CLASSIFICATION DECISION TREE ALGORITHMS IN DATA MINING

1st Miss.SakshiMahurpawar
sakshimahurpawar02@gmail.com
Student, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

2nd Mr. Lowlesh Yadav
lowlesh.yadav@gmail.com
Assistant Professor, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

3rd Mr. Neehal Jiwane
neehaljiwane@gmail.comAssistant Professor, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

ABSTRACT

As the computer technology and computer network technology are developing, the quantum of data in information assiduity is getting advanced and advanced. It's necessary to dissect this large quantum of data and excerpt useful knowledge from it. Process of rooting the useful knowledge from huge set of deficient, noisy, fuzzy and arbitrary data is called data mining. Decision tree bracket fashion is one of the most popular data mining ways. In decision tree divide and conquer fashion is used as introductory literacy strategy. A decision tree is a structure that includes a root knot, branches, and splint bumps. Each internal knot denotes a test on a trait, each branch denotes the outgrowth of a test, and each splint knot holds a class marker. The topmost knot in the tree is the root knot. This paper focus on the colorful algorithms of Decision tree (ID3, C4.5, wain), their characteristic, challenges, advantage and disadvantage.

Keyword: Decision Tree Learning, Algorithm, C4.5, Data Minning, ID3.

1.INTRODUCTION

In order to discover useful knowledge which is asked by the decision maker, the data miner applies data mining algorithms to the data attained from data collector. The sequestration issues coming with the data mining operations are twofold. If particular information can be directly observed in the data, sequestration of the original data proprietor (i.e., the data provider) will be compromised. On the other hand, equipping with the numerous important data mining ways, the data miner is suitable to find out colorful kinds of information underpinning the data. occasionally the data mining results reveals sensitive information about the data possessors. As the data miner gets the formerly modified data so then the ideal was to show the relative performance between formerly used bracket system and the new system introduced. As former studies shows that the ensemble ways give better results than the decision tree system therefore the asked result was inspired thru this concern.

2.METHODOLOGY:

A decision tree is a flowchart suchlike tree structure, where each internal knot represents a test on a trait, each branch represents an outgrowth of the test, class marker is represented by each splint knot (or terminal knot). Given a tuple X, the trait values of the tuple are tested against the decision tree. A path is traced from the root to a splint knot which holds the class vaticination for the tuple. It's easy to convert decision trees into bracket rules. Decision tree learning uses a decision tree as a prophetic model which maps compliances about an item to conclusions about the item's target value. It's one of the prophetic modelling approaches used in statistics, data mining and machine literacy.

Tree models where the target variable can take a finite set of values are called bracket trees, in this tree structure, leaves represent class markers and branches represent convergences of features that lead to those class markers. Decision tree can be constructed fairly fast compared to other styles of bracket. SQL statements can be constructed from tree that can be used to pierce databases efficiently. Decision tree classifiers gain analogous or better delicacy when compared with other bracket styles.

A number of data mining ways have formerly been done on educational data mining to ameliorate the performance of scholars like Retrogression, inheritable algorithm, kudos bracket, k means clustering, associate rules, vaticination etc. Data boobytrapping ways can be used in educational field to enhance our understanding of literacy process to concentrate on relating, rooting and assessing variables related to the literacy process of scholars. Bracket is one of the most constantly. The C4.5, ID3, wain decision tree is applied on the data of scholars to prognosticate their performance. These algorithms are explained below.

3.ID3 ALGORITHM

Iterative Dichotomies 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. It's serially enforced and grounded on Hunt ‟ s algorithm. The introductory idea of ID3 algorithm is to construct the decision tree by employing a top down, greedy hunt through the given sets to test each trait at every tree knot. In the decision tree system, information gain approach is generally used to determine suitable property for each knot of a generated decision tree. thus, we can elect the trait with the loftiest information gain (entropy reduction in the position of outside) as the test trait of current knot. In this way, the information demanded to classify the training sample subset attained from latterly on partitioning will be the lowest. So, the use of this property for partitioning the sample set contained in current knot will make the admixture degree of different types for all generated sample subsets reduced to a minimum. Hence, the use of an information proposition approach will effectively reduce the needed dividing number of object bracket.

