IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-3 E-ISSN: 2347-2693

Survey Report on Various Decision Tree Classification Algorithm Using Weka Tool

P. Tomar[1*], A.K. Manjhvar[2]

[1*]Department of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior
[2]Department of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior

e-mail: tomarprateeksha09@gmail.com, Mob.: 7490490946

Available online at: www.ijcseonline.org

Received: 12/Feb/2017 Revised: 25/Feb/2017 Accepted: 14/Mar/2017 Published: 31/Mar/2017

Abstract **Data mining is the procedure of find or concentrates new patterns from extensive data sets including techniques from data and counterfeit consciousness. Arrangement and gauge are the procedures used to make out imperative data classes and conjecture plausible pattern .The Decision Tree is a critical scientific categorization technique in data mining grouping. It is generally utilized as a part of showcasing, reconnaissance, misrepresentation location, logical disclosure. As the established calculation of the decision tree ID3, C4.5, C5.0 calculations have the benefits of high group speed, solid learning capacity and straightforward development. In any case, these calculations are additionally unacceptable in viable application. Data mining is the method of find or focus new cases from immense instructive accumulations including methodologies from data and fake awareness. course of action and guess are the strategies used to make out basic data classes and gauge conceivable example .The Decision Tree is a basic logical order procedure in data mining portrayal. While using it to arrange, there does exists the issue of inclining to pick trademark which have more values, and neglecting properties which have less values. This paper gives focus on the diverse counts of Decision tree their trademark, troubles, ideal position and injury.. This work shows the strategy of WEKA examination of record converts, all around requested technique of** weka **use, decision of attributes to be mined and examination with Knowledge Extraction of Evolutionary Learning . I took database [1] and execute in weka programming. The complete of the paper shows the relationship among all kind of decision tree figurings by weka mechanical assembly.**

Keywords-Data Minning, Classification Algorithm, Decision Tree, J48, Random Forest, Random Tree, LMT, WEKA 3.7

I. INTRODUCTION

Data mining is a collection of techniques to glean information from data and turn into meaningful trends and rules to improve your understanding. The basic principles of data mining are to analyze the data from different direction, categorize it and finally to summarize it .Today we are living in digital world where data increasing day by day, to get any information from mountain of database is not only difficult but mind blogging also. To deal with this huge data we need data mining techniques. Data mining [2] define as the process of analysing, searching data in order to find contained, but prospective information. Data mining is used to find the hidden information prototype and relationship between the large data set which is very useful in decision creation. The advantages of data mining are analysis routinely, results of analysis is objective, accuracy of data is constant. Data mining also known as knowledge discovery in database (KDD), mainly data mining follows these steps; Data cleaning, Data integration, Data selection, data transformation, data mining, pattern evolution, knowledge evolution data reduction. Data mining having various numbers of techniques which have own speciality, such as clustering, data processing, pattern recognition, association, visualization etc.

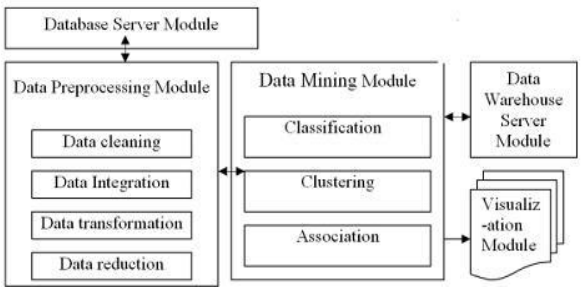

Fig.1. Data mining techniques

II. CLASSIFICATION

Order is conceivably the most ordinarily utilized data mining method. order [4] is the way toward decision an arrangement of models that delineate and separate data classes and ideas,

with the end goal of having the capacity to utilize the model to gauge the gathering whose mark is obscure. There are numerous calculations that can be utilized for characterization, for example, decision trees, neural systems, strategic relapse, and so on. In this work we are using decision tree calculation for characterization.

