An Implementation of Decision tree Learning Algorithm for Information Retrieval and Machine Learning

[1]Ms. X. Jose Suganya, [2]Dr. R. Balasubramanian

[1]Head of the Department, Department of Computer Applications, Shri Shankarlal Sundarbai Shasun Jain College for Women, T.Nagar, Chennai.

[1]Research Guide, JJ College of Arts and Science, Bharathidasan University, Trichy.

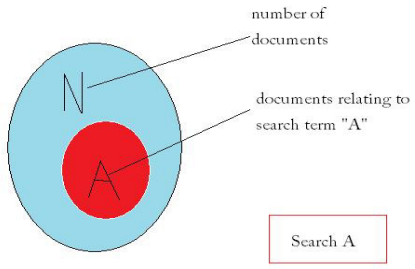
Abstract:

Information systems have enormous volume of data all over the web. Information Retrieval is the act of acquiring the right data at the right time when needed. "Google" is one of the reputed companies leading the Information Retrieval yet it is not seamless and perfect. There are trillions of documents on the web. Shifting through them manually is practically not possible as it is time consuming. Machine learning attempts to organise these resources and retrieve them according to the query provided by the users. ID3 algorithm is one of the machine learning techniques that can help to classify data. The paper will look into, how ID3 fits into Information Retrieval and Machine Learning, how it works, implement it to classify some data and finally conclude on its performance metrics.

Keywords— **ID3, Machine learning, Data mining**

1.INTRODUCTION

ID3 is a mathematical algorithm for building the decision tree invented by J. Ross Quinlan in 1979.Uses Information Theory invented by Shannon in 1948.Builds the tree from the top down, with no backtracking. Information Gain is used to select the most useful attribute for classification. Machine Learning is a part of Artificial Intelligence. In Machine Learning computers are taught to find useful information from data. Machine Learning will allow automatic generation of patterns and rules in large data help predict outcome of queries and also help classify data. Machine Learning has wide range of application like Information Retrieval, Natural Language Processing, Biomedicine, Genetics, and Stock Market etc. We have so much data available on the internet. There are search engines like "Google" and "Yahoo" that try to find what we are looking for in the web. The data on the web is not meaningful at all if it is not available to the people who are looking for it. "Information Retrieval" is the process of finding this pre-stored information.



Suppose we have a set of documents "N" and a subset "A" of "N"

Fig 1. Information retrieval

That includes all the documents that relates to the keyword "A". When a user inputs "A" as the search keyword then the ideal situation in Information Retrieval would be to return all the documents in the set of "A". Let's look at Google, the most popular search engine for Information Retrieval in the web. A query of the phrase "Implementation of ID3 algorithm for classification" in Google returns 40,000 results in English. The Scholar version of Google returns 8000 results. We also find that all the results that it produces are not the exact matches of what we were looking for but the success ratio is acceptable.

There are two main processes involve in Information Retrieval, ***understanding what the user is looking for*** and ***organising the data*** so that all related data are grouped together for correct retrieval. The user in a web search normally inputs search terms in keywords. Web search engines can't fetch good results when natural language is used in the query so keywords are preferred. "User Modelling" can be done to check what the users input and what results they accepted for their query. This will help us determine what the user was looking for when he input a particular search

phrase. Organising the data that relates to the search term is known as Classification.

2.SEMANTIC WEB

One of the main ideas to organise data and the documents in the web is to have Meta data as compulsion. Meta data describes the what 'data' the data actually holds. This will make it very efficient to collect, organise and retrieve data. But there is no authority in the web that enforces the use of Meta data. Hyper Text Transfer Protocol (HTML) files, text files and multimedia files like JPEG, Flash and WMV files allow the use of Meta tags to describe their contents, size, author and other properties. We don't actively use the META tags which makes the goal of making the web semantic more difficult.

