REVIEW ARTICLE OPEN ACCESS

A Survey Ondecision Tree Learning Algorithms for Knowledge Discovery

C. V. P. R. Prasad*, Dr. Bhanu Prakash Battula**
*Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
**Associate Professor, Vignan College, Andhra Pradesh, India.

Abstract
Theimmense volumes of data are populated into repositories from various applications. In order to find out desired information and knowledge from large datasets, the data mining techniques are very much helpful. Classification is one of the knowledge discovery techniques. In Classification, Decision trees are very popular in research community due to simplicity and easy comprehensibility. This paper presentsan updated review of recent developments in the field of decision trees.
Index Terms— Knowledge Discovery, Data Mining, Classification, Decision Trees.

I. INTRODUCTION

In Machine Learning community, and in Data Mining works, Classification has its own importance. Classification is an important part and the research application field in the data mining [1]. With ever-growing volumes of operational data, many organizations have started to apply data-mining techniques to mine their data for novel, valuable information that can be used to support their decision making [2]. Decision tree learning is one of the most widely used and practical methods for inductive inference [3].

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [4]. There are many different data mining functionalities. A brief definition of each of these functionalities is now presented. The definitions are directly collated from [5]. Data characterization is the summarization of the general characteristics or features of a target class of data. Data Discrimination, on the other hand, is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data.

Classification is an important application area for data mining. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model can be represented in various forms, such as classification rules, decision trees, mathematical formulae, or neural networks. Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

Outlier Analysis attempts to find outliers or anomalies in data. A detailed discussion of these various functionalities can be found in [5]. Even an overview of the representative algorithms developed for knowledge discovery is beyond the scope of this paper. The interested person is directed to the many books which amply cover this in detail [4], [5]. This paper presents an updated survey of various decision tree algorithms in machine learning. It also describes the applicability of the decision tree algorithm on real-world data.

II. THE CLASSIFICATION TASK

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science. Identifying the common ―core characteristics of a set of objects that are representative of their class is of enormous use in focusing the attention of a person or computer program. For example, to determine whether an animal is a zebra, people know to look for stripes rather than examine its tail or ears. Thus, stripes figure strongly in our *concept* (generalization) of zebras. Of course stripes alone are not sufficient to form a class description for zebras as tigers have them also, but they are certainly one of the important characteristics. The ability to perform classification and to be able to *learn* to classify gives peopleand computer programs the power to make decisions. The efficacy of these decisions is affected by performance on the classification task.

In machine learning, the classification task described above is commonly referred to as *supervised learning*. In supervised learning there is a specified set of classes, and example objects are labeled with the appropriate class (using the example above, the program is told what a zebra is and what is not). The goal is to generalize (form class descriptions) from the training objects that will enable novel objects to be identified as belonging to one of the classes. In contrast to supervise learning is *unsupervised learning*. In this case the program is not told which objects are zebras. Often the goal in unsupervised learning is to decide which objects should be grouped together—in other words, the learner forms the classes itself. Of course, the success of classification learning is heavily dependent on the quality of the data provided for training—a learner has only the input to learn from. If the data is inadequate or irrelevant then the concept descriptions will reflect this and misclassification will result when they are applied to new data. The popular approach of classification examples are C4.5 [6], CART [7] and REP [8].

III. EVALUATION CRITERIA'S FOR DECISION TREES

To assess theclassification results we count the number of true positive (TP),true negative (TN), false positive (FP) (actually negative, but classifiedas positive) and false negative (FN) (actually positive, butclassified as negative) examples. In this paper, we use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures.

Apart from these simple metrics, it is possible to encounter severalmore complex evaluation measures that have been used in different practical domains. One of the most popular techniques for the evaluation of classifiers is the Receiver Operating Characteristic (ROC) curve, which is a tool for visualizing, organizing and selecting classifiers based on their tradeoffs between benefits (true positives) and costs (false positives).

