

Exploring Machine Learning: The ID3 algorithm

Mohammed Ibrahim
Computer Science Department
College of Science, Swansea University
Swansea, SA2 8PP, UK

December 25, 2011

Contents

1	Introduction	1
1.1	Machine learning	1
1.2	Decision trees	2
1.3	ID3	2
2	Tools	3
2.1	Source control management (Git)	3
2.2	Unit testing (JUnit)	3
2.3	Document writing (Latex)	3
2.4	Document of source code	3
3	Evaluation and application	3
4	Methodology and requirements document	3
4.1	Methodology	3
4.2	Requirements	3
4.3	Risk management	3

1 Introduction

*** yet to complete ***

1.1 Machine learning

Project which has been assigned to me is based on machine learning. To understand the domain knowledge I have been reading books on machine learning. In simple words machine learning can be described as a technique or method that involves making the machine to learn and behave based on training data given and past experience to improve its performance. “Machine learning is programming computers to optimize a performance criterion using example data or past experience”. [3]

Machine learning also covers concepts of artificial intelligence (AI) and these techniques are widely used in every field. Face detection, text recognition, strategy games, web searching application are among few to mention that we see in day-to-day life. Machine learning concepts are applied into many other fields such as mathematics, biology and statistics.

1.2 Decision trees

Decision Trees:

Decision tree is a learning technique that it is used to test an object and analyse it. It returns 'positive' or 'negative' value based on that decision can be taken for a tested object. At a lower level decision trees can be also be represented in the form of if-then rules that can be easily understand

Decision tree classifies in the form of tree structure where each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree mainly classify data using attributes and it consists of decision nodes and leaf nodes. The tree has number of branches representing with the tested attribute values. Leaf node attribute generate uniform result and it doesn't require any additional classification testing.

A leaf node indicates the value of target attribute where as a decision node states some test which can be carried out on single attribute-value, with one branch and sub-tree for each possible outcome of the test.

"Decision Tree classify instances by sorting them down the tree from the root to some leaf node, which provide the classifications of the instances, Each node in the tree specifies a test of some attribute of the instances and each branch descending ".[1] Decision tree are commonly used for obtaining information for the purpose of decision making. The tree starts with root node then user split each node recursively according to decision tree learning algorithm. According to Tom M. Mitchell "Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Decision tree learning is one of the most widely used and practical methods for inductive inference ".[1]

1.3 ID3

ID3 Algorithm

1. A mathematical algorithm for building the decision tree. 2. It has been invented by J. Ross Quinlan in 1979. 3. Uses Information Theory invented by Shannon in 1948. 4. It builds the tree from the top down, with no backtracking. 5. Information Gain is used to select the most useful attribute for classification.

In a case of ID3 algorithm, it takes three set of parameters. (Examples, Target attribute, Attributes) First is an example that represents the training set. Here training set contains both positive and negative samples. Target attribute is the one whose value has to be determined by using decision tree. And third parameter is the list of attributes that will be tested by the decision tree. Attribute selection is an important part of ID3 algorithm. With the attribute selection step, two terms comes into picture: Entropy and Information gain. With the attribute selection process, the algorithm decides which attribute will be appropriate for becoming a node in the tree. For an instance play ball. In this example, outlook, temperature, humidity, wind, play ball are attributes. Out of this attributes play ball is considered as classifier because depending on the value of play ball (yes or no), the decision will be made whether tennis can be played or not.

2 Tools

2.1 Source control management (Git)

2.2 Unit testing (JUnit)

2.3 Document writing (Latex)

2.4 Document of source code

3 Evaluation and application

4 Methodology and requirements document

4.1 Methodology

4.2 Requirements

4.3 Risk management