4.C4.5 ALGORITHM

C4.5 is an algorithm used to induce a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's before ID3 algorithm. The decision trees generated by C4.5 can be used for bracket and for this reason C4.5 is frequently appertained to as a statistical classifier. As unyoking criteria, C4.5 algorithm uses information gain. Threshold is generated to handle nonstop values and also divide attributes with values above the threshold and values equal to or below the threshold. C4.5 algorithm can fluently handle missing values, as missing trait values aren't employed in gain computations by C4.6.

4.1.The algorithm C4.5 has following advantages:

Handling each trait with different cost. Handling training data set with missing trait valuesC4.5 allows trait values to be marked as „? ‟ For missing. Missing trait values aren't used in gain and entropy computations. Handling both nonstop and separate attributes to handle nonstop attributes, C4.5 creates a threshold and also splits the list into those whose trait value is above the threshold and those that are lower than or equal to it. Pruning trees after creation C4.5 goes back through the tree once it has been created, and attempts to remove branches that aren't demanded, by replacing them with splint bumps.

4.2.C4.5's tree construction algorithm differs in several felicitations from CART, for case:

Tests in wain are always double, but C4.5 allows two or further issues. wain uses Gini indicator to rank tests, whereas C4.5 uses information grounded criteria. wain prunes trees with a cost complexity model whose parameters are estimated bycrossvalidation, whereas C4.5 uses a single pass algorithm deduced from binomial confidence limits. wain looks for surrogate tests that compare the issues when the tested trait has an unknown value, on the other hand C4.5 apportions the case probabilistically among the issues.

5.CART ALGORITHM:

It stands for Bracket and Retrogression Trees. It was introduced by Bierman in 1984. It builds both groups and retrogression trees. The bracket tree construction by wain is grounded on double splitting of the attributes. wain also grounded on Hunt ‟ s algorithm and can be enforced serially. Gini indicator is used as blistering measure in opting the splitting trait. wain is different from other Hunt ‟ s grounded algorithm because it's also use for retrogression analysis with the help of the retrogression trees. The retrogression analysis point is used in vaticinating a dependent variable given a set of predictor variables over a given period of time. Wagons supports nonstop and nominal trait data and have total speed of processing.

Maximum Entropy (ME) and Decision Tree (DT) Experiments on PP Attachment

Experiment	Accuracy	Training Time	# of Features
ME Default	82.0%	10 min	4028
ME Tuned	83.7%	10 min	83875
DT Default	72.2%	1 min	
ME IFS	80.5%	30 hours	387
DT Binary	-	1 week +	-
Baseline	70.4%	-	-

Table.1.Maximum Entropy (ME) and Decision Tree (DT) Experiments on PP Attachment

6.DECISION TREE LEARNING SOFTWARE

Some software's are used for the analysis of data and some are used for generally used data sets for decision tree literacy are bandied below

6.1.WEKA:

WEKA (Waikato Environment for Knowledge Analysis) workbench is set of different data mining tools developed by machine literacy group at University of Waikato, New Zealand. For easy access to this functionality, it contains a collection of visualization tools and algorithms for data analysis and prophetic modeling together with graphical stoner interfaces. WEKA supported performances are windows, Linux and MAC operating systems and it provides colorful associations, bracket and clustering algorithms. All of WEKA's ways are rested on the supposition that the data is available as a single flat train or relation, where each data point is described by a fixed number of attributes (typically, numeric or nominal attributes). It also provides preprocessors like attributes selection algorithms and pollutants. WEKA provides J48.