The most commonly used empirical measure, accuracydistinguish between the numbers of correct labels of different classes.The mathematical notation for calculation of accuracy is give below ineq (i),

$$ACC = \frac{TP + TN}{TP + FN + FP + FN} \quad \text{--------- (i)}$$

A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate. The Area under Curve (AUC) measure is computed ineq (ii) and eq (iii) ,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{--------- (ii)}$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad \text{--------- (iii)}$$

On the other hand, in several problems we are especially interestedin obtaining high performance on only one class. For example, in the diagnosis of a rare disease, one of the most important things is to know how reliable a positive diagnosis is. Another important measure used in decision tree is the tree size. The size of the tree is calculated by the depth of the tree and using the number of nodes and leaves.

IV. BENCHMARK DATASETS USED IN DECISION TREE LEARNING

Table 1 summarizes the benchmark datasetsused in some of the recent studies conducted on decision tree learning. The details of the datasets are given in table 1. For each data set, S.no., Dataset, name of the dataset, Instances, number of instances, Missing values, missing values in the dataset, Numeric attributes, number of numerical attributes, Nominal attributes, number of nominal attributes, Classes, number of classes are descried in the table for all the datasets.. The benchmark datasets used in popular experimental study are given in Table 1. The most popular machine learning publicly available datasets are available at Irvine [9].

S.no. Dataset	Instances	Missing values	Numeric attributes	Nominal attributes	Classes
1. Anneal	898	No	6	32	5
2. Anneal.ORIG	898	Yes	6	32	5
3. Arrhythmia	452	Yes	206	73	13
4. Audiology	226	Yes	0	69	24
5. Autos	205	Yes	15	10	6
6. Balance-scale	625	No	4	0	3
7. Breast-cancer	286	Yes	0	9	2
8. Breast-w	699	Yes	9	0	2
9. Colic-h	368	Yes	7	15	2
10. Colich.ORIG	368	Yes	7	15	2
11. Credit-a	690	Yes	6	9	2
12. Credit-g	1,000	No	7	13	2
13. Pima diabetes	768	No	8	0	2
14. Ecoli	336	No	7	0	8
15. Glass	214	No	9	0	6
16. Heart-c	303	Yes	6	7	2
17. Heart-h	294	Yes	6	7	2
18. Heart-statlog	270	No	13	0	2
19. Hepatitis	155	Yes	6	13	12
20. Hypothyroid	3,772	Yes	7	22	4
21. Ionosphere	351	No	34	0	2
22. Iris	150	No	4	0	3
23. Kr-vs-kp	3,196	No	0	36	2
24. Labor	57	Yes	8	8	2
25. Letter	20,000	No	16	0	26

26. Lympho	148	No	3	15	4
27. Mushroom	8,124	Yes	0	22	2
28. Optdigits	5,620	No	64	0	10
29. Pendigits	10,992	No	16	0	10
30. Primarytumor	339	Yes	0	17	21
31. Segment	2,310	No	19	0	7
32. Sick	3,772	Yes	7	22	2
33. Sonar	208	No	60	0	2
34. Soybean	683	Yes	0	35	19
35. Splice	3,190	No	0	61	3
36. Vehicle	846	No	18	0	4
37. Vote	435	Yes	0	16	2
38. Vowel	990	No	10	3	11
39. Waveform	5,000	No	41	0	3
40. Zoo	101	No	1	16	7

The complete details regarding all the datasets can be obtained from UCI Machine Learning Repository [9].

V. RECENT ADVANCES IN DECISION TREES

In Data mining, the problem of decision trees has also become an active area of research. In the literature survey of decision trees we may have many proposals on algorithmic, data-level and hybrid approaches. The recent advances in decision tree learning have been summarized as follows:

VasilePurdilaet al. [10] haveproposed a parallel decision tree learning algorithm expressed inMapReduce programming model that runs on Apache Hadoopplatform and has a very good scalability with dataset size.Dewan Md. Faridet al. [11] haveproposed a new learning algorithm for adaptive network intrusion detection using naive Bayesian classifier and decision tree, which performs balance detections and keeps false positives atacceptable level for different types of network attacks, and eliminates redundant attributes as well as contradictory examples from training data that make the detection model complex.

