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.

Another experimentation in reference [9] uses datasets of 3 different sizes to show the performance of ID3 and C4.5 and in all three cases, C4.5 outperforms ID3 algorithm.

[10] compared 3 algorithms J48, Random Tree and SimpleCART. On the basis of comparison, it was demonstrated that J48 algorithm has worked in a better way for predicting student's post-graduation course.

In another study [11], the popularity of decision tree was shown on the basis of the review of various research papers.

So, there are numerous examples to show the application of various algorithms of the decision tree in data mining in various fields. Also, it can be applied in the field of mobile telecom churn prediction and performance comparison can be done for the various algorithms.

3. METHODOLOGY

This section of the paper describes the methodology adopted for the process of data mining.

3.1 Data Collection And Description

The dataset used in this research is the data of mobile users who have churned or not churned. It is collected from an online survey done amongst the mobile users of different gender, age group and having different network providers.

Table1. User's description and attributes.

S. No	Attribute Name	Description	Data Type
1	Age Group	Age group	Nominal
2	Gender	Gender	Nominal
3	Network Provider	Type of provider(Airtel/ Vodafone/Others)	Nominal
4	Churning Behaviour	Does the user churn(Yes/ No)	Nominal

3.2 Data Pre-Processing

This step involves cleaning the data by removing missing values and filtering the data so that it can be in a format accepted by WEKA.

During this step, the dataset was cleaned by removing missing values and also the age field in the dataset was converted from numeric to nominal, so that it can be accepted by ID3 algorithm in WEKA.

3.3 Data Integration

In this step, the data is gathered from different sources and combined into a common pool. The data collected from online survey was in different excel files. These files were combined and records were concatenated into one single file.

3.4 Data Transformation

This step involves converting the data into the required format. The data file received from online survey was in the excel .xlsx format. No conversion was required but it was saved into .csv (Comma Separated Value) form so that it can be used and processed in WEKA.

3.5 Data Training

In WEKA, the cross-validation method is used where a number of folds 'n' is specified [1]. In this case, the records are shuffled and after that divided into n folds of equal size. Every iteration uses one fold for testing and the remaining n-1 folds for training the classifier. Then results of tests are collected and analyzed to find an average over all folds. It returns the cross-validation estimate of the accuracy[1].

4. IMPLEMENTATION

In this research work, WEKA tool is used for analyzing mobile telecom data. The reason for selecting WEKA tool is because it is an open source software issued under the General Public Licence (GNU). It is a collection of multiple machine learning algorithms to perform data mining. An algorithm can be applied directly to any dataset. WEKA can implement algorithms for data preprocessing, regression, classification, association rules and clustering. It also has a visualization tool for graphical representation [12]. The ID3 and C4.5 algorithms of decision tree were implemented on the collected dataset using the 10 folds cross-validation option under test options. This predictive model will then be useful in predicting the mobile telecom churning.

5. RESULTS AND DISCUSSION

After performing data pre-processing and cleaning and applying the WEKA tool on three datasets consisting of 50, 100 and 150 records respectively, it was observed that in every case C4.5 algorithm outperformed the ID3 algorithm in terms of accuracy (correctly classified instances), Kappa statistics, Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error.

Table2. Accuracy comparison between ID3 and C 4.5 algorithm

Size of Data Set	Algorithm	
	ID3 (%)	C4.5 (%)
50	42	54
100	44	45
150	50.6667	51.3333

Here, the accuracy of the model is defined by the number of instances classified correctly[1]. Also, as per [1], performance can be measured by counting the proportion of correctly predicted examples in an unseen test dataset. This value is the accuracy.

Table3. Summary of the results in WEKA.

Evaluat-ion Criteria	ID3 (50)*	C4.5 (50)*	ID3 (100)*	C4.5 (100)*	ID3 (150)*	C4.5 (150)*
KS	-.1294	0.0103	-0.0969	-0.1326	0.0228	-.0163
MAE	0.5153	0.494	0.511	0.4987	0.486	0.5013
RMSE	0.5565	0.5139	0.5505	0.5086	0.52	0.5089
RAE (%)	111.204	101.1206	105.5488	100.004	99.9543	101.0892
RRSE (%)	115.0724	103.9483	111.7993	101.8037	105.4368	102.201

Here,

KS - means Kappa Statistics

MAE - means Mean Absolute Error

RMSE - means Root Mean Squared Error

RAE - means Relative Absolute Error

RRSE - means Root Relative Squared Error

and

ID3(50)* - means ID3 applied on a dataset of 50 records.

C4.5(50)* -Means C4.5 applied on a dataset of 50 records.