6.2.GATree:

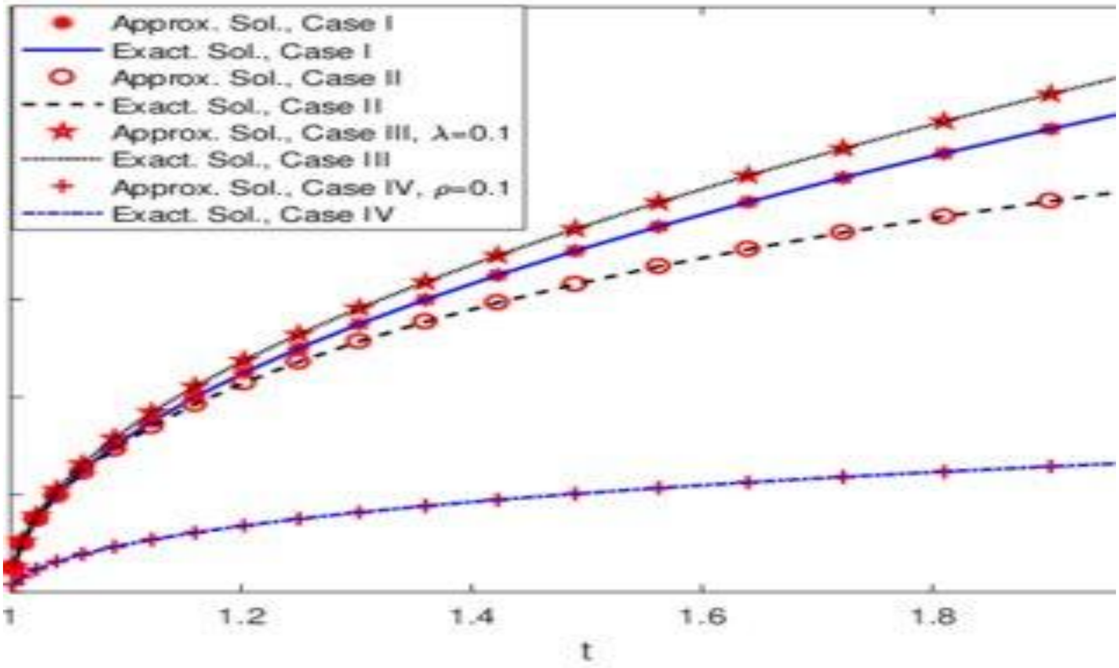
GATree (Genetically Evolved Decision Trees) use inheritable algorithms to directly evolve bracket decision trees. rather of using double strings, it adopts a natural representation of the problem by using double tree structure. On request to the authors, the evaluation interpretation of GATree is now available. To induce decision trees, we can set colorful parameters like generations, populations, crossover and mutation probabilityetc.

6.3.Aliced'ISoft:

Alice d'ISoft software for Data Mining by decision tree is an important and inviting tool that allows the creation of segmentation models. For the business stoner, this software makes it possible to explore data on line interactively and directly. And the evaluation interpretation of Alice d'ISoft is available on request to the authors.

6.4.See5/ C6.0:

See5/ C6.0 has been designed to dissect substantial databases containing thousands to millions of records and knockouts to hundreds of numeric, time, date, or nominal fields. It takes advantage of computers with over to eight cores in one or further CPUs (including Intel Hyperthreading) to speed up the analysis. See5/ C6.0 is easy to use and doesn't presume any special knowledge of Statistics Machine literacy. It's available for Windows Xp Vista/7/8 and Linux.



7.APPLICATIONS OF DECISION TREES IN DIFFERENT AREAS OF DATA MINING

Decision trees have a wide range of applications in different areas of data mining. They are popular due to their simplicity, interpretability, and effectiveness in solving a variety of problems. Here are some common applications in various domains:

7.1.Classification:

Decision trees are frequently used for classification tasks, where the goal is to categorize data into different classes. Examplesinclude:

Email Spam Detection: Deciding whether an incoming email is spam or not based on features like sender, subject, andcontent.

Medical Diagnosis: Diagnosing diseases or conditions based on patient symptoms and test results.