Koushal Kumar[12] haveconducted a study on artificial neural networks and then combined it with decision trees in order to fetch knowledge learnt in the training process. After successful training, knowledge is extracted from these trained neural networks using decision trees in the forms of IF THEN Rules which we can easily understand as compare to direct neural network outputs. Bruno Carneiro da Rocha et al. [13] haveevaluate the use of techniques of decision trees, in conjunction with the managementmodel CRISP-DM, to help in the prevention of bank fraud. Priyanka Saini et al. [14] have conducted a study on the evaluation of decision tree based ID3 algorithm and its implementation with student data example.Mohammad Khanbabaei et al. [15] have proposed a newhybrid classification model which isestablished based on a combination of clustering, feature selection, decision trees, and genetic algorithmtechniques. They used clustering and feature selection techniques to pre-process the input samples toconstruct the decision trees in the credit scoring model. The proposed hybrid model choices and combinesthe best decision trees based on the optimality criteria.

Richard Laishram Singh et al. [16] have made an attempt on building a word sense disambiguation system in Manipuri language. Decision tree model is used to identify conventional positional and context based features are suggested to capture the sense of the words, which have ambiguous and multiple senses.Dianhong Wang et al. [17] have proposed a novel roughest based multivariate decision trees (RSMDT) method in which, the positive region degree of condition attributeswith respect to decision attributes in rough set theory is usedfor selecting attributes in multivariate tests. And a newconcept of extended generalization of one equivalencerelation corresponding to another one is introduced andused for construction of multivariate tests.

Xinmeng Zhang et al. [18] have provided the definition ofsimilarity computation that usually used in data clusteringand apply it to the learning process of decision trees. They also proposed a novel splitting criteria which chooses the split with maximum similarity and the decision tree is calledmstree. At the same time, they suggest the pruning methodology for removing the unnecessary parts of the formed decision tree.José A. Martínez V.et al. [19] have proposed a methodology for themaintenance of trees based on data analysis. Starting from the informationcaptured in the field, they useddifferent techniques and models based onfuzzy logic and genetic algorithms, which keeps maintenance tasks onthe optimal time and place.

Ying Wanget al. [20] have proposed animprovedID3 algorithm and a novel classification attribute selection method based on Maclaurin-Priority Value First method. It adopts the foot changing formula and infinitesimal substitution to simplify the logarithms in ID3. For the errors generated in this process, an oppositeconstant is introduced to be multiplied by the simplifiedformulas for compensation. The idea of Priority Value First is introduced to solve the problems of value deviation. Ida Moghimipouret al. [21] have introducedthe three data mining software,namely SPSS-Clementine, RapidMinerand Weka. They also provided principal concepts of the decision tree method which are the mosteffective and widely used classification methods.

Dong-sheng Liuet al. [22] have proposed a modified decision tree algorithm for mobile user classification, which introducedgenetic algorithm to optimize the results of the decision tree algorithm.They also take the context information as a classificationattributes for the mobile user and they classify the context into public context and private context classes. Xiangxiang Zeng et al. [23] have conducted a study using survey data to builddecision tree models for forecasting the popularity of a number of Chinese colleges in each district. They first extract a feature called"popularity change ratio" from existing data and then use a simplified but efficient algorithm based on "gain ratio" for decision treeconstruction.The finalmodel is evaluated using common evaluation methods.

Win-TsungLo et al. [24] have proposed a design and implement a new parallelized decision tree algorithm on a CUDA (compute unified device architecture), which is a GPGPU solution provided by NVIDIA. In the proposed system, CPU is responsible for flow control while the GPU is responsible for computation.Mutasem Sh. Alkhasawnehet al. [25] have conducted a study on landslides dataset in a wide area of penang island, malaysia using four decision trees models Chi-square Automatic Interaction Detector (CHAID), Exhaustive CHAID, Classification and Regression Tree (CRT), and Quick-Unbiased-Efficient Statistical Tree (QUEST).

