ISSN:0975-9646

Prediction and Comparative Analysis of Students Placements Using C4.5 &C5.0

K.Nasaramma[1], M.Bangaru Lakshmi[2], D.kiranmayi[3]

[1,2,3] *CSE Department, Vignan's Institute of Information Technology*

***Abstract*-Data mining is a knowledge discovery process that analyzes data and extracts useful information. Classification is the supervised learning method which uses decision tree to classify records. Using feature values of records can be classified. Each node in a decision tree represents a feature in an instance to be classified. In this work C 4.5 and C5.0 are compared in terms of accuracy. Among all these classifiers C5.0 gives more accurate results. C5.0 trees are smaller and therefore require less memory than c4.5. Error rate is low so accuracy in result set is high and pruned tree is generated so the system generates fast results as compare with other technique. Our system use C5.0 classifier that Performs feature selection and reduced error pruning techniques.**

***Index Terms*- Data mining, classification, decision tree.**

I. INTRODUCTION

Campus Placements in educational institutions is a significant metric for evaluation of an institution. Nowadays institutions providing technical education are taking care of their student placements. In this regard, prediction of student's placement system based on their performance in curricular and co curricular activities helps to improve their placements by analyzing the students are lagging in particular areas. So, we built a model by using C5.0 data mining algorithm based on
decision tree. The accuracy of C5.0 is compared to C4.5 and the results are discussed.

A. Introduction to Data Mining:

Data Mining is the process of analyzing data to find useful information hidden in data. The mining techniques include prediction, classification, Association analysis and clustering. The future trends and behaviors can be predicted with data mining tools which allow businesses to make decisions. These techniques are applied in various fields like business analysis, financial analysis, weather prediction, fraud detection, social media mining etc. The extracted knowledge can be presented in the form of graphical or statistical forms.

Data mining process consists of five major elements:

- Extract, transform, and load data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

B. Data Mining in Educational institutions:

Campus placements of the student play an important role in an educational institution. Majority of the companies have been focusing on campus recruitment to fill up their positions. The companies identify the talented and qualified professionals before they have completed their education. This method is the best way to work on the right resources at the right time to get good companies at the beginning of their career. This can very well be achieved using the concepts of data mining. For this purpose, we have analyzed the data of students of computer engineering.

II. METHODOLOGY

This predictive model consists of the following phases:

A. Data collection:

The data sets of technical education institutions are collected based on previous placement records this model predicts the data sets of upcoming student batches with relative performances.

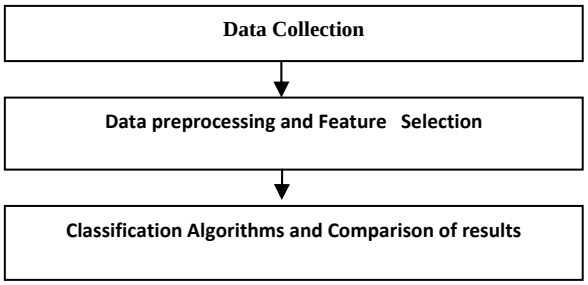
B. *Data preprocessing and Feature Selection*:

In this phase, redundant and missing values are identified and corrected. The features which are significant in prediction is selected by using relevance measures like covariance. For Example: The student's technical knowledge feature is more important than the marks obtained in high school.

C. *Classification Algorithms and Comparison of Results:*

The main attributes considered in building this predictive model are marks scored in various tests of students.
The student's knowledge on programming skills, Numerical ability, reasoning, presentation skills are considered to build a decision tree which leads to classification. The classification of students is identified using c4.5 and c5.0 algorithms. The accuracy of results of both the algorithms on training and test data sets are compared.

Fig1.System Architecture



III. IMPLEMENTATION

A. Data set collection module:

The student's data sets are collected from technical education institutions providing computer science courses. The data comprises of student scores on the attributes like student aggregate marks, programming skills, Numerical ability, reasoning, and presentation skills.

B. Data preprocessing Module:

Data preprocessing is performed on the data sets to remove redundant records, filling missing values by manually checking the attribute values. The exceptional values are identified and corrected to avoid wrong predictions.