3.CLASSIFICATION

Classification is process of grouping together documents or data that have similar properties or are related. Our understanding of the data and documents become greater and easier once they are classified. We can also infer logic based on the classification. Most of all it makes the new data to be sorted easily and retrieval faster with better results.

Probabilistic Models

In Probabilistic Models the probability of each class and features are recorded with the help of the training data set. The outcome of the new data or its classification is based on these probabilistic models. One of the examples of Probabilistic Modelling is the Bayesian Model.

Neural Networks

In Neural Networks, the data and its output (nodes) are inter-connected in a web like structure through programming constructs which mimic the function of the neurons in the human brain. Based on these, when new data is place in one of the nodes, its output can be predicted or it can be classified accordingly.

Decision Trees

A decision tree classifies data using its attributes. It is upside down. The tree has decision nodes and leaf nodes. In Fig 2, "linkFromAcademias" attribute is a decision node and the "author" attribute is the leaf node. The leaf node has homogenous data which means further classification is not necessary. ID3 algorithm builds similar decision trees until all the leaf nodes are homogenous.

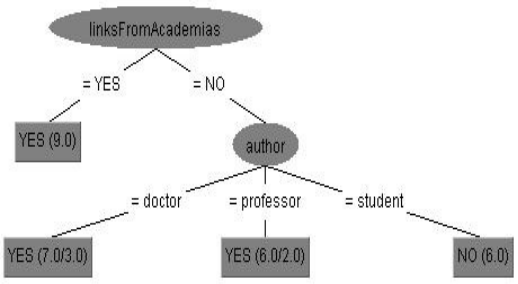


Fig 2. Sample Decision Tree

Training Data and Set

ID3 algorithm is a supervised learner. It needs to have training data sets to make decisions. The training set lists the attributes and their possible values. ID3 doesn't deal with continuous, numeric data which means we have to descretize them. Attributes such age which can values like 1 to 100 are instead listed as young and old.

Attributes	Values
Age	Young, Middle aged, Old
Height	Tall, Short, Medium
Employed	Yes, No

Fig 3 Training Set

The training data is the list of data containing actual values.

Age	Height	Employed
Young	Tall	Yes
Old	Short	No
Old	Medium	No
Young	Medium	Yes

Fig 4 Training Data

Entropy

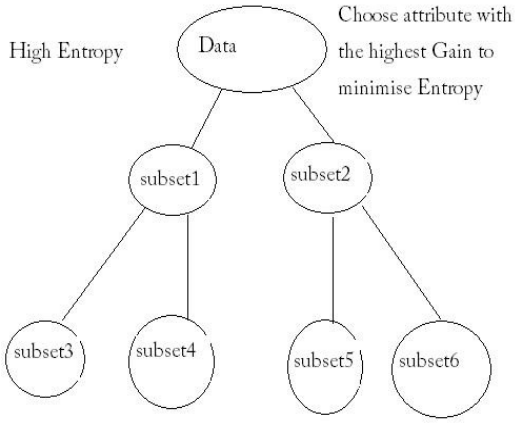


Fig 5. Entropy

Entropy refers to the randomness of the data. It ranges from 0-1. Data sets with entropy 1 means it is as random as it can get. A data set with entropy 1 means that it is homogenous. In Fig 6, the root of the tree has a collection of Data. It has high entropy which means the data is random. The set of data is eventually divided into subsets 3, 4, 5 and 6 where it is now homogenous and the entropy is 0 or close to 0. Entropy is calculated by the formula:

$$E(S) = -(p+)*log2(p+) - (p-)*log2(p-)$$

"S" represents the set and "p+" are the number of elements in the set "S" with positive values and "p –" are the number of elements with negative values.

The purpose of ID3 algorithm is to classify data using decision trees, such that the resulting leaf nodes are all homogenous with zero entropy.

Gain

In decision trees, nodes are created by singling out an attribute. ID3's aim is to create the leaf nodes with homogenous data. That means it has to choose the attribute that fulfils this requirement the most. ID3 calculates the "Gain" of the individual attributes. The attribute with the highest gain results in nodes with the smallest entropy.