ID3(100)*- means ID3 applied on a dataset of 100 records.

C4.5(100)*-Means C4.5 applied on a dataset of 100 records.

ID3(150)* - means ID3 applied on a dataset of 150 records.

C4.5(150)*-Means C4.5 applied on a dataset of 150 records.

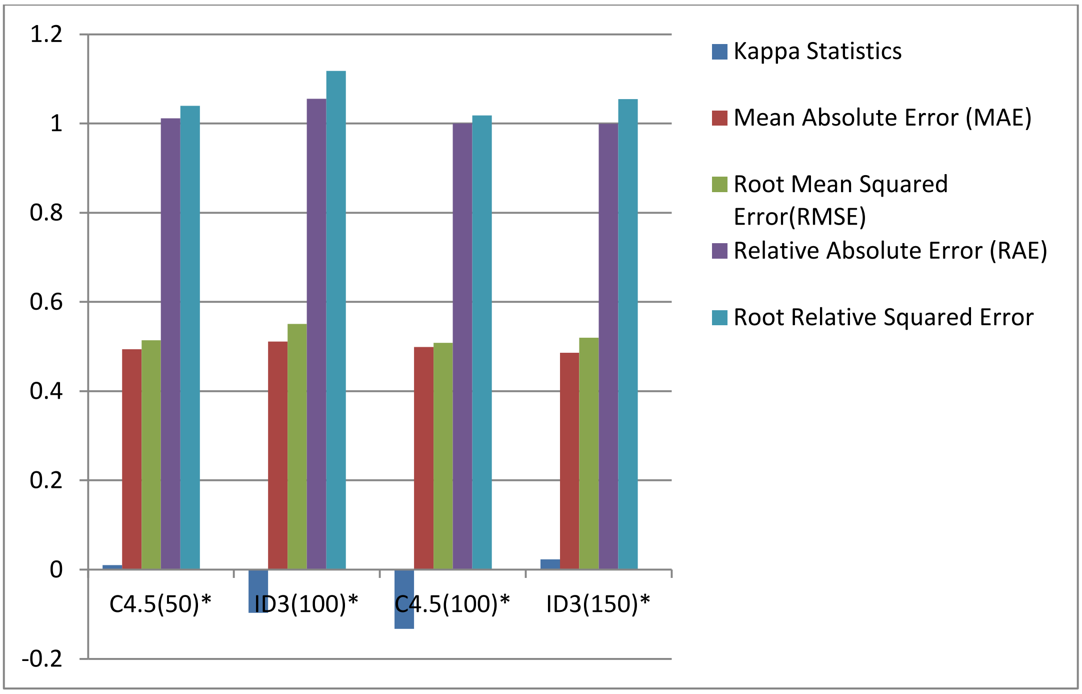

Fig. 1 Summary of the results in WEKA

As shown in the table3 and Fig.1, all values for MAE, RMSE, RAE and RRSE are showing a lesser value in case of the C4.5 algorithm, thus making it a better technique to predict the telecom churn.

Other factors, like TP (True Positive) and FP (False Positive) can also be taken into consideration as follows:-

Table4. Results of TP and FP Rate for ID3 and C4.5

Classi-fier	Data set Size	TP Rate	FP Rate	Preci-sion	Recall	Class
ID3	50	0.538	0.667	0.5	0.538	N C*
ID3	50	0.333	0.462	0.368	0.333	C*
C4.5	50	0.724	0.714	0.583	0.724	N C*
C4.5	50	0.286	0.276	0.429	0.286	C*
ID3	100	0.49	0.587	0.481	0.49	N C*
ID3	100	0.413	0.51	0.422	0.413	C*
C4.5	100	0.679	0.809	0.486	0.679	N C*
C4.5	100	0.191	0.321	0.346	0.191	C*
ID3	150	0.575	0.552	0.554	0.575	N C *
ID3	150	0.448	0.425	0.469	0.448	C*
C4.5	150	0.72	0.735	0.541	0.72	N C*
C4.5	150	0.265	0.28	0.439	0.265	C*

Where, N C* means- Not Churn and C* means- Churn.

