International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

Deployment of ID3 decision tree algorithm for placement prediction

Kirandeep, Prof. Neena Madan
M.Tech (CSE), G.N.D.U, Regional Campus, Jalandhar, Punjab, India

ABSTRACT

This paper details the ID3 classification algorithm. Very simply, ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. The main aim of this paper is to identify relevant attributes based on quantitative and qualitative aspects of a student's profile such as CGPA, academic performance, technical and communication skills and design a model which can predict the placement of a student. For this purpose ID3 classification technique based on decision tree has been used.

I. INTRODUCTION

Classification is the process to map data into predefined groups or classes. Also called supervised learning because classes are determined before examining data. It can also be defined as

D= {t1,t2,……..,tn}
C= {C1,C2,……...,Cm}

where data is defined by D having set of tuples that is assigned to class C.

e.g. Pattern recognition, an input pattern is classified into one of several classes based on similarity.
A bank officer who has the authority to approve the loan of any person then he has to analyze customer behavior to decide passing the loan is risky or safe that is called classification.

II. EASE OF USE

Predicting tumor cells as benign or malignant

Helpful in the field of medical science for predicting whether the tumor cells are malignant or not.

Classifying credit card transactions as legitimate or fraudulent
To check whether the transactions are legal or not.

Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
For classification of proteins on the basis of their properties

Categorizing news stories as finance, weather, entertainment, sports etc.
For categorization of news on the basis of respective classes

CLASSIFICATION ALGOS

Statistical based algorithms: This can be categorized into two categories:

1. Regression
2. Bayesian classification

Regression:
Deals with estimation of output values from input values. Can be used to solve classification problems. Also used for forecasting ,Radio

y=c0+c1x1+……+cnxn

where y is output & c0,c1,….,cn are the coefficients that defines relation between input and output i.e. x&y.

e.g. A simple linear regression problem can be thought as

y=mx+b

This can be equated as partitioning of two classes.If we attempt to fit data that is not linear to the linear model , the result will be poor model of data.

Bayesian classification:

By analyzing each independent attribute, a conditional probability is determined. Consider a data value xi,the probability that a related tuple ti, is in class Cj can be given as

P(Cj|xi)

i.e. P(xi),P(Cj),P(xi|Cj) from these values,Bayes theorem allows to estimate the probability

P (Cj|xi) & P(Cj|ti)

According to the theorem,

1. Estimate P(j) for each class by counting how often each class occurs.
2. The no. of occurrences of xi i.e. each attribute value**.** Similarly P(xi|Cj) can be estimated.
3. Suppose that ti has p independent attribute values

{xi1,xi2,……,xip}

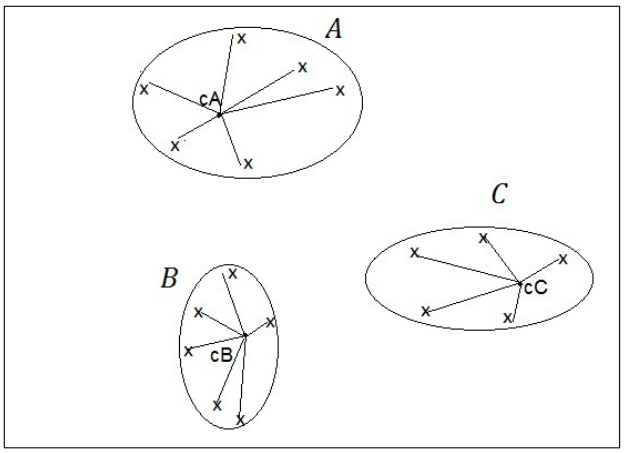
Similarly, P(Cj|ti) can be estimated.

B. **Distance based algorithms:**

Assignment of the tuple to the class to which it is most similar.

Algo:

Input: c1,c2,…….,cm(Centers for each c)
//input tuple
Output: C
//class to which t is assigned
dist=inf ;
for i=1 to m do
if dist(ci,t) < dist ; then
c=i;
dist=dist(ci,t);



C. ***Decision tree based algorithms****: A* 2-Step process includes

1) Construction of tree where each internal node is labeled with an attribute.
2) Leaf node is labeled with class.

THE ID3 ALGORITHM

A technique to build a decision tree based on information theory and attempts to minimize the no. of comparisons.

The ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy (information gain) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set is then split by the selected attribute (e.g. age is less than 50, age is between 50 and 100, age is greater than 100) to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

1. Calculate the entropy of every attribute using the data set.
2. Split the set into subsets using the attribute for which entropy is minimum (equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute.
4. Recurse on subsets using remaining attributes.

ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C:

Step 1: If all instances in C are positive, then create YES node and halt.

If all instances in C are negative, create a NO node and halt.
Otherwise select a feature, F with values v1, ..., vn and create a decision node.
Step 2: Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.
Step 3: Apply the algorithm recursively to each of the sets Ci.

ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute.The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

Data Description

The sample data used by ID3 has certain requirements, which are:

- *Attribute-value description* - the same attributes must describe each example and have a fixed number of values.
- *Predefined classes* - an example's attributes must already be defined, that is, they are not learned by ID3.
- *Discrete classes* - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect.

b) Attribute Selection

How does ID3 decide which attribute is the best? A statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to define gain, we first borrow an idea from information theory called entropy.

Entropy: A formula to calculate the homogeneity of a sample then the entropy S relative to this c-wise classification is defined as

Entropy(e1,e2,….en)=-p1logp1-p2logp2….-pnlogpn

Entropy(S) =∑- p(x) log p(x)

Where Pi is the 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 7 instances with 3 positive and 4 negative instances, the entropy is calculated as

Entropy ([3+,4-]) = -(3/7)log(3/7)-(4/7)log(4/7)=0.016

Thus, the more uniform the probability distribution, the greater is its entropy. If the entropy of the training set is close to one, it has more distributed data and hence, considered as a good training set.

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 defined as difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The motive is to find the feature that best splits the target class into the purest possible children nodes - pure nodes with only one class This measure of purity is called information. It represents the expected amount of information that would be needed to specify how a new instance of an attribute should be classified. The formula used for this purpose is:

G(D, S) = H(D) - ∑P(Di)H(Di)

Reasons to choose ID3

1. Understandable prediction rules are created from the training data.
2. Builds the fastest tree & short tree.
3. Only need to test enough attributes until all data is classified.

III. IMPLEMENTATION

Campus placement is a process where companies come to colleges and identify students who are talented and qualified, before they finish their graduation. The proposed system determines the likelihood of placement based on various attributes of a student's profile. Depending on the parameters , manual classification is done whether the student is placed or not placed. The data set comprises of different quantitative and qualitative measures of 7 students. The attributes such as department of the student, CGPA (Cumulative Grade Performance Average), programming skills, future studies such as planning for a master's degree, communication skills and the total number of relevant internships have been taken into consideration. Based on the training set, information gain and entropy is calculated to

determine the splitting attribute for constructing the decision tree.

Dept.	CGPA	Prog. Skills	Future Stds.	Comm. Skills	Intern ship	Placement
CS	8.0	Good	No	Good	1	Placed
IT	6.7	Avg.	Yes	Avg.	0	Not Placed
CS	7.6	Good	Yes	Good	1	Placed
CS	8.3	Good	No	Avg.	1	Placed
IT	6.5	Poor	Yes	Good	0	Not Placed
CS	8.6	Poor	Yes	Good	1	Not Placed
IT	7.9	Good	Yes	Avg.	1	Placed

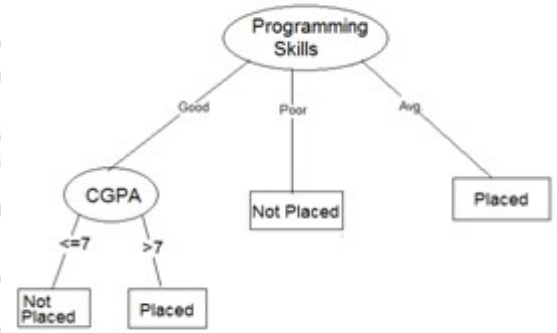
Fig. 1 Student data

The combination of various attributes determines whether the student is placed or not. The quantitative aspects like undergraduate CGPA. The qualitative aspects like communication and programming skills form a backbone for a student to get placed as each recruiting company desires to hire students that have a sound technical knowledge and ability to communicate effectively. The other factors like internships, backlogs, future studies add value only when the prior requirements are met.

