TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

COMPARATIVE ANALYSIS OF DECISION TREE CLASSIFICATION ALGORITHMS

Thesis

Submitted to:

Central Department of Computer Science and Information Technology
Kirtipur, Kathmandu, Nepal

In Partial Fulfillment of the Requirements for

Master's Degreein Computer Science& Information Technology

Submitted by:

Ms.Shikha Karmachrya

Date: 30 May, 2014

Under the Guidance of

Prof. Dr. Subarna Shakya

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk, Nepal

ACKNOWLEDGEMENT

I am highly indebted tomy thesis supervisor**Prof. Dr.SubarnaShakya,**Department of Electronics and Computer Engineering Institute of Engineering, Pulchowk, Nepal for his valuable and constructive suggestion during the planning and development of this research. His willingness to give his time so generously has been very much appreciated.

My gratitude goes to **Prof. Dr. SubarnaShakya**of the **Research Committee**, Central Department of Computer Science and Information Technology, Kirtipur, Kathmandu, Nepal for providing me an opportunity to conduct this research as well as frequent feedbacks.

I would like to gratitude **Mr. IndraChaudhary** for his kind support as well as valuable suggestions and frequent discussion. Similarly, I extend my debt of gratitude to my friends and best wishers who exhorted me for the initiation.

ABSTRACT

In our daily life there is lots of records, phone call records, salary records, homework records, assignment record, personal details record, sales record, song, videos and so on. These all records kept in a table are called data; we have lots of data in different field. Whenever there is data we can have lots of information, patterns, meaning etc. Data mining applications has got rich focus due to its significance of classification algorithms. The comparison of classification algorithm is a complex and it is an open problem. First, the notion of the performance can be defined in many ways: accuracy, speed, cost, reliability, etc. Second, an appropriate tool is necessary to quantify this performance. Third, a consistent method must be selected to compare with the measured values. The selection of the best classification algorithm for a given dataset is a very widespread problem. In this sense it requires to make several methodological choices. So this research focused in the analysis of decision tree classification algorithm in different datasets of multiple attributes and multiple instances. Where analysis was done among five decision tree algorithms (BFTree, J48, RandomTree, REPTree and SimpleCart).J48 was able to classify 82.16% of the data correctly which was best among all in comparison to results of evaluation metrics (Accuracy, Precision, Recall and F-Measure) and SimpleCart was able to build decision tree with small tree size of 17.24 (averaged value).

Keywords:BFTree,CART**,** Data Mining, Decision Tree, J48,RandomTree, REPTree.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATION

API	:	Application Programming Interface
ARFF	:	Attribute-Relation File Format
CART	:	Classification and Regression Tree
CDR	:	Call Detail Record
GATree	:	Genetically Evolved Decision Tree
GNU	:	General Public License
ID3	:	Iterative Dichotomiser
KDD	:	Knowledge Discovery from Data
MARS	:	Multivariate Adaptive Regression Splines
Q0S	:	Quality of Service
REPTree		Reduced Error Pruning Tree
RF	:	Random Forest
RT	:	Random Tree
TN	:	TreeNet
URL	:	Uniform Resource Locator
WEKA	:	Waikato Environment for Knowledge Analysis
WWW	:	World Wide Web