ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 6, June 2012

A Survey on Frequently Used Decision Tree Algorithm and There Performance Analysis

Anshu Tiwari, Dr. Vijay Anant Athavale

Abstract -In the machine learning process, classification can be described as supervise learning algorithm. Any kind of learning involves, three phases according to their use in different fields i.e. experimental data collection, algorithm implementation and model evaluation. A classification technique has various properties that enable the representation of structures. These structures reflect the knowledge of the domain being classified. Industries, education, business and many other domains required knowledge for growth. This knowledge is formed using discovery of the data which is generated by different domain is possible using data mining. Extraction of knowledge from data in a human-understandable structure is the main goal of data mining. In this paper we include the most frequently used algorithms and their performance evaluation, over different size and type of data sets.

Keywords: **Learning, Classification, Knowledge, Human-Understandable Structure.**

I. INTRODUCTION

All The process of data mining consists of three stages:

Exploration: In this stage data preparation mainly include cleaning data, data transformation, selection of subset records and for large data sets with large number of features it also require to do feature selection.

Model building and validation: This is the stage in which we consider a variety of models and select the best one based on their predictive performance i.e. explaining the unpredictability in question and produces good results across samples.

Deployment: Than final stage involves, using the model selected as best in the previous stage is choose and applying it to new data in order to generate predictions or estimates of the expected outcome.

In the field of business information and knowledge management data mining is becoming increasingly popular. Significant consideration has been devoted by the machine learning research community to the task of acquiring "classification knowledge" for which, among a Pre declared set of available classes, the objective is to choose the most appropriate class for a given case.

In this paper we include the following facts Collect the experimental data, Find most frequent and effective algorithms that perform all better over different kind of data sets Implement all the best fit algorithms Process the experimental data using implemented methods Evaluate performance in all cases. And at last simulate the results.

II. BACKGROUND STUDY

All Decision trees have various advantages:

1. Simple to understand and construct. It's easy to understand decision tree models after a short explanation.
2. Requires little data training. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
3. Able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable. Ex: relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.
4. Uses a white box model. If a given situation is visible in a model the justification for the circumstance is easily explained by Boolean logic. An example of a black box model is an artificial neural network since the explanation for the results is complex to understand.
5. Probable to validate a representation using statistical tests. That makes it possible to explanation for the reliability of the model.
6. Robust. Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

Performs well with large data in a short time Large amounts of data can be analyzed using standard computing resources.

1. Decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree.
2. Decision-tree learners can create over-complex trees that do not generalize the data well. This is called over fitting. Mechanisms such as pruning are necessary to avoid this problem.
3. There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems. In such cases, the decision tree becomes prohibitively large. Approaches to solve the problem involve either changing the representation of the problem domain (known as propositionalisation) or using learning algorithms based on more expressive representations (such as statistical relational learning or inductive logic programming).

For data including categorical variables with different number of levels, information gain in decision trees are biased in favour of those attributes with more levels. There are various decision tree tools are available with some limitations. These systems working process is complex thus we propose a new simple and easy working data model for frequently used data mining techniques, which also have been discussed in literature survey. Our complete system is designed using three popular data

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 6, June 2012

mining algorithm. Namely ID3, C4.5, SLIQ. I am using and also calculating parameters like **accuracy**, **memory used**, **build time** and **search time**. These parameters cover all the important aspects of decision trees which help us to evaluate models. I am also including the graphs for analysis of the data size and data type effect on the different models. After all required to enhance performance by boosting of these popular models and compare it with each other. By using these decision models we make selection of tree on the basis of results and conclusion is made which decision tree performs better for particular application.

III. SYSTEM DETAIL

Our complete system is designed in three major modules.

1. Experimental data selection: Different type of data selected as the experimental data set. To get the performance is varies or not according to data. Here we collect data of different size and different types. We use nominal data (i.e., not consisting of numerical values) and numerical data both to evaluate results.
2. Data analysis using the selected data models: Here the implementation of algorithms includes. Data analysis using different algorithm includes simply data analysis and after boosting of algorithm data analysis.
3. Result analysis: Different system generated resultant parameters are generated. Results analysis includes the performance analysis of system with boosting and without boosting of the system. And on different parameters like accuracy, memory uses, time taken to build model and search time. At last we conclude that which model is best for nominal data and which on is best for numeric data.

