Grenze International Journal of Engineering and Technology, June Issue

Decision Tree Algorithm: A Survey

PunamKhandar[1], Chetana Thaokar[2] and Abhijeet Thakare[3]
[1-3]Shri Ramdeobaba College of Engineering and Management, Nagpur, India
Email: khandarp@rknec.edu, thaokarcb@rknec.edu, thakarear@rknec.edu

***Abstract*—Decision tree is one of the popular techniques in the research community to be used to develop decision making approaches. Over the period of almost four decades, a wide variety of applications of decision trees for providing good quality solutions in the different areas of life including technological pitfall, health hazards , social & geographical hitches, etc have been tried by several authors. The kind of work to be stated includes Object recognition using binary DT, Face detection, Diagnosing failures in large Internet sites, Prognostic decision making in breast cancers, Speech emotion recognition, Flood probability mapping, Automated Detection of Deceptive Language, Knowledge Discovery on Climate and Air Pollution and Fault Detection. Some pieces of works have even put commendable effort for development of DT and considerable evolution in the technique is observed. Some noteworthy works are effective decision tree design technique with the help of priori statistics, overcoming the shortcoming of ID3's using a new DT algorithm that united the ID3 and Association Function (AF) as well as an algorithm that combines principle of Taylor Formula with information entropy solution of ID3, alternating decision tree (ADTree) technique, etc. Here we are making a sincere effort to present some important applications of DT discussed by different authors through their valuable work as well as the variety of evolutionary aspects of DT put forward by them for the entire research community.**

***Index Terms*— Decision tree, algorithms, design techniques, knowledge discovery.**

I. INTRODUCTION

Decision trees are a powerful and popular mechanism for describing data. The rules derived from the decision trees can be easily understood and even implemented directly as a database query for retrieval purposes [8].It is a predictive model and used as a tool for decision support. Decision tree is important tool in data mining approaches. It's a tree-like model of decisions that recursively partitions the given space to subspaces that leads to decision making. Such algorithms contain conditional control statements in which each internal node performs a test on an attribute, each branch represents the result of the test, and each terminal node represents a class label. Thus the classification rule is confirmed as the paths from root to leaf. Decision trees are essential supervised paradigm of machine learning model, also popular one. It primly covers the areas like data mining, machine learning, and pattern recognition. Originally developed in decision theory and statistics, DTs were enhanced by researchers in Data science too for exploring large and complex bodies of data that identifies usefulpatterns.
The study contains 7 sections. Sect. 1 is introduction. Sect. 2 states various DT algorithms. Sect. 3 discusses the evolution of DT over the decades. Sect. 4 is about different applications of DT in various domains. Sect. 5 focuses on important methodologies. Sect. 6 discusses the comparative study results. Sect. 7 is about conclusion of the survey.

Grenze ID: 01.GIJET.8.2.47_1

II. VARIOUS DECISION TREE ALGORITHMS

In this section, we briefly discuss about some of the DT algorithms. Among many, some important algorithms of Decision Tree are CART (Classification And Regression Tree), ID3 (Iterative Dichotomiser3), C4.5 (successor of ID3) and Chi-square automatic interaction detection (CHAID).

CART uses partitioning procedure recursively based on an iterative search for best binary splits of data. Classifiers thus obtained are binary trees. The leaves of these binary trees determine the class labeling of the classes. Data resampling techniques used concurrently removes the biased classifier performance measures. [1] This paper uses tree-structured methodology and data resampling technique. CART is used to confirm the previous pattern analysis. The decision tree works on the principle of greedy algorithm which uses the approach of recursive top-down for structuring the tree . Quinlan in 1979 proposed the ID3 algorithm [5], which is the most ideally used algorithm in decision tree. But that algorithm is limited due to tendency to use attributes with many values. Many works have been put forward to address this limitation of ID3.Basically,ID3,C4.5andRandomForestdecisiontreealgorithmsaredifferentintheirfeaturesasbriefednext. ID3 and C4.5 are a single tree oriented algorithms. But the basic difference is that ID3 only work with Discrete or nominal data, whereas C4.5 work with both Discrete and Continuous data. [7] Random Forest on theother hands builds several trees from a single data set, and select the best decision among the forest of trees it generate. Besides, CHAID [32] is a Classification Tree technique which is basically preferred to evaluate complex interactions among predictors along with the simplicity of resultant tree diagram. The total modeling database is formulated as the "trunk" of the tree. CHAID helps in predictive analysis that uses target variable with multiple class categories. This algorithm then creates a first layer of "branches" by displaying values of the strongest predictor of the dependent variable. A criterion variable associated with the rest of variables configures the segments based on dependency of variables using a significantchi-square.

