IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

# A Survey on Different Decision Tree Methods for Solving Classification Issues

**V. Nirmala[1*], A. Nithya[2]**

[1,2]School of Computer Science, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore, Tamilnadu, India

[*]*Corresponding Author: nirmalavelusamy2018@gmail.com*

**Available online at: www.ijcseonline.org**

Accepted: 09/Jan/2019, Published: 31/Jan/2019

***Abstract*—**Data mining has effectively and tremendously enhanced the service in diverse areas, such as health care, business analysis, and social media. It is used to extract useful information from a huge volume of data by using various techniques like pre-processing, feature extraction, feature selection, and classification. One of the important research issues of the data mining and machine learning is a classification model. This model is to learn a classifier from a given trained dataset to predict the class of test dataset. Decision trees have become one of the most well-known classification methods for extracting classification rules from data, on account of their excellent learning capability. This especially focuses on to examine the various decision tree techniques to support data mining environments. The main objective of this survey is to study different decision tree methods used for detecting and solving classification issues. Finally, comparisons are made for different decision tree techniques in data mining.

***Keywords*—** Data mining, Decision tree, Classification, Knowledge extraction, Machine learning.

## I. INTRODUCTION

Classification technology based on IF-THEN rules have received a surge in attention given its potential applications, such as image processing, speech recognition and medical diagnosis [1, 2]. Classification methods are used to detect the casual signals in data while the process of finding casual relationships in data is a type of supervised learning when the outcome variable is fixed. One of the most widely used and a good example of classification methods in data mining [3] is decision trees [4].

Many variations of the decision tree algorithm were proposed in the literature [5]. They include Classification And Regression Tree (CART) [6], Iterative Dichotomizer 3 (ID3) [7], CHi-squared Automatic Interaction Detector (CHAID) [8] and Conditional Inference Trees [9]. A decision tree [10] is a classifier expressed as a recursive partition of the training instances. It is constructed in a top-down manner, in each iteration, the instance space is partitioned by choosing the best attribute to split them [11, 12]. This main contribution of this paper is to analyze the various decision tree techniques in data mining.

The rest of the paper is structured as follows: Section II analysis different decision tree techniques for data mining process. Section III presents the comparison of decision tree techniques in the literature. Section IV presents the conclusion of this survey.

## II. LITERATURE SURVEY

Different imputation approaches (Local Linear Interpolation and Global Statistic Approximation) was applied [13] to deal with complicated types of incomplete data in clinical environments. Significant features were discovered that were relevant to the severity of scoliosis with embedded technique. The scoliosis prediction models with multiple linear regression, k nearest neighbor, tree, support vector machine (SVM) and random forest algorithms were established and compared.

A new approach [14] was developed for securely constructing Random Decision Trees (RDTs) to both horizontally and vertically partitioned data sets. The proposed protocols were implemented and then, the computation and communication cost, and security was analyzed. The major contribution was to realize that RDTs were provided good security with very high efficiency.

A classifier [15] was presented for risk assessment in patients from congestive heart failure. This classifier was processed based on long term Heart Rate Variability (HRV) measures which were utilized for the individuation of high-risk conditions in Congestive Heart Failure (CHF). It was

 

estimated through the New York Heart Association classification (NYHA) scale. If the patients are suffering from CHF, then they were considered at higher risk. On the other hand, if the patient suffering from mild CHF then they were considered at lower risk. The method was utilized for developed the classifier is Classification And Regression Tree (CART).

A discernibility matrix [16] was defined and obtained on an ordinal dataset and complete feature subsets, respectively. A method of fusing complete monotonic decision trees was proposed that omits the procedure of selecting decision trees and determining the number of decision trees. A set of monotonic decision trees was obtained directly and automatically, and they will serve as base decision trees for constructing a decision forest. Although it was included a fewer number of trees, rank was still preserved that was ensured monotonically consistent rules. The proposed approach was decreased the number of base classifiers effectively and then classification model was simplified, and good classification performance was obtained simultaneously.

An enhanced method [17] was proposed based on Binary search on levels (BSOL) by employing a replication control technique for alleviating memory overhead of BSOL without performance penalty. The experimental results show that the memory requirement was decreased with the same speed performance as compared to BSOL. The approach was achieved better speed performance and memory requirement.

A focal-test-based spatial decision tree model and its learning algorithm [18] were discussed. Computational optimization was conducted and a refined algorithm was designed that selectively updates focal values. Both theoretical analysis and experimental evaluation show that the refined algorithm was more scalable than the baseline algorithm. A novel focal test technique with adaptive neighborhoods was designed for avoiding over-smoothing in wedge-shape areas.

A fusing principle image [19] was introduced for combining the base classifiers using the idea of maximal probability that was employed for enhancing the generalization ability of the monotonic classification system. The results show the effectiveness of this method from two viewpoints of the classification accuracy and the mean absolute error. An attribute reduction technique was presented for a monotonic classification task. A fusing method was designed for fusing monotonic decision trees induced by the rank entropy based monotonic decision tree approach.

A novel K-ary partition discretization approach [20] was investigated with no more than K-1 cut points by utilizing expected class number and Gaussian membership functions. For continuous-valued attributes, a novel K-ary crisp decision tree induction was proposed with a Gini index which combined the presented discretization approach. This method was enhanced classification accuracy and decreased Decision Tree scale, particularly in terms of tree depth.