Fig 1. Block Diagram of Proposed System

IV. ASSUMPTIONS AND DEPENDENCIES

The raw data are unstructured and individual listings aren't always clean-cut and complete as far as the fields listed above are concerned, this is problematic. Thus we select data for experiment purpose in ARRF format. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. This document describes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date attributes, and sparse instances).ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information. Hence system can accept file in only ARFF format.

One of the problem with decision tree is that it is not dynamic i.e. the reliability of the information in the decision tree depends on feeding the precise internal and external information at the onset. Another fundamental flaw of the decision tree analysis is that the decisions contained in the decision tree are based on expectations, and irrational expectations can lead to flaws and errors in the decision tree. Although the decision tree follows a natural course of events by tracing relationships between events, it may not be possible to plan for all contingencies that arise from a decision, and such oversights can lead to bad decisions.

V. ALGORITHM USED

A.ID3 Algorithm

INPUT: Experimental data set *D* which is showed by discrete value attributes.

OUTPUT: A decision tree *T* which is created by giving experimental dataset.

i) Create the node N;
ii) If instance is belong to the same class
iii) Then return node N as the leaf node and marked with CLASS *C*;
iv) IF attribute List is null, THEN
v) Return the node *N* as the leaf node and signed with the most common CLASS;
vi) Selecting the attribute with highest information gain in the attribute List, and signing the test_attribute;
vii) Signing the node *N* as the test_attribute;
viii) FOR the known value of each test_attribute to divide the samples;
ix) Generating a new branch which is fit for the test_attribute= *ai* from node *N*;
x) Suppose that *Ci* is the set of test_attribute=*ai* in the samples;
xi) IF *Ci* is null THEN
xii) Adding a leaf node and signed with the most common CLASS;
xiii) ELSE we will adding a leaf node return by the Generate_decision_tree.

B.Supervised Learning in Quest (SLIQ) Algorithm

SLIQ is a decision tree classifier that can handle both numerical and categorical attributes it builds compact and accurate trees. It uses a pre-sorting technique in the tree growing phase and an inexpensive pruning algorithm. It is suitable for classification of large disk-resident datasets, independently of the number of classes, attributes and records.

Tree Building
MakeTree (Training Data *T*)
Partition (*T*)
Partition (Data *S*)
If (all points in S are in the same class)
Then return;
Evaluate Splits for each attribute A;

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 6, June 2012

Use best split to partition S into S1 and S2;
Partition (S1);
Partition (S2);

The *gini* index is used to evaluate the "goodness" of the alternative splits for an attribute

If a data set *T* contains examples from *n* classes, *gini(T)* is defined as

$$\text{gini}(T) = 1 - \sum P_j^2$$

Where pj is the relative frequency of class j in T. After splitting *T* into two subset *T1* and *T2* the gini index of the split data is defined as

$$\text{gini(T)}_{\text{split}} = \frac{|T1|}{|T|} gini(T1) \frac{|T2|}{|T|} gini(T2)$$

The first technique implemented by SLIQ is a scheme that eliminates the need to sort data at each node It creates a separate list for each attribute of the training data. A separate list, called *class list*, is created for the class labels attached to the examples. SLIQ requires that the *class list* and (only) one *attribute list* could be kept in the memory at any time.

VI. COMPARATIVE RESULTS FOR ACCURACY

Here we simulate comparative results for accuracy between ID3, SLIQ and C4.5 in the graph blue line shows ID3, pink for SLIQ and yellow line shows for SLIQ.

Accuracy between ID3, C4.5 and SLIQ

The below given graph show the accuracy of the ID3, C4.5 and SLIQ algorithm. All three algorithm shows one characteristic when size of data set is small then accuracy is high and data set size is large the accuracy is reduced. By using graph we can see at the initial state when the size of data set is too large then SLIQ and C4.5 reflect similar accuracy. And as size reduces the similarity of accuracy pattern between ID3 and C4.5 is much similar.

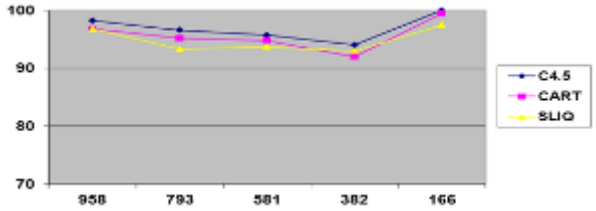

Fig 2: Shows Accuracy between ID3, C4.5 and SLIQ

VII. COMPARATIVE RESULTS FOR BUILD TIME

Here we show the comparison between three algorithms in the domain of time in both manners with boosting and without boosting. To differentiate more accurately ID3 is represented using blue lines, pink lines for C4.5 and Yellow Line represents SLIQ algorithm.