In one of the paper [4], author discusses fuzzy decision trees and their applications. Generally, the pruning of subtree doesn’t support for backtracking. Hence, author has proposed to sets the tree traversal pointers which supports for good backtracking mechanism. Thus creates a branch-bound-backtrack algorithm which has time complexity of O(log n) where n is number ofclasses.

III. LITERATURE SURVEY

Computational problem arises due to data set having large dimensions. If the data samples are taken as sequence of decision rules then they can be precisely classified in to distinctive classes. Decision tree classifier follows this principle. Hence it is better applicable to large dimensional data sets. A correct, flexible and efficient classification scheme can be achieved using decision tree classifier. [2] This paper gives effective decision tree design technique with the help of priori statistics. The technique is automated and it uses canonical transforms and Bays table look-up decision rules. The nodes of the tree are optimally designed using associated decision table. The author obtains decision tree that has 0.75 probability of correct classification where as theoretically 0.79 probability of correct classification can be achieved using full dimensional Bays classifier.About 2009 and so, the author has stated in his work, the shortcoming of ID3’s that occurs while selecting attributes with many values. Hence a new decision tree algorithm that united the ID3 and Association Function (AF) is developed [5].

To target the perfection over deficiency of ID3 algorithm, authors Huang Ming,NiuWenying and Liang Xuhas proposed a new improved classification algorithm that combines principle of Taylor Formula with information entropy solution of ID3 algorithm [10]. So entropy solution of ID3 is simplified and uses weights N to be assigned to entropy of ID3.. It thus handles the deficiency of ID3 algorithm as stated in the previous case. The score analysis is used to analyse the improved algorithm. It is evident from the experimentation results that simplified entropy with weight algorithm takes less time as compared to ID3 algorithm for building up the tree and accuracy is increased by3%.

optimized solutions for different kinds of applications. Here not much significance is given to the size and architecture of the network before the learning process. A decision tree structure is used to organize the perceptron units. The tree grows during the learning process (beginning with only one initial cell). With the growth of the tree, the leaf cells are divided in smaller and smaller parts of the initial training set.

IV. METHODOLOGIES USED

Many different techniques are proposed till now for solving the problems in different domains using variety of decision trees. We present here the prominent methodologies to unmasking the power of DT exploited over the long period of time in decision making, knowledge discovery, etc.

Offshore emergency conditions are changing in nature. It poses the high risk, time pressure and uncertainty to the individuals on board. This study[12] looks into how different emergency scenarios attributes impact people's quality of beginning route after training. An experimental study was administered during a virtual environment (VE) with 17 naive participants. Using a lecture based training (LBT) approach, the contestant were trained to assemblage during emergencies. Behaviourof the participants was observed during both training and testing scenarios, and human performance data were collected. The collected data were divided into training and testing data sets. A decision tree algorithm is provided with the training data. It then gives rise to a set of decision rules that describes the way people use different attributes of emergency scenarios to choose an egress route. It was observed that albeit the participants were given an equivalent training scenarios, on many occasions, they apparently interpreted the given knowledge differently. Thus, the perception of the attributes during a scenario can vary from individual to individual which leads to the variation in their decision trees within the testing data set. Outcomes of the decision trees were compared to the observed outcomes of the participants, thereby providing a basis to calculate the prediction accuracy of the trees during emergencies using a lecture based training (LBT) approach. The prediction accuracy of the decision trees is 95% on average. The trees predicted the performance of 11 participants with 100% accuracy (3 out of 3 testing scenarios), and 2 participants with 67% accuracy (2 out of 3 testingscenarios).
In one of the articles[13], author proposed the development of an automated system for analysis of urban traffic safety management and road conditions. C4.5 is found to be the best among the considered algorithms. In performing the analysis of the road conditions, traffic intensity was first calculated according to the generalformula

$$N= \sum^{n}_{i=1} (N_i - K_i) \qquad (1)$$

where N_i = traffic intensity of cars of given type, cars /hour
K_i = modular ratio for group of cars
n = number of cars types.

