



**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 Degree in 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 to my thesis supervisor **Prof. Dr. Subarna Shakya**, 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. Subarna Shakya** 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. Indra Chaudhary** 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**

<b>TITLE</b>	<b>PAGE</b>
<b>COVER PAGE</b>	
<b>ACKNOWLEDGEMENT</b>	<b>II</b>
<b>ABSTRACT</b>	<b>III</b>
<b>TABLE OF CONTENTS</b>	<b>IV-VI</b>
<b>LIST OF TABLE</b>	<b>VII</b>
<b>LIST OF FIGURE</b>	<b>VIII</b>
<b>LIST OF ABBRIVATION</b>	<b>IX</b>
<b>CHAPTER 1            INTRODUCTION</b>	<b>1 -3</b>
1.1.    Background of the Study	1
1.2.    Statement of the Problem	1
1.3.    Objectives of the Study	2
1.4.    Limitation of the Study	2
1.5.    Structure of the Report	2
<b>CHAPTER 2            LITERATURE REVIEW</b>	<b>4-11</b>
2.1.    Data Mining	4
2.1.1. Classification	5
2.2.    Application Programming Interface (API) for Data Mining	9
2.2.1. WEKA	10
2.3.    Related Works	10
<b>CHAPTER 3            RESEARCH METHODOLOGY</b>	<b>12 - 18</b>
3.1.    Background	12

3.2.	Source of Data	12
3.3.	Algorithms	12
	3.3.1. BFTree	12
	3.3.2. J48	13
	3.3.3. RandomTree	15
	3.3.4. REPTree	15
	3.3.5. SimpleCart	16
3.4.	Experiments Protocol	17
	3.4.1. Validation	18
	3.4.2Experiment Setup	18
<b>CHAPTER 4</b>	<b>EXPERIMENT AND RESULT ANALYSIS</b>	<b>19-28</b>
4.1.	Background	19
4.2.	Tool	19
4.3.	List of Data	20
4.4.	Experiments and Results	20
	4.4.1. 5-fold cross-validation	21
	4.4.2. Confusion Matrix	21
	4.4.3. Accuracy	21
	4.4.4. Precision	23
	4.4.5. Recall	24
	4.4.6. F-Measure	25
	4.4.7. Tree Size	26
	4.4.8. Averages ofEvaluation Metrics	27
4.5.	Context Analysis	27

<b>CHAPTER 5</b>	<b>CONCLUSION AND FUTURE WORK</b>	<b>29</b>
5.1.	Conclusion	29
5.2.	Future Works	29
<b>REFERENCES</b>		<b>30-31</b>
<b>APPENDIX</b>		<b>32-36</b>

## LIST OF TABLE

<b>TABLE</b>	<b>TOPIC</b>	<b>PAGE</b>
Table 3.1:	Experimental Parameters	18
Table 4.1:	List of Data	20
Table 4.2:	Confusion Matrix	21
Table 4.3:	Result taking Accuracy	22
Table 4.4:	Result taking Precision	23
Table 4.5:	Result taking Recall	24
Table 4.6:	Result taking F-Measure	25
Table 4.7:	Result taking Tree Size	26
Table 4.8:	Result taking averages of Evaluation Metrics	27

## LIST OF FIGURE

<b>FIGURE</b>	<b>TOPIC</b>	<b>PAGE</b>
Figure 2.1:	Data mining as confluence of multiple disciplines	4
Figure 2.2:	Decision tree example	5
Figure 2.3:	Possibility of attribute as node	7
Figure 4.1:	Graph of table 4.3	22
Figure 4.2:	Graph of table 4.4	23
Figure 4.3:	Graph of table 4.5	24
Figure 4.4:	Graph of table 4.6	25
Figure 4.5:	Graph of table 4.7	26
Figure 4.6:	Graph of table 4.8	27

## 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
QoS	: 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