C. Classification Module:

The classification of tuples is carried out based on decision tree algorithms C4.5 and C 5.0. The data sets are divided into training set and test data set.Training data set is used to build a learning model and test set is used to validate the model[1]. The following figure depicts the decision tree based on C4.5 algorithm.The attribute selection of C4.5 is based on data which splits sets efficiently based on splitting criterion Gain Ratio[2]. The attribute which is

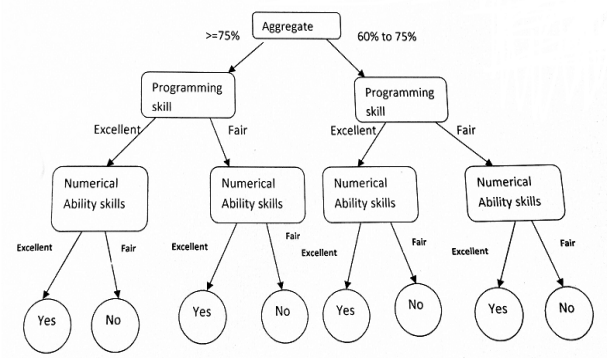


Fig2: Decision tree for placement Data

having higher gain ratio is chosen as a best split in each step, during construction of a decision tree.

Gain Ratio(A)=Gain(A)/Split Info(A)

C5.0 initially build a rule set for each possible value of input attributes. By using boosting technique it improves accuracy by modifying the initial tree[3]. The number of iterations to improve the tree can be chosen by the user. By increasing number of iterations learning takes longer time.

IV. RESULTS AND ANALYSIS

The collection of data set is divided into two sets, in which training set consists of two-third of records and one-third as test set. The features considered for building a predictive model are marks scored in qualifying exam, performance in programming skills and computing skills. The classification is performed by generating decision tree and the accuracy of the classifier is measured as number of tuples classified correctly. Accuracy is given by using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents True Positives, TN represents True Negatives(TN), FP represents False Positives and FN represents False Negatives.

Accuracy of the C4.5 Analysis: The accuracy of C4.5 classifier when test data is given as input is 88.8%

Analysis of C4.5	
True Positives	323
True Negatives	32
False Positives	27
False Negatives	18
Accuracy	88.8%

Accuracy of the C5.0 Analysis: The accuracy of C5.0 classifier when test data is given as input is 95%.

Analysis of C5.0	
True Positives	327
True Negatives	53
False Positives	13
False Negatives	7
Accuracy	95%

CONCLUSION

In every institution placement analysis is very crucial, which is beneficial for students seeking placement after completion of their course work. In this regard our work shows C5.0 classifies the student's records accurately, which helps placement co-coordinator to identify the students who are not up to the mark in specific categories like programming knowledge or computing skills. The accuracy obtained for C4.5 is 88.8% and C5.0 is 95%, which shows that C5.0 is better if used for placement analysis compared to C4.5 algorithm.

REFERENCES

1. Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong
2. A comparative study of decision tree ID3 and C4.5 Badr HSSINA,Abdelkarim MERBOUHA, Hanane ZZIKOURI, Mohammed ERRITALI TIADlaboratory,ComputerSciencesDepartment, Faculty of sciences and techniques Sultan Moulay Slimane University Beni-Mellal, BP: 523, MoroccoM. Govindarajan, Text Mining
3. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, International Journal of Computer Applications (0975 – 8887) Rutvija Pandya Diploma Computer Engineering Department, Gujarat.
4. Results and Placement Analysis and Prediction using Data Mining and Dashboard, International Journal of Computer Applications (0975 – 8887) Volume 137 – No.13, March 2016 Siddhi Parekh Ankit Parekh , Ameya Nadkarni,Riya Mehta B.E. Students, Dept. of Information Technology St. Francis Institute of Technology Mumbai, India
5. Performance Analysis of Data Mining Techniques for Placement Chance Prediction V.Ramesh, P.Parkavi, P.Yasodha, International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011 1 ISSN 2229-5518
6. Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms International Journal of Computer Applications (0975 – 8887) Volume 108 – No 15, December 2014 .