Exploring Machine Learning: The ID3 algorithm

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

January 3, 2012

Contents

1 Introduction

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".

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 ". 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] The 3 widely used decision tree learning algorithms are: ID3, ASSISTANT and C4.5. Based on research have decided to do an implementation on ID3 algorithm.

1.3 ID3

ID3 Algorithm:

ID3 algorithm deals with the generation of decision tree, J.Ross Quinlan developed ID3 at the University of Sydney in the year 1983. The fundamental idea of ID3 is to build the decision tree by using top-down greedy search through the given sets to test each attribute at every tree node. In 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)

Source control management (Git):

Git is a powerful, fastest, sophisticated distributed version control system this is quickly replacing subversion in open source and corporate programming communities. It is written in C language and it is active from several years. It designed to handle extremely large projects with speed and efficiency, but just as well suited for small personal repositories; it is especially popular in the open source community, serving as a development platform for projects like the Linux Kernel, Ruby on Rails, WINE or X.org.[4] Birth of GIT: In the year 2002 Linus Benedict Torvalds uses Bit-Keeper for tracking Linux when it gets better, he writes his own Source Control Management, GIT. Later GIT officially used to track Linux and released GIT 1.5.0 version in the year 2007. I will be using 1.7.8 version.

Repository: A repository is a set or collection of commits, the work which you have done past it shows in an archive and looks like project's working tree in your machine or someone else's. It holds a set of branches and tags, to identify certain commits by name.

The index : Git does not commit changes directly from the working tree into repository. Changes are first registered in the index,it is the way of confirming changes one by one before doing any commit.

Working tree :the directory in a file system called working tree which has repository by stating extension .git and it includes all the files and sub directories in that directory.

Commit : the word "commit" is often used by git, other revision control systems use the words "version".

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