An Emperical Study on Decision Tree Classification Algorithms

Lakshmi.B.N[1] Dr. Indumathi.T.S[2] Dr. Nandini Ravi[3]

Abstract
The increasing data with technological advancement has put-forth a challenging situation for researchers to identify the most appropriate field to select, manage, make sense and use; reliable, novel, potentially useful, understandable, valid and ultimate data patterns. Data mining is one such field that can provide a solution to the problems faced to manage the large amounts of available data. Classification is a branch of data mining that is being extensively and efficiently used to manage the large amount of data through many levels of abstractions. There are many optimized methods of classification in data mining. Decision Tree is one of the most effective methods of classification to approach large amounts of data in comparison to other available methods. In this paper it is intended to survey a few Decision Tree Classification Algorithms like CART, ID3, C4.5, CHAID and MARS. The paper provides a brief description of the basics concepts in the section I, considers the reviews of other authors about the selected algorithms in section II, describes and compares the decision tree classification algorithms in III, based on all the reviews, comparison and analysis concludes the paper highlighting the pros and cons of each algorithm.

Keywords: Data Mining, Classification, Decision Trees, CART, ID3, C4.5 and CHAID.

I. INTRODUCTION

The technological advancement throughout the world is producing large amounts of data difficult to manage and maintain, thus challenging researchers to identify the most appropriate field to select, manage, make sense and use; reliable, novel, potentially useful, understandable, valid and ultimate data patterns.

Lakshmi.B.N*, Department of Computer Science and Engineering, VIAT, PG Research Center, VTU-RRC, Muddenahalli, Chikaballapura, Karnataka, India. Ph: 8147949104.*

Dr.Indumathi.T.S*, PG Co-ordinator, VIAT, PG Research Centre, VTU-RRC, Muddenahalli, Chikaballapura, Karnataka, India.*

Dr.Nandini Ravi*, MBBS, MD (Obs & gyn), Dhruva Nursing Home, Hoskote, Karnataka, India.*

Data from the real world has a lot of discrepancies and inconsistencies that are in need of maintenance and management. Data mining is one of the field in Information Communication Technology (ICT) that can provide a helping hand to manage, make sense and use these huge amounts of data by sorting out the discrepancies and inconsistencies. Data Mining is an important technique for managing data with which any of the technique may be integrated depending on the kind of data to be mined by extracting useful, logical and meaningful information and patterns from the huge data. The main aim of the technique to find information that can next be used to develop meaningful data and make accurate decisions and develop new systems. Data mining extracts the hidden predictive information from huge databases and is a powerful new technology with a great potential helping to focus on most important and required information in the data warehouses[1]. Data mining tools predict future trends and behaviours, thus allowing to make proactive, knowledge- driven decisions thus resolving time consuming question by scouring databases for hidden patterns, finding information that are predictive that may else be missed even by experts in some cases[1]. Data mining generally is considered as a process of data analysis from different perspectives and summarizing this data into useful information utilizable to raise revenue, cut costs or both. Here users are allowed to analyze data from various angles or dimensions, categorize it and summarize identified relationships. In the recent era data mining applications are available on all size systems and platforms. The most common techniques in data mining for identifying hidden patterns and information in data are classification and clustering analyses. Classification and clustering though seem similar, are different techniques. Classification routines in data mining use a variety of methods and the method used affects the way data is classified. There are several types of classification methods that include decision tree induction, Bayesian networks, k-nearest neighbour* classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques Classification technique is one of the data mining technique capable of

ISSN: 2278 – 7798

processing a wider variety and amount of data and is high in popularity [2]. Classification is a process of assigning an object to a specific class based on its similarity to examples of other objects previously seen called the training data. Classification comes with a degree of certainty i.e. It may be probability of object belonging to a class or some other measure of how closely an object may resemble other examples from the class. Decision Trees classification algorithms are one of the most well accepted classification method due it their high quality, efficiency, possibility of multi-level classification of huge data and capacity to handle continuous, numeric and noisy data. A decision tree is a flow chart like structure consisting of a single root node, internal nodes, branches and leaf nodes. Here each internal node is a selection of a attribute from a number of attribute alternatives and the first selection is the root node of decision tree and the internal nodes that follow up as branch nodes are the selection between a number of alternatives and each leaf node represents the result/decision. Each branch/internal node can have two or more branches depending upon the selected algorithm. WEKA (Waikato Environment for Knowledge Analysis) is a free open source software specialized for Data mining which is a popular suite of machine learning written in Java developed at the University of Waikato. WEKA consists of a set of visualization tools and algorithms for data analysis and predictive model having graphical user interfaces for easy functionality [3]. In this paper it is intended to survey a few Decision Tree Classification Algorithms like CART, ID3, C4.5 and CHAID. The paper provides a brief description of the basics concepts in the section I, considers the reviews of other authors about the selected algorithms in section II, describes and compares the decision tree classification algorithms in III, based on all the reviews, comparison and analysis concludes the paper.

