Decision Tree – A Numerical Approach

Nishtha Srivastava
Dept of Computer Department
Vadodara Institute Of Engineering

Abstract- *Educational data mining is the method of applying data mining tools and techniques to analyze data at educational organizations. In this paper, we have used educational data mining to predict students' campus placement based on their grades in previous subjects like B.Tech, M.Tech etc.This may be lead to efficient use of student education data base for placement and non-placement classification. The model which we have used to predict is ID3 decision tree algorithm.*

Keywords- Data Mining, Classification, Decision Tree, ID3, Entropy, Information gain.

I. INTRODUCTION

Campus placement plays a very vital role in student's life. This paper suggests a system which helps in deciding the likelihood of a student getting placed. We have used the concept of Classification. Classification is notion of data mining.Decision tree use a tree-like graph of decisions and their possible outcomes, including chance event outcomes, resource costs and utility.

II. BACKGROUND STUDY AND RELATED WORK

A. ID3 Algorithm

In decision tree learning, ID3 (Iterative Dichotomiser)is an algorithm invented by Ross Quinlan used to create a decision tree from the dataset.[3]To model the classification process, a tree is built using the decision tree approach.Once a tree is created, it is tested to each tuple in the database and this results in classification for that tuple.

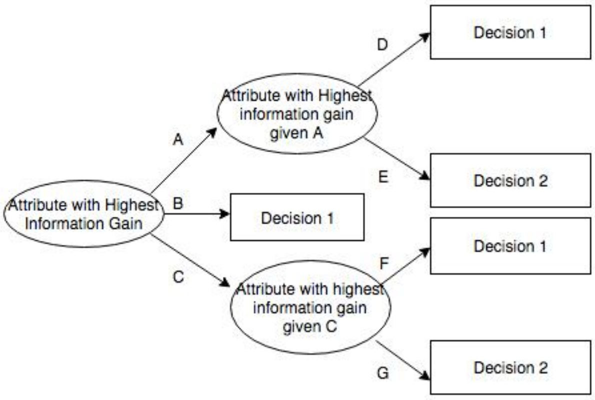

Attribute with Highest Information Gain
A
B
C
Attribute with Highest information gain given A
Decision 1
Attribute with highest information gain given C
D
E
F
G
Decision 1
Decision 2
Decision 1
Decision 2

Fig 1 Decision Tree works on the concept of information gain

The following issues are faced by most decision tree algorithms[4]:

- To choose splitting attributes
- Number of splits to be taken based on entropy.
- The stopping criteria

The basis for decision tree is Entropy. Our main motive is to map all examples to different categories based upon different situations of the condition attribute set. The algorithm prefers information gain as attribute selection criteria. Usually the attribute that has got the maximum information gain is picked as the splitting attribute of the current node. Branches can be set up based on various values of the attributes and the process above is recursively performed on each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain and Gain Ratio are used.

B. Entropy

Entropy is the impurity associated with an arbitrary collection of illustrations. It's the lowest number of bits of information needed to encode the classification of an arbitrary member of collection S. It is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples.
Entropy is given as,

Entropy(s) =$\sum$- Pi log$_2$Pi

Pi = probability of S belonging to class i.

Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. For e.g. if training data has 14 instances with 5 positive and 9 negative instances, the entropy is calculated as

Entropy ([5+,9-]) = -(5/14)log(5/14)- (9/14)log(9/14)
=0.9402

C. Information Gain

The decision tree is built in a top-down fashion. ID3 chooses the splitting attribute with the highest gain in information, where gain is the difference between how much information is needed after the split. This is calculated by calculating the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. Suppose the Information has Gain (G, A) of an attribute A, it is to a collection of examples in G, which is defined under as follows
:

Gain(G,A) = Entropy (G) - Σ |Gv|÷|G| Entropy (Gv)

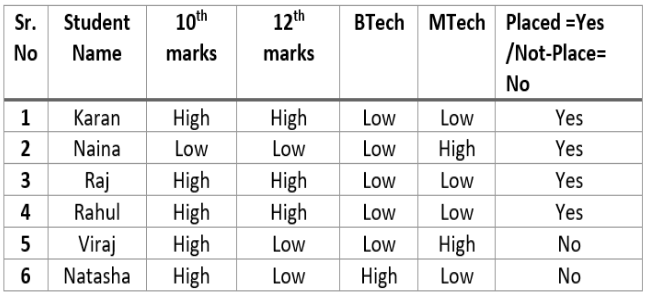
Σ = Value (A) is the set of all possible values for attribute A.
Gv = Gv is the subset of G for which Attribute A has v , i.e., Gv = {g = Gv | A(g) = v }).

The first calculation or term in information Gain is the entropy of the original collection G and the secondterm is the desired value of the entropy after G is partitioned using attribute A. The expected entropy described by second term is the sum of the entropies of each subset, the fraction of examples |Gv|÷|G| that belong to Gain (G, A) is therefore the optimal reduction in entropy happen by knowing the value of attribute A.

The attribute with highest value of information gain is used as the splitting node thereby constructing the tree in top down manner.

III. PROPOSED SYSTEM

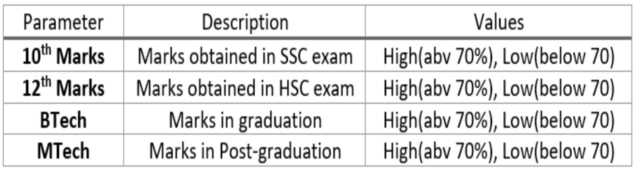
A decision tree is then implemented to determine the possible outcome if a student is placed or not. Below is a set of student data ,considered as the base set for the proposed system. The data comprises of 6 students. The attributes such as SSC marks, HSC marks,B. Tech grade and MTech grades have been taken into consideration. Based on the training set as in Fig 2, information gain and entropy is calculated to determine the splitting attribute for constructing the decision tree.