Further calculates the Carrying Capacity of the Road and the Traffic Load Factor. Throughput of the multi-lane road P, units/hour, is determined by the formula, taking into account the effect of the controlled intersection:

$$P = P_i * K_m * \alpha \qquad (2)$$

where P_i = lane capacity, units/hour
K_m = coefficient of multibandness;
α = coefficient considering the influence of signal-controlled junction.

The assessment of the complexity of the transport node on a five-point system is based on the calculation of the complexity index:

$$m= n_d + 3n_m + 5n_i \qquad (3)$$

where n_d = number of deviation points
n_m = number of merging points;
n_i = number of points of intersection.

Algorithm C4.5 creates the classifier within the sort of a choice tree. It enables to work with numeric attributes. Also empowers to build a tree from an incomplete training sample.
On the other hand, as stated previously, ID3 algorithm suffers from the limitation of inclining to choose attributes with many values. In the proposed work[5], a new decision tree algorithm combining ID3 and Association Function(AF) is presented. AF serves to overcome the ID3's deficiency of tending to take value with more attributes along with representing the relations between all elements and their attributes. AF algorithm is proposed as follows: Suppose A is an attribute of data set D, and C is the category attribute of D. The relation degree function between A and C are often expressed as follows:

$$AF(A) = \frac{\sum_{i=1}^{n} |x_{i1} - x_{i2}|}{n} \qquad (4)$$

Where x_{ij}(j = 1, 2 represents two sorts of cases) indicates that attribute A of D takes the i^{th} value and category attribute C takes the sample number of the j^{th} value, n is the number of values attribute A takes. Then, the normalization of relation degree function value is followed. Suppose that there are m attributes and every attribute relation degree function value are AF(1),AF(2),AF(m), respectively. Thus there is

(5)---------------------------------- $V(k) = \frac{AF(k)}{AF(1) + AF(2) + \cdots + AF(m)}$

$$Gain'(A) = (I(s_1, s_2, \cdots, s_m) - E(A)) \times V(A) \qquad (6)$$

Which $0 < k \leq m$. Then, equation (3) can be modified as
Gain'(A) are often used as a replacement criterion for attribute selection to construct decision tree consistent with the procedures of ID3 algorithm. Namely, decision tree are often constructed by selecting the attribute with the most important Gain'(A) value as test attribute. By this manner , the disadvantage of using ID3 are often overcome.

Decision trees can be classified as univariate (axis-parallel or orthogonal) and multivariate (oblique) decision trees. In univariate decision trees, among several features the best split is selected supported by some impurity criteria. Impurity criteria can be used as gini index or entropy can be used feature selection for splitting criteria, while as in multivariate decision trees, all or part of the features are evaluated for the best split. Mainly univariate decision tree can approximate any oblique decision boundary using a large number of stair-like decision boundaries. In this work [15], splitting plane is decided based on multivariate features in each non leaf node. Two bounded support vector machine is employed to obtain two clustering hyperplanes. Each of this hyperplane is closer to data points of 1 group and as far as possible from the data points of other group. Using these hyper planes, decision tree is created by splitting each non-leaf node. At each non-leaf node of a decision tree, hyperplane separates the data. Then the separation at the child nodes in the next level is simpler. These models showson an average 84%accuracy.

Computer-Aided Molecular Design (CAMD) is very popular technique for defining and designing reactions. Mostly the work done uses the structures of reactants or solvents. Quantitative Structure Property Relationship model needs to be developed which takes in to an account reactant structures and solvent effect on reaction rate. So computer aided system becomes important way forpredicting the constant rate. So diels- alder reaction is used for the study of solvents and reactors. In this study[11] Principal component analysis wasapplied to identify principal elements of the reactants and solvents. Decision tree is then generated taking principal elements and provides best possible initial population for genetic algorithm. Every time initial node was changed to get different initial population. Lastly genetic algorithm is applied for best fit to maximize r2 value.