The Classification procedure includes following strides:

- Assemble preparing data set.
- Identify class quality and classes.
- Identify valuable quality for course of action.
- Learn a portrayal using preparing cases in Training set.
- Use the propagation to group the unidentified data tests

III. DECISION TREE

Decision trees are a method for in the interest of a grouping of guidelines that prompt a class or esteem. Decision Tree is flowchart like tree structure.

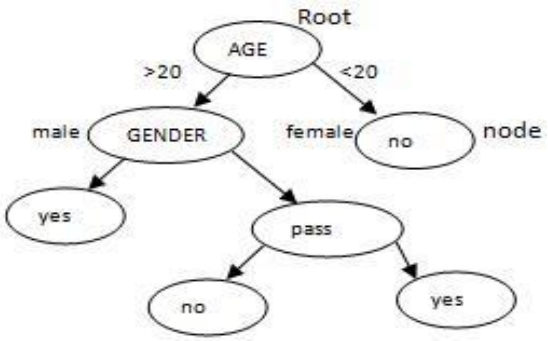

Fig.2. Decision tree

The decision tree comprises of three essentials, root hub, inner hub and leaf hub. Best most central is root hub. Leaf hub is the terminal central of the structure and the hubs in the middle of is known as the inner hub. Each inner hub means test on a property, each branch speaks to a result of the test, and each leaf hub holds a class mark. Different decision tree calculations are utilized as a part of order like ID3,AD Tree,REP,J48,FT Tree, LAD Tree, decision Trample, LMT ,unintentional woods ,coincidental tree etc. In this work following trees take for examination

AJ48- A prophetic machine-learning shape which decide the target value of a new sample based on distinctive trait values of the accessible data is J48 decision tree the distinctive trait signify by the interior hubs of a decision tree, the branches between the nodes tell us the conceivable qualities that these attributes can have in the investigational samples, while the terminal nodes disclose to us the last estimation of the destitute variable

LMT- A arrangement model with an unrelated supervised preparing calculation that joins strategic expectation and decision tree learning is logistic model tree (LMT). Logistic model treesuse a decision tree that has straight relapse models at its leaves to provide a section insightful direct relapse model.

Random Forest- chance forest is anensemble learning strategy for categorization, regression and other Tasks, that operate by build a multitude of decision trees at preparing time and outputtingthe class that is the mode of the order or mean prediction of the individual trees. Random backwoods adjust for decision trees' propensity for over fitting to their preparation set. Random forests are a method for averaging various profound decision trees, prepared on various parts of the same training set, with the goal of diminishing the Fluctuation. This come at the expense of a small increase and some loss of interpretability, yet for the most part enormously helps the arranging of the blade portrayal.

Random tree- - A random tree is a collectionof tree predictors that is called timberland. It can deal with both arrangement and relapse issues. The order works as follows: The random trees classifier takes the input highlight vector, orders it with each tree in the timberland, and outputs the class label .

IV. WEKA

WEKA might be a data mining programming delivered by the school about Weka to done New Zealand that mechanical get together. Data mining computations using the java tongue. Weka is a perspective in the verifiable background of the data mining What's more machine. Taking in Scrutinize people group, an outcome it might be the equitable toolbox that need grabbed such wide determination. Weka might be A winged creature. Purpose of Newzealand. WEKA might be a propelled trademark for Creating machine Taking in (ML) frameworks and there. Demand ought to true data mining issues. It will be an amassing about machine taking in computations to data mining. Errands. That WEKA undertaking focuses ought to give a careful aggregation about machine taking in counts and data pre. Changing instruments ought to experts. Those counts would clearly to a database. WEKA executes computations to data. Pre-handling, characterization, relapses, gathering Also participation rules; it in like manner fuses perception instruments.

. WEKA might be open hotspot programming issued under general populace allow those. Data record regularly used

Toward Weka might be On ARFF record for-tangle, which includes from asserting unprecedented labels, will demonstrate distinctive things in the data record first: quality names, quality sorts, Also quality qualities and the data. To working from guaranteeing WEKA we not convincing reason. UI of the client and gives numerous offices .The GUI Chooser comprises of four catches—one for each of the four noteworthy Weka applications.