II. LITERATURE SURVEY

Data mining is a process of analyzing data from different perspectives and gathering the knowledge from it. Various studies have been carried out that focus on data mining specially classification algorithms. One of the most efficient, easy to implement and effective classification method to mine the data from large database is decision tree construction method. Different decision tree algorithms applied for various datasets are considered in this section and explained. Sneha Soni in [4] have presented a well known datamining classification algorithm named CART which is one of the best known methods for machine learning and computer statistical representation. Here the paper shows results of multivariate dataset encompasing the simultaneous observation and analysis of more than one statistical variable and CART result is represented as a decision tree or by flow chart. Chaitrali S. Dangare et al [5] in their paper have analyzed prediction systems for heart disease using a variety of input attributes which account to 15 medical attributes to predict the likelihood of the patient getting a heart disease. The researchers use data mining classification techniques Decision Trees, Naive Bayes, and Neural Networks on Heart disease database and compare their performance based on accuracy of predicting heart disease. D.Lavanya et al in [6] their paper have studied a hybrid approach wherein with CART classifier feature selection and bagging techniques have been considered to evaluate the performance based on accuracy and time for various breast cancer datasets. Lior Rokach in [7] the paper present an updated survey of current methods for decision tree construction in a top-down manner. A unified algorithmic framework is suggested for presenting the decision tree classification algorithms and describes various splitting criteria and pruning methodologies. Elakia et al in [8] have designed a system to justify that various data mining classification algorithms can be used on educational databases to suggest career options for high school students and predict potentially violent behaviour among students by including additional parameters with academic details using a data mining tool called rapid miner. T. Santhanam et al in [9] have provided a study that used data mining modeling techniques to examine blood donor classification. The authors have used CART decision tree algorithm implemented in WEKA and analyzed standard UCI ML blood transfusion dataset. The accuracy of the algorithm was also analyzed. K.Sudhakar et al in [10] have used data mining techniques such as Decision Trees, Naive Bayes, Neural Networks, Associative classification and Genetic Algorithm to analyze heart disease database. Matthew N. Anyanwu et al in [11] have reviewed the serial implementations of decision tree algorithms, and identified commonly used ones. To evaluate performance of the commonly used serial decision tree algorithms the authors have used experimental analysis based on sample data records (Statlog data sets). Anju Rathee et al in [12] have explained and applied ID3, C4.5 and CART decision tree algorithms on students' data to predict their performance. Comparison and evaluation of all these algorithms based on the performance and results on already existing

ISSN: 2278 – 7798

datasets is done. Gilbert Ritschard et al in [13] have discussed the origin of tree methods and surveyed the earlier methods that led to CHAID decision tree classification algorithm. The authors have explained functioning of CHAID and briefed about the differences between the original method and the proposed extension method of CHAID. Smart drill in [14] has provided a basic introduction to CHAID decision tree classification algorithm. Leland Wilkinson et al in [15] discuss pitfalls in the use of classification and regression tree methods and specially highlight their suitability. S. Koyuncugil et al in [16] have presented a data mining model for detecting financial and operational risk indicators by CHAID decision tree algorithm. Belaid et al in [17] have proposed a technique for logical labelling of document images which makes use of decision tree based approach to learn and recognize the logical elements of a page. The authors employ a data mining method namely Improved CHi-squared Automatic Interaction Detection" (I-CHAID).

III. DECISION TREECLASSIFICATION ALGORITHMS

A. ID3 Algorithm:

ID3 (Iterative DiChotomiser 3) is a decision tree classification algorithm originally developed by J.Ross Quilan in 1975. ID3 is a supervised learning algorithm which builds a decision tree from a given data set, resulting in a tree used to classify future datasets. This algorithm is used in machine learning and natural language processing domains. Here in ID3 each and every node corresponds to a splitting attribute and every branch is a possible value of that attribute. In the decision tree ID3 constructs at every node a splitting attribute is selected that is most informative among other attributes not yet considered in the path from root node of constructed tree. The criterion of information gain is utilized by the ID3 algorithm to determine the goodness of a split. The splitting attribute is decided based on the attribute with greatest information gain and dataset is split for all various values of the considered attribute. The entropy and information gain considered by the ID3 algorithm are explained as follows:

Entropy is the measure of disorder or impurity in the dataset. Entropy can be generalized from boolean to discrete-valued target functions. Entropy comes from information theory. Higher the entropy more is the information content. When a node in a decision tree is used to partition the training sample data instances into smaller subsets the entropy changes typically.

