International Journal of Computer Applications (0975 – 8887)
Volume 167 – No.11, June 2017

The Analytical Comparison of ID3 and C4.5 using WEKA

Vani Kapoor Nijhawan
Assistant Professor
VIPS, GGSIPU
Delhi

Mamta Madan, PhD
Professor
VIPS, GGSIPU
Delhi

Meenu Dave, PhD
Professor
Jagan Nath University
Jaipur

ABSTRACT
Data mining means to find out some useful information from a big warehouse of data and the process is aimed at unfolding old records and identifying novel patterns from the data. Data mining is used for classification and prediction. Many techniques and algorithms are available for mining the data. Out of many techniques, the decision tree is the simplest. This paper focuses on comparing the performance accuracy of ID3 and C4.5 techniques of the decision tree for predicting customer churn using WEKA. The data used for this research work has been collected by designing a survey form and getting it filled by around 150 mobile phone users belonging to a different gender, age groups and having different types of connection providers. For the data analysis in WEKA, the cross-validation method is used where a number of folds n (10 as standard as per the software) is used. From the results, it is observed that C4.5 algorithm exhibits better performance than ID3.

Keywords
Data mining, Decision tree, ID3, C4.5

1. INTRODUCTION
In the telecom sector, churning is a process that happens when a customer leaves the current network provider and goes to some other one because of their type of connection or some other reasons. For the purpose of analysis, data has been collected in the form of a survey being done on the users of different age groups and having different types of connections. So, the need is to analyze the collected data, to find some kind of a pattern, which can be used for future predictions. The major challenge for the companies is to identify the customers who are about to churn and to retain them by offering few schemes in which they may be interested. For this prediction, decision tree technique can be applied, due to its advantages.

1.1 Decision Trees
Decision trees are popularly used for prediction and classification. It is a simple and powerful way of knowledge representation [2]. The Decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [3]. Decision tree technique results in a set of If-then rules that are easy to understand and clear. They yield fast results.

Advantages of Decision Tree

- It is easy to understand and cheap to implement.
- Most decision tree algorithms can be applied both serially and parallelly. Parallel implementation of decision tree algorithms leads to the fast generation of results, especially for large datasets [4]. However, a serial implementation of decision tree algorithm is easy to implement and desirable when small or medium data sets are involved.

There are four more popularly used algorithms of decision tree i.e. ID3, CART, CHAID, C4.5. Out of these, this paper focuses on ID3 and C4.5.

1.1.1 ID3
The ID3 algorithm is a simple decision tree generating algorithm introduced by Quinlan Ross in the year 1986. It is the forerunner to the C4.5 algorithm. It applies top to down approach based on divide and conquers strategy. This does tree construction in two phases, i.e. tree building and pruning. An information gain measure is used to choose the splitting attribute amongst all attributes. It accepts categorical attributes only for designing a tree. It does not give accurate results when there is noise [5].

1.1.2 C4.5
This algorithm is a descendant of ID3 designed by Ross in 1993.It is also referred to as the J48 algorithm. Like ID3, it is also implemented serially [6], but it has more advantages over ID3. Some of them are:-

- It can handle both categorical as well as discrete data.
- The decision tree algorithm C 4.5 was one of the first algorithms, which can handle missing values. Quinlan (author of the algorithm) [7], has explained, how C 4.5 handles missing values. Missing attribute values are simply not used in gain and entropy calculations [8].
- C4.5 does tree pruning, by going back through the tree after its creation. It attempts for removing branches which are not of help by replacing internal nodes with leaf nodes [8][6].

2. RELATED WORK
[5] explored three algorithms of the decision tree, namely, ID3, C4.5, CART and compared their performance in the field of education data mining and have shown in their analysis that C4.5 is better than ID3, but CART is better than C4.5.

[6] have done a data mining for predicting typhoid fever after collecting data from a well-known Nigerian hospital and their wok shows that out of the three techniques i.e, ID3, C4.5, and MLP, MLP gives the best results but C4.5 also gives better results as compared to ID3.