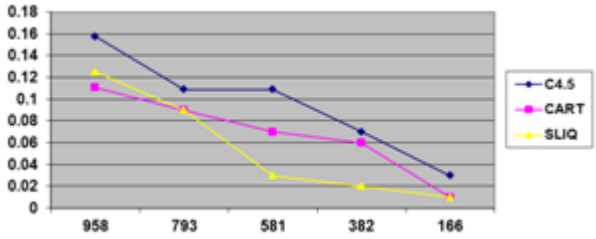

Fig 3: Shows Build time between ID3, C4.5 and SLIQ

As we can see in the graph in the initial state when the size of data set is large then ID3 consumed more time but C4.5 and SLIQ takes less time then ID3. But when the size of data set is reduces build time of the all the system is reduces and reflect more similar results.

VIII. COMPARATIVE RESULTS FOR MEMORY

By the study we can clearly see this all three algorithms are use quit different memory uses. It is found that with decreasing the size of data memory usage decreases constantly.

Memory between ID3, C4.5 and SLIQ

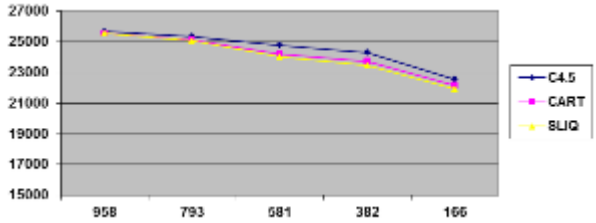

Fig 4: Shows Memory uses between ID3, C4.5 and SLIQ

IX. COMPARATIVE RESULTS FOR SEARCH TIME

Here below given graphs are for all algorithm selected and with boosting and without boosting.

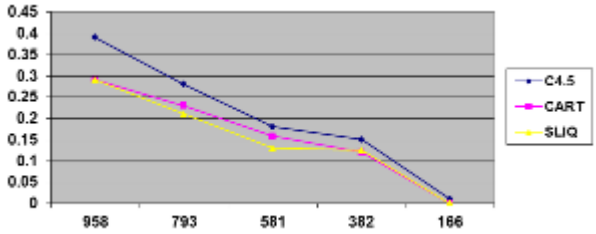

Fig 5: Shows Search time between ID3, C4.5 and SLIQ

Here we can clearly see search time continuous decreases as size of data set decreases. But ID3 takes more time then C4.5 and SLIQ. And if we compare SLIQ and C4.5 then we found that both algorithms consume similar time and most of time SLIQ performs better results.

Accuracy between ID3, C4.5 and SLIQ

The below given graph show the accuracy of the ID3, C4.5 and SLIQ algorithm. With numeric dataset SLIQ perform better than ID3 and C4.5 one of the reason is that SLIQ first sort the data and then build the model.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 6, June 2012

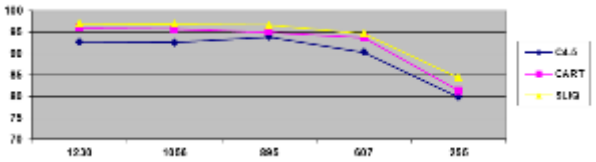

Fig 6: Shows Accuracy between ID3, C4.5 and SLIQ

Time between ID3, C4.5 and SLIQ

The below given graph show that with small data size all algorithms taking same time with increase in data size SLIQ is taking more time than ID3 and C4.5.

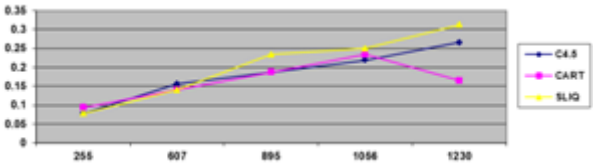

Fig 7: Shows Build time between ID3, C4.5 and SLIQ

X. COMPARATIVE RESULTS FOR SEARCH TIME

Here below given graphs are for all algorithm selected.

Time between ID3, C4.5 and SLIQ

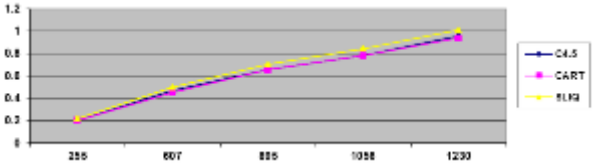

Fig 8: Shows Search time between ID3, C4.5 and SLIQ

XI. COMPARATIVE RESULTS FOR MEMORY USAGE

Memory between ID3, C4.5 and SLIQ

By the study we can clearly see this all three algorithms are use quit different memory uses. It is found that the memory usage of SLIQ is more than ID3 and C4.5.