Sr. No	Student Name	10th marks	12th marks	BTech	MTech	Placed =Yes /Not-Place= No
1	Karan	High	High	Low	Low	Yes
2	Naina	Low	Low	Low	High	Yes
3	Raj	High	High	Low	Low	Yes
4	Rahul	High	High	Low	Low	Yes
5	Viraj	High	Low	Low	High	No
6	Natasha	High	Low	High	Low	No

Fig 2: Student Data Set

The combination of various attributes determines whether the student is placed or not. Attributes and their values are shown in Fig 3.



Parameter	Description	Values
10th Marks	Marks obtained in SSC exam	High(abv 70%), Low(below 70)
12th Marks	Marks obtained in HSC exam	High(abv 70%), Low(below 70)
BTech	Marks in graduation	High(abv 70%), Low(below 70)
MTech	Marks in Post-graduation	High(abv 70%), Low(below 70)

Fig 3 Attributes and their values

According to our student data set as shown in Fig 2, we have set of 6 examples with 4 “yes” and 2 “no”.

So, entropy is calculated as:

Entropy(S) =$\sum$- Pi log$_2$Pi

Entropy (4+,2-) = -P$_+$log$_2$P$_+$ - P$_-$log$_2$P

Entropy (4+,2-) = -(4/6) (log (4/6) ÷log 2)- ((2/6) log (2/6) ÷log 2)) = 0.9182

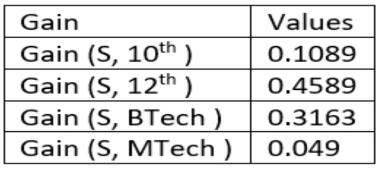
In B.Tech attribute there are two possible values such as Low and High.

B.tech = Low is of occurrence 5
B.tech = High is of occurrence 1
B.tech = In this Low, out of 1 there are 1 ‘Yes’ and 0 ‘No’
= [1+,0-]
B.tech = In this High, out of 5 there are 3 ‘Yes’ and 2 ‘No’
= [3+,2-]

Entropy (Low) = - (1/1) log2 (1/1) – (0) log2 (0)
= 0
Entropy (High) = - (3/5) log2 (3/5) – (2/5) log2 (2/5)
= 0.97095

Gain (S, 10th) = Entropy (S) – (5/6) × Entropy (Low) – (1/6) × Entropy (High)
= 0. 9182– (5/6) × 0 – (1/6) × 0.97095– (3/12) × 0
= 0.1089

Similarly , we can calculate Gain (S, 12th), Gain (S, BTech) and Gain (S, MTech). Please refer Fig 4 for the same.



Gain	Values
Gain (S, 10th)	0.1089
Gain (S, 12th)	0.4589
Gain (S, BTech)	0.3163
Gain (S, MTech)	0.049

Fig 4 Gain and Values

Attribute having the maximum Information gain becomes the root node. Sohere, our root node will be 12thMarks. This process repeatedly going on until all data do not classify perfectly or we run out of attributes.

In the final stage, wemake decision tree as shown in Fig 5.

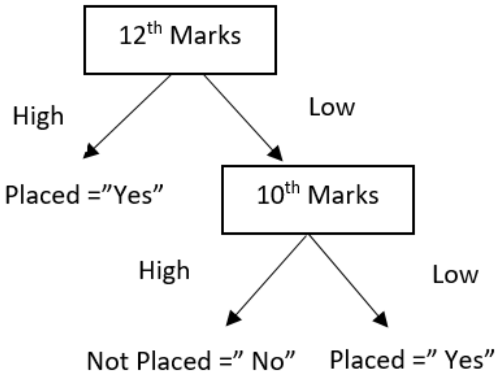


Fig 5 Decision Tree

The subsequent nodes of decision tree at each level are determined by the value obtained in information gain.

IV. CONCLUSION

In this paper ID3 classification algorithm is used to generate decision rule. The generated decision rule can be used to predict a student’s campus placement. The result of this algorithm can be used by the placement-in-charge to identify those set of students that are likely to face problems in campus placement. The classification model can play an important role in increasing the placement statistics.

REFERENCES

[1] Recommender Systems by Perm Melville and Vikas Sindhwani *IBM T.J. Watson Research Center, Yorktown Heights, NY 10598*

[2] Er.Paramjit kaur 1 , Er. Kanwalpreet Singh Attwal ," Comparative

[3] Analysis of Decision Tree Algorithms for the Student's Placement Prediction", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015 . D. D. B. Rakesh Kumar Arora,

[4] “Placement Prediction through Data Mining,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 7, july 2014. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao,

[5] " Predicting Students' Performance using ID3 and C4.5 classification algorithm", International 000, Dhawan, Rohit

About The Author

Professor Nishtha Srivastava is currently working as an Assistant Professor at Vadodara Institute of Engineering, Vadodara. She has completed her MTech from Nirma Institute of Engineering in CSE(Computer Science and Engineering) in 2017.She also has 2 years working industrial experience in Oracle , Bangalore and Rishabh Softwares , Vadodara. You can contact her at: nishtha8080@gmail.come/15mcec28@nirmauni.ac.in.