Survey on ID3 Algorithm in Data Mining

Dipu Akter Shila

*B.Sc.Engg in CSE*

*Bangladesh University of Business & Technology*

Dhaka, Bangladesh [dipu.cse.bubt97@gmail.com](mailto:ashiqashiqur@gmail.com)

Md.Tariqul Islam

*B.Sc.Engg in CSE*

*Bangladesh University of Business & Technology*

Dhaka, Bangladesh

ts.masum849@gmail.com

Md. Hasanul Banna

*B.Sc.Engg in CSE*

*Bangladesh University of Business &*

*Technology*

Dhaka, Bangladesh

***Abstract*— In the classification task of decision tree learning, ID3 (Iterative Dichotomiser 3) is a famous algorithm that is used. The ID3 algorithm is a top-down approach algorithm that Quinlan created in 1983 to create a decision tree. In the present era, the ID3 algorithm plays an important role in every field. It is used to identify computer crime forensics, Information Asset Identification, knowledge acquisition for tolerance design, etc. We describe the basic cases and current research contents of the ID3 algorithm in this paper. First of all, we discuss data mining and classification and next represent the theoretical concept of the ID3 algorithm.**

***keywords*— Classification, Decision tree, Id3, Entropy, Information gain**

1. INTRODUCTION

Data mining is the elimination of hidden predictive information and unknown data which are difficult to detect statistically by analyzing the enormous data sets. Knowledge and information underlying this information do not know it in advance, but it is useful. Data mining is a quantitative approach that allows users to report and control errors in a friendly and even more standard way. It is a very important step that collectively tests a lot of Regular data. One of the important tasks of data mining is data pre-processing. In principle, Data mining applies to any kind of data collection. Concerning the produced model, Prediction and classification techniques are used in which classification technique prophesies the unknown values. To search correlations and routines in data, extract knowledge in the reports of rules and define the value of the dependent variables, a repository of data mining techniques can be executed. The following is a list of the data mining techniques used in the field as Naive Bayes, Decision tree, Artificial neural network (ANN), K- nearest neighborhood (KNN), Support vector machine (SVM), ID3 algorithm, etc. Data mining is also called knowledge discovery in databases (KDD) that implies an iterative method of data cleaning, data integration, data selection, pattern recognition, and data mining knowledge recognition. Association, classification, clustering, statistical analysis, and prediction are comprised of data mining.

1. *Classification*

Classification is a supervised learning method that has expressed the companies in an already established class of recruiting new ones and Considering the attributes of the entity. The main objective of the classification is to precisely predict the destination class for each input data. The classification algorithm uses a test data set and training data set to create a Nobel system that provides us to more accurately perform. The ID3 algorithm is also a data classification technique for indexing data that provide a discrete output.

1. *Decision Tree*

The most popular model is the decision trees approach in the classification. It predicts data from the training data set. With a class label, every leaf node is linked. Here is an illustration of a decision tree where the root node is age and child nodes are old and young.

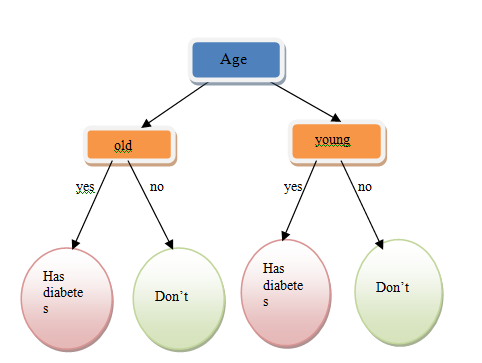


Fig. 1. Determining a decision tree whether a person has diabetes or not.

1. ID3 ALGORITHM

The ID3 algorithm is a top-down approach algorithm that Quinlan created in 1983 to create a decision tree. It is a famous algorithm which follows Occam's razor policy. It does not support the backtracking method.

During building a decision tree, the ID3 decision makes use of two concepts [1]:

1. Entropy

2. Information Gain

Allow us to know about information entropy and information gain.

1. *Entropy*

Entropy is the possibility measure of uncertainty. Entropy's responsibility is to control or how a decision tree breaks the data.When entropy is zero, then the dataset is perfectly classified.

The equation of information entropy is-

(1)

1. *Information Gain*

By decreasing the weighted entropies of each branch, information Gain is estimated for a split.Information gain is also called Kullback-Leibler divergence

The equation of information gain is:

(2)

Here,

Training set,A=}.

|A| =training sample number.

s = ( 1,2,3,…..,(|j|) ) and = (1,2,….,|A|) .

H = .

= denotes sample subsets.

(*j* =1*,* 2 *· · · |A|*) refers to the probability of each class for every current sample set.