Customer Churn Prediction: Predicting whether customers are likely to leave a subscription service or stay based andcontent.

7.2.Regression:

Decision trees can be applied to regression problems where the goal is to predict a continuous numeric value. Examplesinclude:

Stock Price Prediction: Forecasting the future prices of stocks based on historical data and various indicators.

Real Estate Price Estimation: Predicting property prices based on attributes like location, size, and features.

Demand Forecasting: Estimating future demand for products in sales and supply chain management.

7.3.AnomalyDetection:

Decision trees can be used for identifying outliers or anomalies in datasets. This is useful in:

Credit Card Fraud Detection: Detecting unusual or fraudulent transactions by comparing them to typical spending patterns.

Network Intrusion Detection: Identifying unauthorized or suspicious activities in computer networks to enhance cybersecurity.

Quality Control in Manufacturing: Detecting defective products on the production line based on sensor data.

7.4.RecommendationSystems:

Decision trees can power recommendation engines that suggest products, content, or services to users. Examples include:

Movie Recommendations: Recommending movies to users based on their past viewing history and preferences.

Ecommerce Product Recommendations: Suggesting products to online shoppers based on their browsing and purchasehistory.

Music Recommendations: Recommending songs or playlists to users on music streaming platforms.

7.6.CustomerSegmentation:

Decision trees can be used for customer segmentation, dividing customers into different groups for targeted marketing or personalized services. Applications include:

Market Segmentation: Dividing customers into segments with similar characteristics for tailored marketing strategies.

Personalized Marketing: Delivering customized ads and content to users based on their demographics and behavior.

Healthcare Patient Clustering: Grouping patients with similar medical conditions and risk factors for specialized treatmentplans.

7.6.NaturalLanguageProcessing(NLP):

Decision trees are applied in NLP for tasks such as sentiment analysis, text classification, and information and content:
Sentiment Analysis: Classifying text data (e.g., social media posts or product reviews) as positive, negative, or neutral sentiment.
Spam Detection: Identifying spam or malicious text messages and comments.
Information Retrieval: Categorizing documents or web pages based on their content for search engines and contentindexing.

7.7. DataPreprocessing:

Decision trees are used for feature selection and data cleansing before applying more advanced machine learning Rather:
Feature Selection: Identifying the most relevant features in a dataset to improve model performance and reduce dimensionality.
Data Cleaning: Detecting and handling missing values, outliers, and noisy data to enhance data quality.

7.8.FinancialRiskAssessment:

In the financial sector, decision trees are employed for credit scoring, loan approval, and risk assessment. Examples include:
Credit Scoring: Evaluating an individual's or company's creditworthiness to make lending decisions.
Loan Default Prediction: Predicting the likelihood of borrowers defaulting on loans.

8.CONCLUSION

This paper studied colorful decision tree algorithms. The effectiveness of colorful decision tree algorithms can be anatomized grounded on their delicacy and time taken to decide the tree. This paper provides scholars and experimenter some introductory abecedarian information about decision tree algorithms, tools and operations.

REFERENCES

1. Anju Rathee, Robin prakash mathur, “Survey on Decision Tree classification algorithm for the evaluation of student performance” International Journal of Computers & Technology, Volume 4 No. 2, MarchApril, 2013, ISSN 22773061
2. S.Anupama Kumar and Dr. Vijayalakshmi M.N. (2011) “Efficiency of decision trees

in predicting student‟s academic performance", D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335343, 2011.