Let S be a data set, let p be the fraction of positive valued training data samples and q be the fraction of negative training data samples then entropy is given by

$$Entropy(S) = -p\log_2(p) - q\log_2(q) \quad (1)$$

Information Gain: Information gain is a measure of change in entropy. It gives the importance of a considered attribute and is used to decide the ordering of attributes in nodes of a decision tree. Consider S to be a set of data samples, A an attribute, S_v the subset of S with A=v and Values(A) set of all possible values of A, Information Gain is given by

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$

(recall $|S|$ denotes the size of set S) (2)

In ID3, for each remaining attribute entropy is calculated and the least entropy attribute is used to split the set S. Classification improves with the entropy in other words higher the entropy the classification improves. The advantages and disadvantages of ID3 decision tree algorithm are it is robust to errors in the set of training data samples, training is reasonable fast, very fast classification of new data samples. Some of the disadvantages of ID3 decision tree algorithm are its difficult to extend to real-valued target functions, the algorithm needs to adapt to continuous attributes, there can be a issue of over-fitting of data samples from the set.

B. C4.5 Algorithm :

C4.5 decision tree algorithm is also proposed by Ross Quinlan in 1993 and is an extension of ID3 accounting for unavailable values, continuous attributes value ranges, pruning of decision trees, rule derivation and to overcome the limitations of ID3 algorithm. This algorithm introduces a number of extensions to the original ID3 algorithm.C4.5 handles both continuous and discrete attributes by creating a threshold and splitting the list into attribute values above, equal or below the threshold considered. Missing values in the training data set samples are handled by C4.5 algorithm by not using the gain and entropy calculations. Pruning trees once created by going back through the tree and removing branches that aren't helpful and replacing them with leaf nodes is performed by this algorithm and it also handles differing cost attributes. An open source implementation of C4.5 algorithm is called J48 in the WEKA data mining tool. C4.5 algorithm follows the same steps for building decision trees from a set of training data samples as ID3 by using the concept of information gain and entropy, wherein the splitting criteria is the normalized

ISSN: 2278 – 7798

information gain i.e. Entropy difference. A few base cases are considered by the C4.5 algorithm: (1) If all data samples considered in the list are of the same class, a leaf node is created for the decision tree thus choosing that class. (2) If no information gain is provided by any features a decision node is created higher up the tree using expected value of the class. (3) If a previously unseen class of data sample is encountered higher up the tree a decision node is created using a expected value. C4.5 algorithm follows a post pruning approach also called pessimistic pruning. C4.5 learns a mapping from attribute values to classes is learnt by C4.5 algorithm with which new unseen data samples can be classified. The C4.5 algorithm reduces the classification errors caused by specialization in the training data samples by pruning the completed decision tree to make it more general. C4.5 decision tree algorithm generates a small, very accurate and a simple decision tree. C4.5 considers a different measure known as Gain Ratio given by

$$GainRatio\ (A) = \frac{Gain\ (A)}{SplitInfo\ (A)} \quad (3)$$

Si= { S1, S2,.....Sn} = partitions of S based on values of attribute A. |Si| is number of cases in partition Si. |S| is total number of cases in S.

$$SplitInfo\ (A) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} \log_i \frac{|Si|}{|S|} \quad (4)$$

After finding the best split, the tree continues to be grown recursively. C4.5 decision tree algorithm is gives better classification than ID3 decision tree algorithm.

C. CART Algorithm :

Classification and Regression Tree is one of the classification method that constructs decision trees to classify data samples by knowing the number of classes in advance. The CART method was developed by Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone in 1984. CART algorithm a datamining classification algorithm is a well known and one of the best machine learning and computer statistical representation methods. This is a robust and binary recursive partitioning methodology wherein parent nodes are split into two child nodes exactly and repeats the process by treating every child node as a parent. CART algorithm presents its result in the form of a decision tree, diagram or flow chart. CART algorithm generates a branch in an attribute by considering a measure called GINI index. The attribute with the least or minimum GINI index after splitting is chosen. If S is a data sample and S1.....Sk a target attribute GINI index is given by

$$\mathbf{Gini(S)} = \sum_{i=1}^{K} \frac{|\mathbf{S_i}|}{|\mathbf{S}|}\left(\mathbf{1} - \frac{|\mathbf{S_i}|}{|\mathbf{S}|}\right) = \sum_{i \neq j} \frac{|\mathbf{S_i}| \times |\mathbf{S_j}|}{|\mathbf{S}|^2} \quad (5)$$

CART algorithm includes maximum tree construction, right tree size selection and new data classification using the constructed tree. This is a flexible method for binary tree construction. Classification Tree analysis is followed when data belongs to the class of the predicted outcome and Regression tree analysis is followed when the predicted outcome is considered a real number. When the target attribute value is ordered it is a regression tree and when the value is discrete is called classification tree. In CART algorithm the variable space is recursively split based on the impurity of the variables to determine the split to build the tree. CART algorithm offers few advantages like it is non-parametric, doesn't require in advance selection of variable , can handle outliers and adjust to time. Some disadvantages of CART algorithm are it produces decision that may be unstable and the splitting is performed by one variable only.