To calculate Gain we use:

$$Gain(S, A) = Entropy(S) - S\ ((|S_v| / |S|) * Entropy(S_v))$$

In the formula, 'S' is the set and 'A' is the attribute. 'SV ' is the subset of 'S' where attribute 'A' has value 'v'. '|S|' is the number of elements in set 'S' and '|Sv|' is the number of elements in subset 'Sv'.

ID3 chooses the attribute with the highest gain to create nodes in the decision tree. If the resulting subsets do not have entropy zero or equal to zero then it chooses one of the remaining attribute to create further nodes until all the subsets are homogenous.

Weaknesses of ID3 Algorithm

ID3 uses training data sets to makes decisions. This means it relies entirely on the training data. The training data is input by the programmer. Whatever is in the training data is its base knowledge. Any adulteration of the training data will result in wrong classification. It cannot handle continuous data like numeric values so values of the attributes need to be discrete. It also only considers a single attribute with the highest attribute. It doesn't consider other attributes with less gain. It also doesn't backtrack to check its nodes so it is also called a greedy algorithm. Due to its algorithm it results in shorter trees. Sometimes we might need to consider two attributes at once as a combination but it is not facilitated in ID3. For example in a bank loan application we might need to consider attributes like age and earnings at once. Young applicants with fewer earnings can potentially have more chances of promotion and better pay which will result in a higher credit rating.

Using ID3 to Classify Academic Documents in the Web

Google has a scholar section where academic documents are available for students. It tries to classify if a document has any academic merit in it. The use of ID3 for classifying the scholar will be tested by identifying the attributes/ value pairs of academic documents and creating a training data set.

Attributes of a Web Document

A web document has many properties like, author, date created, description and tags. To identify if a document is academic or not, I have created a test scheme. The author should be an academic himself, like a doctor, professor a student. It is impossible to check the author's titles but it's just an assumption that we can. All journal articles are academic articles. Websites of academic institutions have ".ac" or ".edu" and governmental organizations have ".gov" in their domain names. So all documents published in those websites by a doctor, professor or students are considered academic. If a document is published in a ".com" website and written by any of the academics, it can only be considered scholar material if it is referred or linked to be by an academic or governmental website.

Google has more than five hundred million hyper-links in database to help them rank

pages. The document that the hyper-links point to are given ticks and ranked by the number of the ticks they get. So analysing if the document containing the hyper-links is from an academic or governmental website shouldn't be troublesome.

Training Set

Attribute	Possible Values
linksFromAcademias	YES,NO
Author	doctor, professor, student
Domain	com,edu,ac,gov
linksTo	null, low, high
journalArticle	YES,NO
scholarMaterial	YES,NO

Fig 7 Training Set to check Academic documents

The 'linksFromAcademia' attribute is to check if the document containing the hyper-link to the document is from an academic website or not. The 'author' attribute has values doctor, professor and student. It means documents authored by people other than this will not be considered at all. The 'linkTo' attribute shows the number of hyperlinks in the document. ID3 can't handle continuous numeric data so the possible values have been descretized as null, high and low. The 'journalArticle' checks if the document is from Journal or not. And finally the 'scholarMaterial' attribute is the decision attribute to decide if the document is academic or not.