The attributes and the possible values are explained below

TABLE I: Attributes and their values

Parameter	Description	Possible Values
Department	Deptt. Of Student	{CS,IT}
CGPA	CGPA(Out of 10)	Numeric{<=10}
Pgmg. Skills	Proficiency in C,C++ & Java	{Good, Avg,Poor}
Future Studies	Whether the student is planning	{Yes,No}
Comm. Skills	Proficiency in Comm. Skills	{Good,Avg,Poor}
Internship	Internships	{Yes,No}
Placement	Whether student is Placed or not	{Yes,No}



The root node chosen here is Programming Skills.And Further classification is done by calculating information gain and entropy for each attribute.

Attributes	Entropy	Gain
Department	CS=0.24 IT=0.26	G(Dept.,S)=0.25
CGPA	<=7=0 >7=0.05	G(CGPA,S)=0.02
Prog. Skills	Good=0.28 Avg=0.25 Poor=0.25	G(Prog. Skills,S)=0.51
Future Studies	Yes=0 No=0.2	G(Future Studies,S)=0.143
Comm. Skills	Good=0.28 Avg=0.28 Poor=0	G(Comm. Skills)=0.28
Internships	Yes=0.05 No=0.25	G(Internship,S)=0.194

Consider the attribute future studies; it has two possible classes viz. Yes and No. There are five students who wish to pursue future studies and remaining two out of seven who do not have any plans to opt for higher studies.

Higher value of entropy indicates higher degree of distribution of information among classes.

The lowest value of information gain is obtained for programming skills. Thus it is chosen as the root node.

Further, the next lowest value (CGPA) is taken as the split node for next level. The subsequent nodes of decision tree at each level are determined by the value obtained in information gain.

Advantages of Decision Tree:

1. For data preparation, decision trees need less effort from users. To overcome scale differences between parameters - for example if there is a dataset which measures revenue in millions and loan age in years, say; this will require some form of normalization before it can fit a regression model and interpret the coefficients. Such variable transformations are not required with decision trees because the tree structure will remain the same with or without the transformation.
2. However, decision trees do not require any assumptions of linearity in the data. Thus, they can be used in scenarios where known parameters are nonlinearly related.
3. The best feature of using trees for analytics is that they are easy to interpret and explain. Decision trees are very intuitive and easy to explain.

IV. RESULTS

The splitting node is based upon Information gain, i.e. Programming skills in this case. Table III indicates the department and the CGPA of the students who have good programming skills. Students having good programming skills are only considered as EGood=0.28, whereas EAverage=0.25, EPoor=0.25

Department	CGPA	Prog. Skills
CS	8.0	Good
CS	7.6	Good
IT	7.9	Good
CS	8.3	Good

The next splitting attribute based upon Information Gain is CGPA. The students having good programming skills and CGPA >7 are considered. As, E>=7=0.05 and E<7=0.

Department	CGPA	Prog. Skills
CS	8.0	Good
IT	7.9	Good
CS	8.3	Good

CONCLUSION

In this paper ID3 classification algorithm is used to generate decision rule. The classification model can play an important role in increasing the placement statistics. It can be concluded that classification algorithms can be used successfully in order to predict student placement. I will use it with MATLAB implementation tool and then results will be compared with another algorithm.

Further the implementation will be done in development and application of novel computational techniques for the analysis of large datasets.

ACKNOWLEDGEMENT

I express my sincere gratitude towards our guide Ms. Neena Madaan who assisted me throughout my work. I thank her for directing me to the right tract and for the motivation and prompt guidance she has provided whenever I needed it.

REFERENCES

1. https://en.wikipedia.org/wiki/ID3_algorithm
2. Key advantages of using decision trees for predictive analyticsSimafore [Online]
3. http://www.simafore.com/blog/bid/62333/4-key-advantages-ofusing-decision-trees-for-predictive-analytics
4. Er. Kanwalpreet Singh Attwal ," Comparative Analysis of Decision Tree Algorithms for the Student's Placement Prediction", International:
5. Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015
6. D. D. B. Rakesh Kumar Arora, "Placement Prediction through Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 7, july 2014.
7. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, " Predicting Students' Performance using ID3 and C4.5 classification algorithm", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2015
8. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.