Suduan Chen et al. [26] have conducted a study for forecasting fraudulent and non-fraudulent financial statement happened between years 1998 to 2012, using improved decision tree models. MoulinathBanerjeeet al. [27] have investigatedthe problem of finding confidence sets for split pointsin decision trees (CART). Their main results establish the asymptotic distribution of the least squares estimators and some associated residual sum of squares statistics in a binary decision tree approximation to a smooth regression curve. Tarun Chopraet al. [28] have evaluated the performance of the proposed approach based on Stochastic Gradient Boosted Decision Trees based method on the DAMADICS benchmark problem. S.V.S. Ganga Devi[29] hasproposed a modified Fuzzy Decision Tree for Fruit data classification and the fuzzy classification rules are extracted.Pravin N. Chunarkar [30] haspresented an updated survey of current methods for constructing decision tree for classifying brain tumour dataset. The main focus is on solving the cancer classification problem using single decision tree classifiers (CART and Random algorithm) showing strengths and weaknesses of the proposed methodologies when compared to other popular classification methods.

Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

VI. CONCLUSION

In this paper, the state of the art methodologies to deal with decision tree has been reviewed. In recent years, several methodologies integrating solutions to enhance the induced classifiers are proposed. In brief we can say that this study summarizes the recent developments in the field of decision trees.

REFERENCES

[1.] Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

[2.] Huimin Zhao and Atish P. Sinha, An Efficient Algorithm for Generating Generalized Decision Forests, IEEE Transactions on Systems, Man, and Cybernetics —Part A : Systems and Humans, VOL. 35, NO. 5, Page no: 287-299, Septmember 2005.

[3.] M. Mitchell. Machine Learning. McGraw Hill, New York, 1997.

[4.] David Hand, Heikki Mannila, and Padhraic Smyth. Principles of Data Mining. MIT Press, August 2001.

[5.] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, April 2000.

[6.] J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann, 1993.

[7.] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[8.] J. Quinlan. Induction of decision trees, Machine Learning, vol. 1, pp. 81C106, 1986.

[9.] A. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci. edu/~mlearn/MLRepository.html

[10.] VasilePurdila, Ştefan-Gheorghe Pentiuc" MR-Tree - A Scalable MapReduce Algorithm for Building Decision Trees", Journal of Applied Computer Science & Mathematics, no. 16 (8) /2014, Suceava.

[11.] Dewan Md. Farid, NouriaHarbi, and Mohammad Zahidur Rahman" Combining naive bayes and decision tree for adaptive intrusion detect", International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.

[12.] Koushal Kumar"Knowledge Extraction from Trained Neural Networks", International Journal of Information & Network Security (IJINS) Vol.1, No.4, October 2012, pp. 282~293 ISSN: 2089-3299.

[13.] Bruno Carneiro da Rocha, Rafael Timóteo de Sousa Júnior" Identifying bank frauds using CRISP-DM and decision trees", International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010.

[14.] Priyanka Saini ,Sweta Rai, Ajit Kumar Jain" Decision Tree Algorithm Implementation Using Educational Data", International Journal of Computer-Aided technologies (IJCAx) Vol.1,No.4,April 2014.

[15.] Mohammad Khanbabaei and Mahmood Alborzi" THE USE OF GENETIC ALGORITHM, CLUSTERING AND FEATURE SELECTION TECHNIQUES IN CONSTRUCTION OF DECISION TREE MODELS FOR CREDIT SCORING", International Journal of Managing Information Technology (IJMIT) Vol.5, No.4, November 2013. DOI : 10.5121/ijmit.2013.5402

[16.] Richard Laishram Singh, Krishnendu Ghosh, Kishorjit Nongmei kapam, Sivaji Bandy opadhyay" A DECISION TREE BASED WORD SENSEDISAMBIGUATION SYSTEM IN MANIPURILANGUAGE", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014.