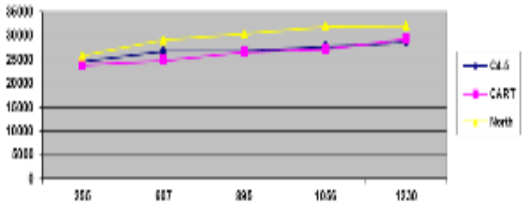

Fig 9: Shows Memory Usage between ID3, C4.5 and SLIQ

By the study we can clearly see this all three algorithms are use quit different memory uses. It is found that the memory usage of SLIQ is more than ID3 and C4.5.

XII. CONCLUSION

The main goal of our paper is to identify the correct and best suit decision tree, in our complete implementation of proposed work we found these results that are listed below.

1. Design and development of performance evaluation system based on decision tree algorithms.
2. Evaluating the performance of the algorithms on two different types of datasets and comparative analysis is performing on the basis of selected parameters.
3. We have analyzed the effect of data size on selected algorithm and found that on changing the size parameters changes with it.
4. We have made the following conclusions on the basis of results obtained.

REFERENCES

[1] Re Optimization of ID3 and C4.5 Decision Tree, Devashsish Thakur, Nisarga Markandaiah, Sharan Raj D Department of Computer Science PESIT Bangalore, Indiadevashish.thakur11@gmail.com

[2] Identifying Markov Blankets with Decision Tree Induction Prepublication Version, Lewis Frey Frey@vuse.vanderbilt.edu Department of Biomedical Informatics, Vanderbilt University, 2209 Garland Avenue; Department of Electrical Engineering and Computer Science, Box 1679 Station B, Nashville, TN 37232USADuglas Fisher DFisher@vuse.vanderbilt.edu Department of Electrical Engineering and Computer Science, Vanderbilt University, Box 1679 Station B, Nashville, TN 37235 USA

[3] IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 978-1-4577-0590-8/11/$26.00 ©2011 IEEE MIT, Anna University, Chennai. June 3-5, 2011

[4] IEEE transactions on geoscience and remote sensing, VOL. 47, NO. 12, DECEMBER 2009 A New Method for Correcting Scan SAR Scalloping Using Forests and Inter-SCAN Banding Employing Dynamic Filtering Masanobu Shimada, Senior Member, IEEE

[5] IEEE transactions on knowledge and data engineering, vol. 14 no 2, march/ April 2002, efficient c4.5, Salvatore ruggieri

[6] Change Detection in the Amazon Rainforest with Radiometric Rotation Technique RCEN Multi-spectral Case Study: Guarayos – Bolivia. 1University of Applied Sciences Fh – Eberswalde – Forestry Faculty Alfred – Moller – Str.1 D – 16225 Eberswalde – Brandenburg – Germany {hferrufino, tzawila}@fh-eberswalde.de 2INPE – Institutor National de Pressurises Espaciais Av. dos Astronauts, 1758 CEP.: 12.227-010 São José dos Campos – SP., Brazil jroberto@dsr.inpe.br

[7] DATA fusion study between polar metric sar, hyper spectral and lidar data for forest information David G. Goodenough1, 2, Hao Chen1, Andrew Dyk1, Geordie Hobart1,2, Ashlin Richardson1

[8] An Improved Algorithm of Decision Tree for Classifying Large Data Set Based on Rainforest Framework, Thangaparvathi.B, Assistant Professor, Dept. of Computer Science, and Thiagarajar College of Engg. - Madurai. btpcse@tce.edu, Anandhavalli.D, PG Student, Dept. of Computer Science, Thiagarajar College of Engg. - Madurai. anandhavallid@tce.edu

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 6, June 2012

[9] World Academy of Science, Engineering and Technology 35 2007, Text Mining Technique for Data Mining Application M. Govindarajan

[10] Text Mining Technique for Data Mining Application M.Govindarajan,

[11] A research on application of data mining in Cyber-Pursuit Yue He† , Jingsi Liu , Lirui Lin School of Business and Administration, Sichuan University, Chengdu 610064, P. R. China (Received 29 December 2009, Revised 8 March 2010, Accepted 28 May 2010).