D. CHAID Algorithm :

Chi-square Automatic Interactive Detector Algorithm shortly known as CHAID Algorithm is a classification decision tree technique developed by Gordon V Kass in 1980 to evaluate complex interactions among predictors and display modeling results in tree diagrams which are easy to interpret. CHAID algorithm is one decision tree classification algorithm which works good with all kinds of categorical variables and continuous variables. This algorithm uses Chi-square splitting criteria for tree construction. This is one algorithm which can used for various tasks like prediction, detection of interaction between variables and classification. This can be considered an extension of Automatic Interaction Detection commonly known as AID and THeta Automatic Interaction Detection commonly known as THAID methods. This algorithm is one of the oldest classification tree methods which creates predictors by dividing continuous distributions into a number of categories with an equal number of observations. Selects the least significant category with respect to the dependent variable. P-value of the predictor variable with smallest adjustment is chosen as the split and this is the variable that will yield the most significant split and this is continued until no further splits are possible. There are many advantages CHAID algorithm provides such as

ISSN: 2278 – 7798

easy interpretation and produces a highly visual output. It produces reliable output but requires rather huge data sample sizes as it uses by default multiway splits and respondent groups can become very small when quiet small data sample sizes are used. This algorithm is also a non-parametric decision tree algorithm.

IV. CONCLUSION

The paper provides a survey on some of the efficient decision tree classification algorithms such as ID3, C4.5, CART and CHAID. This survey may inspire more researchers to use the following algorithms to solve many research problems put-forth by the available huge amounts of data for knowledge discovery. These algorithms are some of the most influential among the datamining classification decision tree algorithms. The algorithms are reviewed in literature survey and a description of each algorithm is provides and some of the advantages and disadvantages are discussed on the algorithms. These algorithms can be used to solve problematic topics in datamining classification research and development.

REFERENCES

[1]. An Introduction to Data Mining, Discovering hidden value in your data warehouse. http://www.thearling.com/ text/dmwhite/dmwhite.htm

[2]. Survey of Classification Techniques in Data Mining by Thair Nu Phyu,University of Computer Studies, Pakokku, Myanmar (Email: Thair54@gmail.com) in Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong with ISBN: 978-988-17012-2-0.

[3]. A Decision Tree for Weather Prediction by Elia Georgiana Petre, Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, e-mail: elia_petre @yahoo.com in Vol. LXI No. 1/2009, in page no: 77 - 82.

[4] Implementation of multivariate data set by CART algorithm by Sneha Soni in International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 455-459

[5] Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques by Chaitrali S. Dangare and Sulabha S. Apte in International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[6] Ensemble Decision Tree Classifier For Breast Cancer Data by D.Lavanya and Dr.K.Usha Rani in International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012

[7] Decision Trees by Lior Rokach and Oded Maimon, Department of Industrial Engineering, Tel-Aviv University in Chapter 9.

[8] Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students by Elakia, 2Gayathri, 3Aarthi, 4Naren J in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4649-4652 with ISSN: 0975-9646

[9] Application of CART Algorithm in Blood Donors Classification by T. Santhanam and Shyam Sundaram in Journal of Computer Science 6 (5): 548-552, 2010, ISSN 1549-3636, © 2010 Science Publications.

[10] Study of Heart Disease Prediction using Data Mining by K.Sudhakar and Dr. M. Manimekalai in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014, ISSN: 2277 128X. Available online at: www.ijarcsse.com

[11] Comparative Analysis of Serial Decision Tree Classification Algorithms by Matthew N. Anyanwu and Sajjan G. Shiva in International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (3).

[12] Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance by Anju Rathee and Robin prakash mathur in International Journal of Computers & Technology, Volume 4 No. 2, March-April, 2013, ISSN 2277-3061, Council for Innovative Research, www.cirworld.com.

[13] CHAID and earlier supervised tree methods by Gilbert Ritschard, Dept of Econometrics, University of Geneva, Switzerland, Juillet 2010.

[14] A Basic Introduction to CHAID by Smart Drill Data Mining, Data-driven Decision Support.

[15]Tree Structured Data Analysis: AID, CHAID and CART by Leland Wilkinson in Sun Valley, ID, Sawtooth/SYSTAT Joint Software Conference.

[16]Risk modeling by CHAID decision tree algorithm by A.S. Koyuncugil and N. Ozgulbas, ICCES, vol.11, no.2, pp.39-46, Copyright© 2009 ICCES.

[17]Improved CHAID Algorithm for Document Structure Modelling by Belaid.

ISSN: 2278 – 7798