The catches can be utilized to begin the following applications:

- **Explorer** : It is the primary interface in Weka. It has an arrangement of boards, each of which can be utilized to play out a specific errand .Once a dataset has been stacked, one of alternat boards in the Explorer can be utilized to perform promote examination.
- **Experimenter:** A situation for performing examinations and leading measurable tests between learning plans.
- **Knowledge Flow:** This condition underpins basically an indistinguishable capacities from the Explorer however with a drag – and drop interface. One preferred standpoint is that it underpins incremental learning.
- **Simple CLI:** Provides a basic command-line interface that allows coordinate execution of WEKA commands for working frameworks that don't give their own particular command line interface.

Fig. 3. WEKA tool front view

A. Execution in weka- It is a step by step process. First is data loading, Data can be loaded from various sources, including files, URLs and databases. WEKA has the capacity to read in .csv format. Firstly we take excel datasheet from real world, the first row contains the attribute names (separated by commas) followed by each data row with attribute values listed in the same order (also separated by commas), convert in .csv file format. Than go to the explore button on weka and save this .csv file. Once data loaded into WEKA, the data set automatically saved into ARFF format.

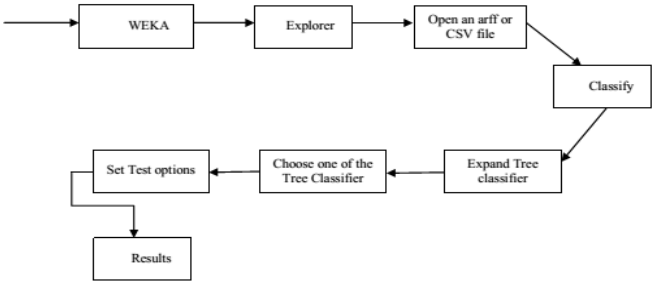

Fig. 4. Execution in weka tool

Choosing the Data from File, After data is stacked, WEKA will perceive the properties and amid the scan of the data will figure some essential insights on each trait. The list of perceived properties, while the beat boards demonstrate the names of the base connection (table) and the current working connection. Left board will Demonstrate the basic statistics on that trait Click on any property . For clear cut properties, the recurrence for each trait esteem is shown; while for persistent qualities we can acquire min, max, and mean, standard deviation, and so on. Prepare the Data to Be Mined, Selecting Attributes From test data document, each record is exclusively distinguished by property and using the Attribute channel in WEKA. In the "Channels" board, tap on the channel catch (to one side of the "Include" catch). This will demonstrate a popup window with rundown accessible channels. Look down the rundown and select weka. Filters. Attribute Filter" After setting channels, go to the grouping catch and tap on it .This will demonstrate a popup window with a rundown of order calculation, expand choice tree on this and select the tree which one u need to try.

- N – Total number of classified instances.
- True Positive (TP) – correctly predicted of positive classes .
- True Negative (TN) – correctly predicted of negative classes.
- True Negative (FP) – wrongly predicted as positive classes.
- True Negative (FN) – total wrongly predicted as negative classes.
- False Positive Rate (FPR) – negatives in correctly classified/total negatives.
- True Positive Rate(TPR) – positives correctly classified.
- Accuracy (A): It shows the proportion of the total number of instance predictions which are correctly predicted .

$$A = \frac{TP + TN}{N}$$

- **Receiver Operating Characteristic (ROC) Curve**: It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is plotted along the y axis and FPR is plotted along the x axis. Performance of each classifier represented as a point on the ROC curve. .
- **Precision(P):** It is a determine of exactness. It is the ration of the predicted positive cases that were correct to the total number of predicted positive cases.

$$P = \frac{TP}{TP + FP}$$

Recall(R): Recall determines completeness. It is the proportion of positive cases that were correctly recognized to the total number of positive cases. It is also known as sensitivity or true positive rate (TPR).

