ID3 - Decision tree learning

Mykola Shevchenko 130708081

Abstract— **This paper describes how the ID3 algorithm is used to classify a set of data. ID3 consists in building a simple tree with branches that lead to a particular decision (leafs).**

I. INTRODUCTION

The process to classify a set of data consists of two main processes. First, the training mechanism builds a decision tree with a given set of training data. This information contains all values of each attribute, that will lead to a final decision, also called target attribute. Therefore all leaves in the tree will contain one of the values of the target attribute. The classifying data set will contain all of the attributes that will be used to determine the value of the target attribute of that data.

II. TRAIN IMPLEMENTATION

A. Recursive training

The recursive implementation of the algorithm allows to treat children nodes as root nodes. Therefore, the function begins by creating a root node that soon will be a child node. The children of the current root node will be mapped to the values of the attribute with the best gain. The best gain is calculated with a special formula described further in Section II.B. Subsequently, this attribute will be used as a criterion to split the training data. This means that the new training data will contain all examples that match one of the values of that attribute. Therefore, the new child will be added as one of the children of the root node. The recursive call then is made with the new childs training data for further splitting, if required.

B. Information Gain

The information gain of an attribute is calculated by subtracting the root node entropy of the training set and the information gain of that attribute. The root entropy is calculated by counting sequences of target attribute values [*Section II.C*]. The gain of an attribute is calculated in a similar way. Once the training set is split by an attribute, its entropy is calculated and the information gain is known. The attribute with best information gain is therefore used to create new children of the root node.

C. Entropy

The entropy is calculated on a data set. The process consists in counting occurrences of target attribute values in the training data. The probability of each value of the target attribute is then given by dividing the occurrences of each value by the total amount of occurrences. The entropy is given by the following formula where p is the calculated probability:

$$\sum -p * \log p / \log 2$$

The value of the entropy will depend on how good that attribute is to split its root training set. If the split data contains the same value of target attributes then the score is maximum = 1, otherwise that set will require further splitting. This is the concept of homogeneity of a training set.

D. Homogeneous training set

A training set is homogeneous when all of the examples' target attribute contains the same value. This means that the training set doesn't require any more splitting. This function call decides when to create a leaf node for the decision tree.

III. CLASSIFY IMPLEMENTATION

A. Classification

The classification of data means simply run through the tree comparing the example attributes value with the children values. This will determine what branch to navigate through and by simply following this method, the tree will lead to a leaf node. Therefore, a recursive methodology is implemented with the current children passed recursively as a parameter to further classify the example with attributes (with the best information gain).

IV. CONCLUSIONS

The current algorithm classifies easily any type of training sets with multiple values of target attribute. The classified examples final decision are printed on the screen. ID3 is used in especially in category classification, biometry and machine learning to predict future behaviour depending on past history. Therefore, it allows to classify any kind of data that can be represented in graph as a dimensional point.

ACKNOWLEDGMENT

I would like to thank my dear Yasmin for correcting grammatical errors in this paper.

REFERENCES

[1] Decision Tree Learning, Simon Colton, 2004, accessed 24th March 2016, ¡http://www.doc.ic.ac.uk/ sgc/teaching/pre2012/v231/lecture11.html¿