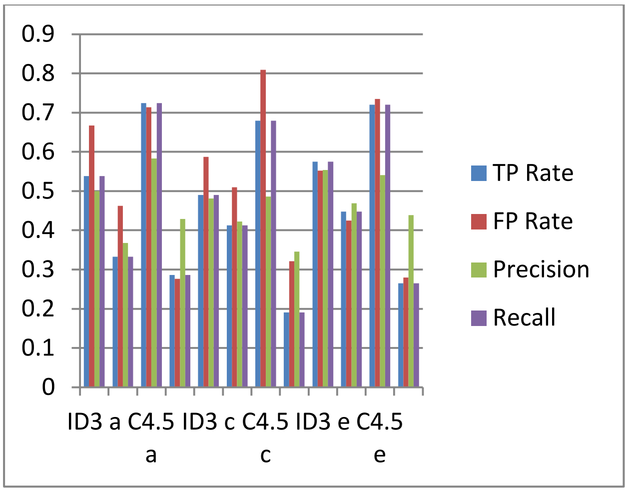

Fig. 2 Results of TP and FP Rate for ID3 and C4.5

Here,

ID3 a - means ID3 applied on a dataset of 50 records showing Not Churn records.

ID3 b - means ID3 applied on a dataset of 50 records showing Churn records.

C4.5 a -Means C4.5 applied on a dataset of 50 records showing Not Churn records.

C4.5 b -Means C4.5 applied on a dataset of 50 records showing Churn records.

ID3 c - means ID3 applied on a dataset of 100 records showing Not Churn records.

ID3 d - means ID3 applied on a dataset of 100 records showing Churn records.

C4.5 c -Means C4.5 applied on a dataset of 100 records showing Not Churn records.

C4.5 d -Means C4.5 applied on a dataset of 100 records showing Churn records.

ID3 e - means ID3 applied on a dataset of 150 records showing Not Churn records.

ID3 f - means ID3 applied on a dataset of 150 records showing Churn records.

C4.5 e-Means C4.5 applied on a dataset of 150 records showing Not Churn records.

C4.5 f-Means C4.5 applied on a dataset of 150 records showing Churn records.

It can be clearly seen from the above table that the TP rate and FP Rate of C4.5 are more than of ID3 for all the cases.

6. CONCLUSION AND FUTURE SCOPE

Under this research work, two techniques of the decision tree for data mining have been studied, on three sets of data having 50,100 and 150 records.

As per the observations, it is seen that C4.5 gives better results as compared to the ID3 algorithm for these data sets. Also, it can be said that C4.5 can be applied on a dataset to predict the churning behavior of mobile phone users in an efficient manner.

To improve the classification accuracy, further research work can be done using different mining algorithms like MLP of Neural network or others.

In future, the research work can also be carried out using the same dataset, by exploring some other mining tool like RStudio.

7. REFERENCES

[1] WEKA Manual for Version 3-7-8 Remco R. Bouckaert Eibe Frank Mark Hall Richard Kirkby Peter Reutemann Alex Seewald David Scuse, 2013. Available at: http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf as on 10-12-2016.

[2] Surjeet K. Y., Saurabh P., (2012).Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" WCSIT, ISSN: 2221-0741 Vol. 2, No. 2, 51-56.

[3] Han J.and Kamber M., (2011). Data Mining: Concepts and Techniques, Morgan Kaufmann Publish.

[4] Shafer, J., Agrawal, R., Mehta, M. Fast serial and parallel classification of very large databases. In Proc. of the 22nd Int'l Conference on Very Large Databases. 1996.

[5] T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.Prasanna Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data", I.J.Modern Education and Computer Science, 2013. Published Online June 2013 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2013.05.03

[6] O..O. Adeyemo, T. .O Adeyeye, D. Ogunbiyi (2015). Comparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever, IEEE African Journal of Computing & ICT ISSN: 2006-1781 Vol 8. No. 1.

[7] J. Ross Quinlan. (1993). C4.5: Programs for Machine Learning. Morgan Kaufman.

[8] Pallavi Mude, Rahila Sheikh, "Study of Decision Tree Classification Algorithms using Matrimonial System", International Journal of Computer & Organization Trends,Volume 18,No. 1, March 2015.

[9] Badr HSSINA, Abdelkarim MERBOUHA,Hanane EZZIKOURI,Mohammed ERRITALI, "A comparative study of decision tree ID3 and C4.5, " International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2014.

[10] Jaimin N. Undavia, Dr. P.M.Dolia and Dr. AtulPatel, " Comparison of Decision Tree Classification Algorithm to Predict Student's Post Graduation Degree in Weka Environment", International Journal of Innovative and Emerging Research in Engineering,Vol 1, Issue 2, 2014.

[11] Dr. Mamta Madan Dr. Meenu Dave Vani Kapoor Nijhawan, " A Review on: Data Mining for Telecom Customer Churn Management", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, Issue 9, 2015.

[12] "Machine Learning with WEKA" WEKA Explorer Tutorial for WEKA Version 3.4.3, Svetlana S. Aksenova, 2004. Available at: https://www.scribd.com/document/247244990/WEKA-Tutorial-Presentation as on 10-12-2016.