.

$$R = \frac{TP}{TP + FN}$$

- **F-Measure:** The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

F-measure=2*recall*precision / Precision+recall

V. COMPARISON OF VARIOUS ALGORITHMS

Decision tree	Tp rate	Ft rate	precision	recall	f-measure	Roc curv-e area	class	Time taken
J48	1	0	1	1	1	1	Y	0.14
	1	0	1	1	1	1	N	
Random forest	0.838	0.014	0.969	0.838	0.899	0.964	Y	0.07
	0.986	0.016	0.924	0.924	0.954	0.962	N	
Random tree	0.838	0.014	0.969	0.838	0.899	0.976	Y	0.01
	0.986	0.0162	0.924	0.986	0.954	0.971	N	
LMT	1	0.014	0.974	1	0.987	1	Y	6.9
	0.986	0	1	0.986	0.993	0.99	N	
Decision stump	1	0	1	1	1	1	Y	0.18
	1	0	1	1	1	1	N	

Table 1: Final statistic of decision tree

Random tree	0	37	0	la	YES	0
J48	0	0	74	lb	NO	0
LMT	0.0433	37	0	la	YES	0.0433
Random forest	0	1	73	lb	NO	0
Decision stump	0.2242	35	2	la	YES	0.2242

Table 2- Compression of weighted avg. for decision tree

VI. CONCLUSION

Comes to fruition exhibits that Decision stump grouping estimation takes minimum time to organize data yet gives less exactness. J48 have quite good accuracy with a little augment in time used for characterization. Most outrageous accuracy given by LMT, but time taken to make game plan model is significantly higher than various classifiers or we can state greatest in each one of the classifiers in the greater part of cases. Rest of the models also lies amidst the best and most exceedingly terrible ones .In this paper Decision Tree portrayal figuring's examining and side interest strategy to clear up the results.

The specific approaches for classification are depicted; we developed the WEKA method relies on upon picking the archive and picking credits to change over .csv file to level record and discussed parts of WEKA execution. Our work extends to utilize the execution of various dataset. Each decision tree order the data effectively and incorrectly instance. We can use these decision tree computations in remedial, keeping cash, securities trade and diverse region

REFERENCES

[1] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann publisher, Third editon -2001 , ISBN: ISBN: 978-0-12-381479-1.

[2] Swasti singhal and monika jena, "*a study on weka tool for data pre-processing, classification and clustering*", international journal of innovation technology and exploring enginnering, Vol.2, Issue.6, pp.250-253, 2013 .

[3] King, M., A., and Elder, J., F., "*Evaluation of Fourteen Desktop Data Mining Tools*", IEEE International Conference on Systems, mans, cybernetics, SMC, Newyork, oct 11th and 14th ,1998, ISBN:0-7803-4778-1.

[4] N. Landwehr , M. Corridor, and E. Forthcoming, ―Logistic model trees,‖ Mach. Learn., vol. 59, no. 1–2, pp. 161–205, 2005. .

[5] L. Breima , "*Random forests, Mach. Learn*", Springer, volume- 45, Issue no- 1, Page no-(5–32), Oct 2001.

[6] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "*Weka in Data Mining and Knowledge Discovery Handbook*", Springer, pp. 1305 –1314, 2005.

[7] Pallavi, Sunila Godara , "*A Comparative Performance Analysis of Clustering Algorithms*", International Journal of Engineering Research and Applications , Volume- 1, Issue no- 3, Page no-(441-445), ISSN: 2248-9622.

[8] E. Straight to the point, M. Corridor, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "*Weka,in Data Mining and Knowledge Discovery Handbook*", Springer, 2005, pp. 1305 – 1314.

Authors Profile

Ms. Prateeksha Tomar pursed Bachelor of Engineering from ITM universe in 2011. She is currently pursuing Master of Technology from Madhav Institute of Technology and Science,gwalior from branch cyber security.

Mr Amit Kumar Manjhvar pursed Bachelor of Engineering from branch Computer Engineering and Master of Technology in Software System. He is currently working as Assistant Professor in Department of Computer Science Engineering, Madhav Institute of Technology and Science, Gwalior.

$$A = \frac{TP + TN}{N}$$