3. Devi Prasad bhukya and S. Ramachandram " Decision tree induction An Approach for data classification using AVL –Tree", International journal of computer and electrical engineering, Vol. 2, no. 4, August, 2010.
4. Jiawei Han and Micheline Kamber Data Mining: Concepts and Techniques, 2ndedition.
5. Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
6. Quinlan, J.R., C4.5 Programs For Machine Learning.Morgan Kaufmann Publishers, San Francisco, Ca, 1993.
7. Introdution To Data Mining By Tan, Steinbach, Kumar.
8. Mr. Brijain R Patel, Mr. Kushik K Rana, "ASurvey on Decision Tree Algorithm for Classification", © 2014 IJEDR, Volume 2, Issue 1.
9. Prof. Nilima Patil and Prof. Rekha Lathi(2012), Comparison of C6.0 & CART Classification algorithms using pruning technique.
10. Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
11. Neha Midha and Dr. Vikram Singh, "A Survey on Classification Techniques in Data Minng", IJCSMS (International Journal of Computer Science & Management Studies) Vol. 16, Issue 01, Publishing Month: July 2016.
12. Juan Pablo Gonzalez and U. Ozguner (2000). Lane detection using histogrambased segmentation and decision trees. Proc. of IEEE Intelligent Transportation Systems.
13. M. Chen, A. Zheng, J. Lloyd, M. Jordan and E. Brewer (2004). Failure diagnosis using decision trees. Proc. of the International Conference on Autonomic Computing.
14. Francesco Bonchi, Giannotti, G. Manco, C. Renso, M. Nanni, D. Pedreschi and S. Ruggieri (2001). Data mining for intelligent web caching. Proc. of International Conference on Information Technology: Coding and computing, 2001, pp. 599 603.
15. Ian H. Witten; Eibe Frank, Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition".
16. ELOMAA, T. (1996) Tools and Techniques for Decision Tree Learning.
17. R. Quinlan (2004). Data Mining Tools See5 and C6.0 Rulequest Research (1997).
18. S. K. Murthy, S. Salzberg, S. Kasif And R. Beigel (1993). OC1: Randomized induction of oblique decision trees. In Proc. Eleventh National Conference on Artificial Intelligence, Washington, DC, 1115th, July 1993. AAAI Press, pp. 322327.

19. Dipak V. Patil and R. S. Bichkar (2012). Issues in Optimization of Decision Tree Learning:
20. L. Hyafil and R. L. Rivest, " Constructing optimal binary decision trees is NP-complete ," Information Processing Letters, Vol. 5, No. 1, 15-17 (1976).
21. L.N. Kanal, "Problem-solving methods and search strategies for pattern recognition," IEEE Trans. Pattern Anal. Mach. lnteU. PAMI -1,193-201 (1979).
22. B. Kim and D. A. Lanclgrebe," Hierarchical decision tree classifiers in high-dimensional and large class data," Ph.D. Thesis and Technical Report TR-EE-90-47, School of EE, Purdue University (1990).
23. P.R. Krishnaiah, Ed. "On hierarchical classifier and interactive design, in Applications of Statistics," Amsterdam, The Netherlands: North-Holland, 1971, pp 301-321.
24. D.E. Knuth, "Optimum binary search trees," ACTA Informatica, vol. 1, 14-25 (1971).
25. D.E. Knuth, "The art of computer programming,l: fundamental algorithms," Addison-Wesley 1968.
26. A.V. Kulkarni and L. N. Kanal, "An optimization approach to hierarchical classifier design," Proc. 3rd Int. Joint Conf. on Pattern Recognition, San Diego, CA, 1976.
27. A.V. Kulkarni and L. N. Kanal, "Admissible search strategies for parametric and non-parametric hierarchical classifiers," Proc. 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, 1978.
28. A.V. Kulkarni, "On the mean accuracy of hierarchical classifiers," IEEE Trans. Comput. C-27, 771-776 (1978). A.V. Kalkarni, "Optimal and heuristic synthesis of hierarchical classifiers," Ph.D. dissertation, Univ. of Maryland, College Park, Comput. Sci. Tech. Rep. TR-469, 1976.
29. M.W. Kurzynski, "Decision rules for a hierarchical classifier," Pattern Recognition Lett. 1,305-310, (1983).
30. M.W. Kurzynski, "The optimal strategy of a tree classifier," Pattern Recognition 16, 81-87 (1983).