Another method to measure information entropy and information gain:

(3)

And,

(4)

(5)

u=no of positive attributes

v=no of negative attributes

1. PROCEDURE

1. Take the sample training data set.

2. Using (3) and (4), calculate the information entropy of each attribute

3. Measure the information gain to use (5) rule

4. Including the highest Information Gain, determine the attribute to divide by accepting the attribute

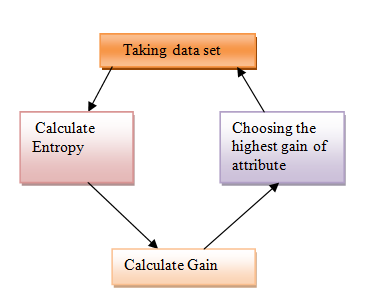


Fig. 2. Basic steps of ID3 Algorithm.

EXPERIMENT AND ANALYSIS

*Sample Training Data Set*

For example, from the record below have to judge if a person has diabetes or not:

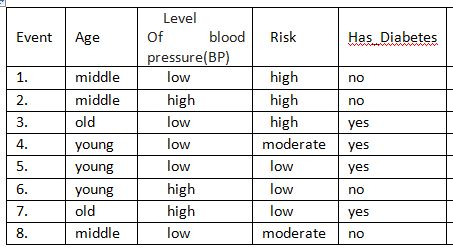


Fig. 3. Sample Training Data Set

*B. Analysis of this example*

We need to calculate information entropy and next measure information gain. To use the latest rule, we calculate these.

Information Entropy Calculate.

Here, give us some training sample and Has\_Diabetes consists of 8 columns with two attributes: yes and or no.

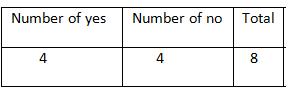


Fig. 4. Calculate no of yes & no

Information Entropy of Has\_Diabetes,

IE(Has\_Diabetes) =

= 1

Age attributes on has\_diabetes

Age consists of 8 columns with three attributes: middle, old and young. Let's start with ‘Age’ attribute, calculating its IE(Has\_Diabetes | Age) and IG(Has\_Diabetes | Age):

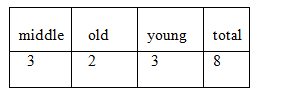


Fig. 5. Calculate number of age attributes

We can see that there are three attributes of age, so first need to calculate  (Has\_Diabetes | Age=middle), (Has\_Diabetes | Age=old), (Has\_Diabetes | Age=young) accordingly.

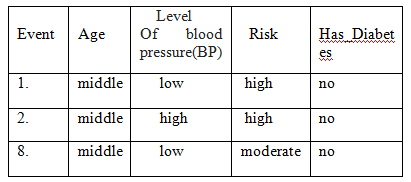
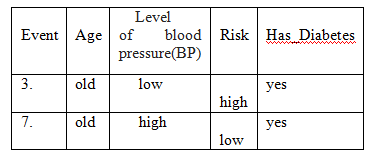


Fig. 6. Middle age attributes on has\_diabetes

There are 3 occurrences of middle age. Has\_Diabetes of 3 items are no and 0 items are yes as shown below-

IE(Has\_Diabetes | Age=middle) =

= 0

 Fig. 7. Old age attributes on has\_diabetes

There are 2 occurrences of old age. Has\_Diabetes of 2 items are yes as shown below-

IE(Has\_Diabetes | Age=old) =

= 0

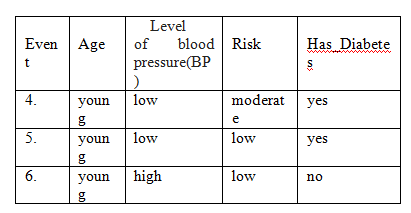


Fig. 8. young age attributes on has\_diabetes

There are 3 occurrences of young age. Has\_Diabetes of 1 item is no and 2 items are yes as shown below-

IE(Has\_Diabetes | Age=young) =

= 0.92

Now, calculate Information Gain of Age-

Gain(Has\_Diabetes | Age)=1-(0×+0×+)

=1-.345

=.655

To get information gain, every column should be calculated in the same way.

IG(Has\_Diabetes,Risk) = 0.06

IG(Has\_Diabetes,Level of BP) = 0.05

Compare all the information gain and we see that information gain of age has the highest score which is the parent node of the decision tree.

So-

The first decision tree is-

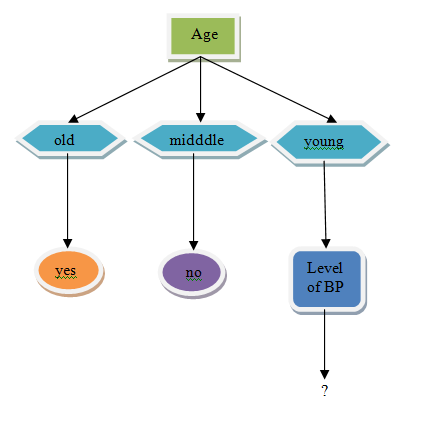


Fig. 9. First decision tree

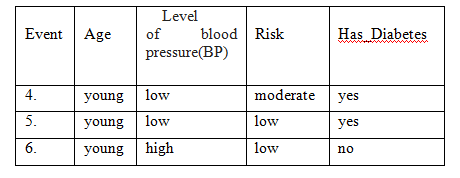


Fig. 10.Again spilt sample training data set

Information Entropy of young attribute is:

IE(Has\_Diabetes | Age=young) = 0.92

Split young attribte with risk and level of BP

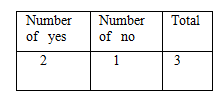


Fig. 11. Calculate number of age attributes

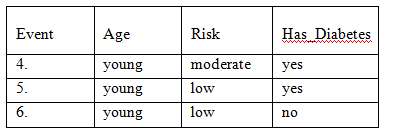


Fig. 12. spilt level of Risk attributes on young age

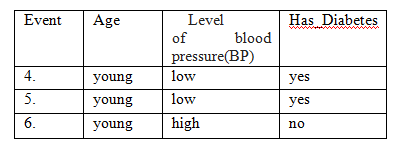


Fig. 13. spilt level of level of BP attributes on young age

IG(Risk) = 0.25

IG(evel of BP) = 0.92

Compare all the Information Gain and the Information Gain of the level of BP is greater than the Information Gain of Risk. That's why we choose the Level of BP node.

The final decision tree is-

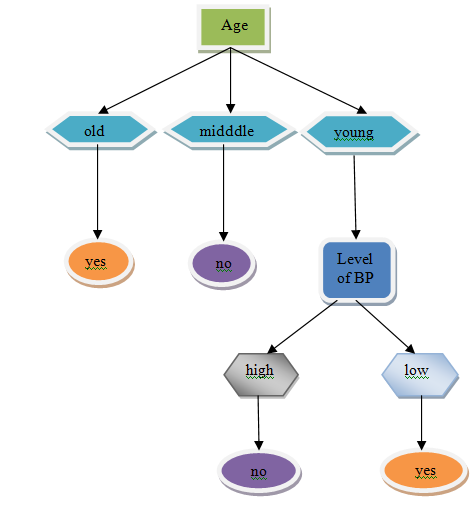


Fig. 14. Final decision tree

Apply if-else rules on the final decision tree for a better understanding.

1.IF “Age”=”old” THEN ”Has\_Diabetes”=”yes”.

2.IF “Age”=”middle” THEN ”Has\_Diabetes”=”no”.

3.IF “Age”=”young” and “level of BP” = “low” THEN ”Has\_Diabetes”=”yes”.

4.IF “Age”=”young” and “level of BP” = “high” THEN ”Has\_Diabetes”=”no”.

APPLICATION OF ID3 ALGORITHM

ID3 algorithm is a kind of decision tree algorithm which is widely used in every areas.Give some examples for the application of id3 algorithm.

**1.Application of ID3 in Diabetes**

Diabetes is a life long disease that the human race is found challenges that have existed ever since Year. ID3 Algorithm applied for predicting diabetes disease. This algorithm can also be applied in other fields like weather forecasting, business transactions, agricultural fields, etc.

**2.Application of ID3 in the Educational Sector**

Id3 algorithm is used in educational sectors for analysis, clustering, archiving. The summary presented is often related to techniques described in a variety of research.To use id3 algorithm, easily decide a decision because it gives fastest and smallest tree .

ADVANTAGES OF ID3 ALGORITHM

* Easily understand the final decision.
* Works with whole data sets.
* It can be explained at a sight.
* Create a tree very fast.
* For each attribute, estimate the information spread.

DISADVANTAGES OF ID3 ALGORITHM

* The biggest problem of the id3 algorithm is the overfitting problem.
* It cannot work with multiple attributes.
* It doesn't search for any missing values.
* Without pruning carelessly, It builds complex trees.

CONCLUSION

In this research paper, we represented the ID3 algorithm. ID3 algorithm is efficient in classifying huge datasets and the simplest algorithm. This paper provides some basic fundamental ideas about the ID3 algorithm like pros and cons, procedures, applications. For mining a data set, It is the most efficient algorithm.

REFERENCES

[1]. DavinderKaur, Rajeev Bedi, Dr. Sunil Kumar Gupta (2015),"REVIEW OF DECISION TREE DATA MINING ALGORITHMS:ID3 AND C4.5", International conference on Information Technology and Computer Science,(ICITCS)2015.

[2]. Himani Sharma, Sunil Kumar(2016),"A Survey on Decision Tree Algorithms of Classification in Data Mining"International Journal of Science and Research (IJSR), 2016.

[3]. R. Quinlan,“Induction of decision trees”,Machine Learning, Vol. 1, No.1, pp.81-106, 1986.

[4]. L.Surya Prasanthi, R.Kiran Kumar,"ID3 and Its Applications in Generation of Decision Trees across Various Domains- Survey", (IJCSIT) International Journal of Computer Science and Information Technologies,Vol.6 (6),2015.

[5]. Quinlan JR (1987) Simplifying decision trees. International Journal of Man-Machine Studies*,*27(3):221–234.

[6] WANG Ying-ying, LI Yi-bin\*, RONG Xue-wen ” Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree” Chinese Automation Congress (CAC),2017