V. COMPARATIVE STUDY

Various comparative studies were performed to compare the different DT algorithms. Also based on the precious work of many different researchers, we have stated the comparison of DT with some other competitive techniquestoo.Table I discusses about one of the study based on comparison of Naive Bayes and DT for intrusion detection system.

TABLE I. COMPARISION OF NAÏVE BAYES AND DT FOR INTRUSION DETECTION

	Winning Strategy	Decision Tree	Naïve Bayes
Normal	99.50%	99.50%	97.68%
Dos	97.10%	97.24%	96.65%
R2L	8.40%	0.52%	8.60%
U2R	13.2%	13.60%	11.84%
Probing	83.3%	77.92%	88.33%

It can be seen that decision trees give slightly better results. However, from computing point of view, naive Bayes can be constructed mostly faster than decision trees. It is concluded that learning and classifying with naive bayes is generally 7 times faster than that with decision trees. The reciprocity between decision trees and naive Bayes can be used to develop a meta classifier to provide good results in each category of attacks.

However , as per the work proposed by Sunakshi [14] for predicting gender from document, the author used Logistic regression, Multinomial NB (naive Bayes), SVC, Random Forest as classification algorithms. The result of their comparison is presented in the Table II.

It can be concluded that logistic regression is better than other classifiers and MultinomialNB also exhibits good outcome. In the paper [7] based on comparative analysis of ID3, C4.5 and random forest, the prediction accuracy is as shown in the Figure 1.

TABLE II. COMPARISION OF ACCURACY FOR VARIOUS CLASSIFIERS

Classifier	Accuracy(%)	Parameter used
Logistic Regression	64.00	penalty=12
Multinomial NB	63.55	alpha=0.5
SVC number	62.00	Default parameters
Random Forest	62.00	Default parameters

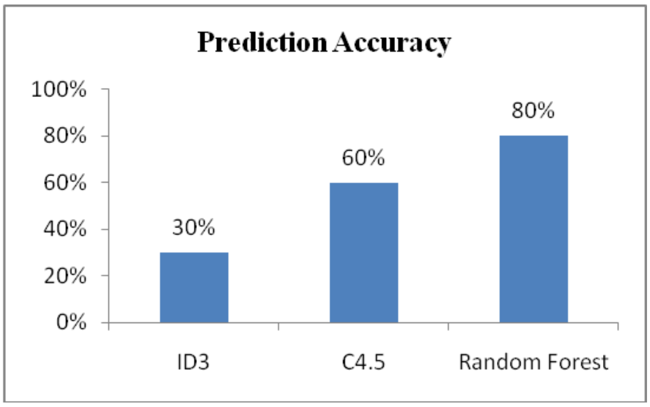

Figure 1. Prediction Accuracy Comparision

From the test results it is apparent that from the above three different forms of decision trees, Random Forest could generate better decision.

VI. CONCLUSION

While comparing Naive Bayes and DT, decision trees is found to give apparently better results. However, naive Bayes can be constructed faster than decision trees[14]. It is further found that the Decision Tree is more correct and more appropriate in estimating the location in indoor environment compared to other approaches like Nearest Neighbour, Neural Network[3]. In one of the work [9] to train surrogate DT from CNN and to build a visual analysis system for exploration, author proposes CNN2DT. It gives global and local interpretations of the decision process of CNNs. Users can reason about the misclassification for single images and find the discriminating capacity of various models. The limitation observed for this method is that the structures of decision trees are slightly unstable when using different amounts of train samples, which is mainly due to the rule extraction algorithm. Moreover, if a decision tree classifier is well designed, the resultant classification scheme is observed to be accurate, flexible, and computationallyefficient[2].

The drawback of Random Forest [14] is that it creates a lot of trees which can make the algorithm slower for prediction. But the prediction accuracy of RF is better than C4.5 and ID3 [7].The shortcoming of ID3's occurs while selecting attributes. It is often biased to select attributes having more values however, which are not necessarily the best attributes [5,6,10].