[17.] Dianhong Wang, Xingwen Liu, Liangxiao Jiang, Xiaoting Zhang, Yongguang Zhao" Rough Set Approach to Multivariate Decision Trees Inducing?", JOURNAL OF COMPUTERS, VOL. 7, NO. 4, APRIL 2012.

[18.] Xinmeng Zhang, Shengyi Jiang "A Splitting Criteria Based on Similarity in Decision Tree Learning", JOURNAL OF SOFTWARE, VOL. 7, NO. 8, AUGUST 2012.

[19.] José A. Martínez V., Fredy H. Martínez S." An intelligent decision support tool for management of tree maintenance in power systems" Electronic Vision - year 7 number 1 pp. 111 - 124 january -june of 2014.

[20.] Ying Wang, Xinguang Peng, and Jing Bian" Computer Crime Forensics Based on Improved Decision Tree Algorithm", JOURNAL OF NETWORKS, VOL. 9, NO. 4, APRIL 2014.

[21.] Ida Moghimipour, Malihe Ebrahimpour" Comparing Decision Tree Method Over Three Data Mining Software", International Journal of Statistics and Probability; Vol. 3, No. 3; 2014

[22.] Dong-sheng Liu, Shujiang Fan" A Modified Decision Tree Algorithm Based on Genetic Algorithm for Mobile User Classification Problem", Scientific World Journal, Volume 2014, Article ID 468324, 11 pages, http://dx.doi.org/10.1155/2014/468324, Hindawi Publishing Corporation.

[23.] Xiangxiang Zeng, Sisi Yuan, You Li, and Quan Zou" Decision Tree Classification Model for Popularity Forecast of Chinese Colleges" Journal of Applied Mathematics, Volume 2014, Article ID 675806, 7 pages,http://dx.doi.org/10.1155/2014/675806, Hindawi Publishing Corporation

[24.] Win-Tsung Lo, Yue-Shan Chang, Ruey-Kai Sheu, Chun-Chieh Chiu and Shyan-Ming Yuan," CUDT: A CUDA Based Decision Tree Algorithm", 偃e Scientific World Journal, Volume 2014, Article ID 745640, 12 pages, http://dx.doi.org/10.1155/2014/745640. Hindawi Publishing Corporation.

[25.] Mutasem Sh. Alkhasawneh, Umi Kalthum Ngah, Lea Tien Tay, Nor Ashidi Mat Isa, Mohammad Subhi Al-Batah" Modeling and Testing Landslide Hazard Using Decision Tree", Journal of Applied Mathematics, Volume 2014,ArticleID 929768, 9 pages, http://dx.doi.org/10.1155/2014/929768,HindawiPublishingCorporation

[26.] Suduan Chen, Yeong-Jia James Goo, Zone-De Shen" A Hybrid Approach of Stepwise Regression, Logistic Regression, Support Vector Machine, and Decision Tree for Forecasting Fraudulent Financial Statements", Scientific World Journal, Volume 2014, Article ID 968712, 9 pages, http://dx.doi.org/10.1155/2014/968712, Hindawi Publishing Corporation.

[27.] Moulinath Banerjee, Ian W. Mckeague," Confidence Sets For Split Points In Decision Trees", *The Annals of Statistics*, 2007, Vol. 35, No. 2, 543–574, DOI: 10.1214/0090536060000 01415, Institute of Mathematical Statistics, 2007.

[28.] Tarun Chopra, JayashriVajpai" Fault Diagnosis in Benchmark Process Control System Using Stochastic Gradient Boosted Decision Trees", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue-3, July 2011.

[29.] S.V.S. Ganga Devi" Fuzzy Rule Extraction For Fruit Data Classification", Compusoft, An international journal of advanced computer technology, 2 (12), December-2013 (Volume-II, Issue-XII).

[30.] Pravin N. Chunarkar" An Efficient Approach of Decision Tree for Classifying Brain Tumors", International Journal Of Engineering Sciences & Research Technology, [Chunarkar, 3(2): February, 2014].