This model will play important role in improving the overall placements of the institute. This will result in significant improvement in quality of admissions and academic results in subsequent years. Like this these institutes will be able to sustain their existence competing with the good Indian and Foreign institutes.

REFERENCES

[1]. Arora K. Rakesh, Badal Dharmendra," Subject Distribution using Data Mining ", IJRET Vol. 2, Issue 12, December 2013
[2]. [Online]http://www.estard.com/decisiontree/decision_trees_definition.asp
[3]. [Online]http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf
[4]. Gupta K G. "Introduction to Data Mining with case studies" , PHI
[5]. Moertini, Veronica S. "Towards the use of C4.5 algorithm for classifying banking dataset." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013
[6]. [Online] Available: http://www.gtbit.org/downloads/ dwdmsem6/dwdmsem6lman.pdf
[7]. Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories "Data Mining Applications in Higher Education".
[8]. Juneja, Deepti, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb. 2013.
[9]. [Online]http://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching__Recall_Precision.pdf
[10]. [Online] http://weka.wikispaces.com/Primer
[11]. Arora K Rakesh, Badal Dharmendra, "Location wise student admission analysis", International Journal of Computer Science, Information Technology and Security, Dec 2012.
[12]. Arora K. Rakesh, Gupta K. Manoj ,"Data Mining: Scope Out Valuable Resources From Mountains Of Information", IITM Buisness Review Journal, July 10