REFERENCES

[1] Kamil A. Grajski, Leo Breiman, Gonzalo Viana Di Prisco, Walter J. Freeman: Classification Of EEG Spatial Patterns With A Tree Structured Methodology: Cart. IEEE Transactions On Biomedical Engineering, Vol. Bme-33, No. 12, December1986.

[2] Peter Argentiero, Roland Chin, And Paul Beaude: An Automated Approach To The Design Of Decision Tree Classifiers. IEEE Transactions OnPattern Analysis And Machine Intelligence, Vol. Pami-4, No. 1, January1982.

[3] Osama Mohamed Badawy ,Moustafa Ahmed Bani Hasan: Decision Tree Approach To Estimate User Location In WlanBased On Location Fingerprinting, 24th National Radio Science Conference (Nrsc 2007), March 13-15,2007.

[4] Robin L. P. Chang AndTheodosiosPavlidis: Fuzzy Decision Tree Algorithms. IEEE Transactions On Systems, Man, And Cybernetics, Vol. Smc-7, No. 1,Pp 28-35, January1977.

[5] Chen Jin, Luo De-Lin, Mu Fen-Xiang: An Improved ID3 Decision Tree Algorithm. Proceedings Of 2009 4th International Conference On Computer Science & Education, IEEE, Pp 127-130, 978-1-4244-3521-0/09,2009.

[6] Chai Rui-Min AndWang Miao: More Efficient Classification Scheme For Id3. 2nd International Conference on Computer Engineering And Technology, IEEE, Pp.329-332,978-1-4244-6349-7/10,2010.

[7] ShijuSathyadevanAnd RemyaR. Nair: Comparative Analysis Of Decision Tree Algorithms: Id3, C4.5 And Random Forest. Springer India, L.C. Jain Et Al. (Eds.), Computational Intelligence InData Mining - Volume 1, Smart Innovation, Systems And Technologies 31,2015.
[8] Shuyuan Mary Ho, Jeffrey T. Hancock, Cheryl Booth, Xiuwen Liu, MuyeLiu, ShashankaS. Timmarajus, Mike Burmester, Real or Spiel? A Decision Tree Approach for Automated Detection of Deceptive Language-Action Cues.49th Hawaii International Conference On System Sciences, Ieee, 1530-1605/16,2016.
[9] Shichao Jia, Peiwen Lin, ZeyuLi, Jiawan Zhang, Shixia Liu: Visualizing Surrogate Decision Trees Of Convolutional Neural Networks. The Visualization Society Of Japan, J Vis Https://Doi.Org/10.1007/S12650-019-00607-Z,2019.
[10] Huang Ming,Niuwenying& Liang Xuhas:Application Of Information Theory To The Construction Of Efficient DecisionTrees.
[11] ShounakDatta, Vikrant A. Dev, Mario R. Eden :Hybrid Genetic Algorithm-Decision Tree Approach For Rate Constant Prediction Using Structures Of Reactants And Solvent For Diels-Alder Reaction, JournalComputersAnd Chemical Engineering(2017)Http://Dx.Doi.Org/10.1016/J.Compchemeng.2017.02.022
[12] Mashrura Musharraf, Jennifer Smith, Faisal Khan, Brian Veitch :Identifying Route Selection Strategies In Offshore Emergency Situations Using Decision Trees, Elseveir Journal Reliability Engineering And System Safety (2018) Https://Doi.Org/10.1016/J.Ress.2018.06.007
[13] Alla G. Kravets, Dmitry A. Skorobogatchenko, Natalia A. Salnikova, Nazim Y. Orudjev and, Olga V. Poplavskaya: The Traffic Safety Management System in Urban Conditions Based on the C4.5 Algorithm, 2018 Moscow Workshop on Electronic and Networking Technologies (MWENT), 978-1-5386-3498-1/18,2018IEEE.
[14] Nahla Ben Amor, Salem Benferhat, ZiedElouedi: Naive Bayes vs Decision Trees in Intrusion Detection Systems, 2004 ACM Symposium on AppliedComputing.
[15] M. A. Ganaie, M. Tanveer, P.N. Suganthan: Oblique Decision Tree Ensemble Via Twin Bounded Svm, International Journal Of Expert Systems With Applications,https://Doi.Org/10.1016/J.Eswa.2019.113072.