Experimental Results

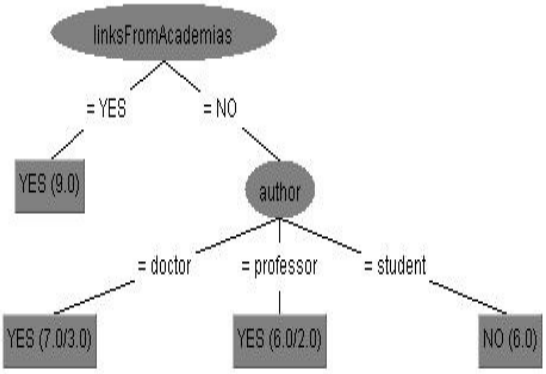
Fig 6. WEKA output

The 'linksFromAcademias' attribute had the highest gain so the node was created based on that attribute. The sub-set with values 'YES' for the 'linksFromAcademias' node was homogenous which meant there was no need to classify it further. There were 9 instances with the 'YES' value in that particular subset. The subset with 'NO' values for the 'linksFromAcademias' had nineteen instances and they were not homogenous (the entropy was still high). The subset with the 'NO' value was then divided by using the 'author' attribute because it had the highest gain. In the resultant subset the subset values was homogenous but the subsets with doctor and professor values were not. So the subset with doctor and professor values were divided on basis of the 'journalArticle' attribute which resulted into a perfectly homogenous. The tree had in total five nodes and eight leaves.

Weka producing the following evaluation report:

=== Evaluation on training set ===

=== Summary ===



Correctly Classified Instances	28	100%
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0%	
Root relative squared error	0%	
Total Number of Instances	28	

=== Confusion Matrix ===

```
 a  b   <-- classified as
17  0 |  a = YES
 0 11 |  b = NO
```

Inferring rules can be described by the following if else statements:

If linksFromAcademias = 'YES' then
{
Document = academic
}
Elseif author = student then
{
Document != academic (not equal to)
}
Elseif author = doctor
{
Elseif domain = 'edu', or 'gov' or 'ac' then
{

```
            Document = academic
}
            Elseif domain = com
{
Elseif journalArticle = 'YES' then
 {
Document = academic
 }
Else
 {
Document != academic (not equal to)
}
            }
}
Elseif author = professor then
{
            Elseif domain = 'edu', or 'gov'
or 'ac' then
             {
            Document = academic
}
            Elseif domain = com
 {
 Elseif journalArticle = 'YES' then
 {
Document = academic
 }
Else
 {
Document != academic (not equal to)
}
            }
}
```

The above nested if else statements actually states the rules to classify academic documents based on the training data set.

CONCLUSION

Classification is very essential to organise data, retrieve information correctly and swiftly. Implementing Machine learning to classify data is not easy to give the huge amount of heterogeneous data that's present in the web. ID3 algorithm depends entirely on the accuracy of the training data set for building its decision trees. The ID3 algorithm learns by supervision. It has to be shown what instances have what results. Due to this ID3 algorithm, I think, cannot be successfully classify documents in the web. The data in the web is unpredictable, volatile and most of it lacks Meta data. The way forward for Information Retrieval in the web, in my opinion would be to advocate the creation of a semantic web where algorithms which are unsupervised and reinforcement learners are used to classify and retrieve data.

5.REFERENCE

1. H. Hamilton, E.Gurak, F. Leah, W. Olive, (2000-2) "*Computer Science 831: Knowledge Discovery in Databases* "[Online] Available from: http://www2.cs.uregina.ca/~dbd/cs831/index.html

2. Bhatia, MPS and Khalid, Akshi Kumar (2008). *"Information retrieval and machine learning: Supporting technologies for web mining research and practice." Webology*, **5**(2), Article 55. [Online] Available from: http://www.webology.ir/2008/v5n2/a55.html

[3] ID3 algorithm. Page last modified January 13, 2012. Retrieved March 31, 2012, from http://en.wikipedia.org/wiki/ID3_algorithm.

[4] ID3 algorithm. Page last modified January 13, 2012. Retrieved March 31, 2012, from http://en.wikipedia.org/wiki/ID3_algorithm.

[5] CSE5230 Tutorial: The ID3 Decision Tree Algorithm. Monash University, Semester 2, 2004. Retrieved April 10, 2012, from http://www.csse.monash.edu.au/courseware/cse5230/2004/assets/decision treesTute.pdf.

[6] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[7]Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[8] http://www.kdnuggets.com/.