A novel algorithm (VFC4.5) [21] was presented for building decision trees. It was proposed an adaptation of the way C4.5 finds the threshold of a continuous attribute. Instead of finding the threshold that was maximized gain ratio, this paper was proposed for simply decreasing the number of candidate cut points by using the arithmetic mean and median for enhancing a reported weakness of the C4.5 algorithm that it deals with continuous attributes.

The AFS theory [22] was utilized for determining the fuzzy membership functions automatically according to the raw data distribution; this was decreased the subjectivity of the formation of fuzzy numbers. An aggregate objective function was constructed for guiding the polymerization process of the fuzzy concepts. A novel Fuzzy Rule Extraction Algorithm was developed that only involves the value of one parameter, H. The threshold was optimized for balancing the classification accuracy with the tree size through a genetic algorithm.

The preeminence algorithm [23] was proposed becomes clearer in microarray gene expression data and a large amount of datasets that decreases the size of large training set concept was proposed. It was utilized a two-fold SVM and applied a data filter using a decision tree that scans the entire data acquire a small subset of data points. It was captured the pattern of the data and it was provided enough information for obtaining good performance.

The statistical query analysis in building decision trees [24] was discussed. The algorithms were presented to build private decision trees and ensemble under differential privacy. In the process of building a private tree, internal nodes were chosen using the noisy maximal vote. A budget allocation strategy was developed so that less noise will be added in larger depth to balance between the true counts and noise. For leaf nodes, the vote of every class was masked with Laplacian noise. The ensemble model was introduced for boosting the accuracy and decreasing the variance. The final classification outcome was set to be the label vote of multiple private trees.

A new model called Casual Decision Trees (CDTs) [25] to find and represent casual relationships in data. The graphical representation of the casual relationship among a set of predicator attributes and an outcome attribute was obtained by this CDT model. In addition to this, methods were designed to build a CDT which used a divide and conquer

 

strategy to build a normal decision tree. In order to select branching attributes of CDT, a criterion was employed. CDT was based on well established partial association tests and potential outcome model, ensuring the casual semantics of the tree.

## III. COMPARISON OF VARIOUS TECHNIQUES

This section provides an overview of the advantages and disadvantages in various decision tree techniques.

**Table 1. Comparison of different decision tree techniques**

| Ref No | Datasets | Merits | Demerits | Performance metrics |
|---|---|---|---|---|
| [13] | Clinical scoliosis dataset | Highly interpretable and viable to support the decision-making in clinical environments | A Small amount of clinical data is collected in this approach | Best Feature Selection = 3 for Random Forest algorithm<br>Mean absolute error = 3.632<br>Root mean square error = 5.309<br>Mean absolute percentage error = 0.231<br>Pearson correlation coefficient = 0.852 |
| [14] | UCI Machine Learning Repository: Mushroom, Nursery, Image Segmentation, and Car. | Strong privacy and less computation | Low accuracy | CAR dataset<br>Building time = 509.52s<br>Classification time = 0.091s<br>Accuracy = 71.6% |
| [15] | Holter databases | High sensitivity rate | Low accuracy rate and low specificity rate | Accuracy = 85.4%<br>Precision = 87.5%<br>Sensitivity = 93.3%<br>Specificity = 63.6%<br>Area under the curve = 78.5% |
| [16] | UCI datasets (Car, Bankruptyrisk, Adult, etc)<br>Real world dataset (Student Score) | Good classification performance and reduce the number of base classifiers | The computing cost of ordinal discernibility matrix and discernibility function might be expensive | Student Score dataset<br>Classification accuracy = 0.857±0.046<br>Mean absolute error = 0.143±0.046 |
| [17] | Several filter sets generated by ClassBench | Reduce the memory requirement and better speed performance | - | Filter set = IPC 100K<br>Memory accesses = 14<br>Memory requirement = 2.3 MB |
| [18] | Real world datasets | Improve the classification accuracy and reduce the computational time | Low precision, recall and F-measure | Scene = 2<br>Precision = 0.76<br>Recall = 0.75<br>F-measure = 0.75<br>Autocorrelation = 0.92 |
| [19] | Adult, Bankruptyrisk, Wine, Car, Student score, etc | Improve classification performance of monotonic decision trees. | Low performance of the fused learning system | Dataset = Adult<br>Accuracy = 0.774±0.001<br>Mean absolute error = 0.226±0.001 |
| [20] | Australian, Cancer, Ecoli, Pima, Spectf, Yeast, Transfusion, etc | Low complexity, reduce the decision tree scale | Low testing accuracy | Dataset = Australian<br>Testing accuracy in standard deviation = 5.36<br>Testing accuracy in average deviation = 75.22<br>Depth = 6.00<br>Nodes = 98.40 |
| [21] | Australian, Airlines, Banana, Bands, Flags, Marketing, Lung cancer, etc | Improve the accuracy results | Low sensitivity and specificity | Dataset = Airlines<br>Testing accuracy = 66,317<br>Sensitivity = 0.661 |

 

| | | | | |
|---|---|---|---|---|
| | | | | Specificity = 0.633<br>Training time = 357.79 |
| [22] | ALLAML, Average, iris, wine, wdbc, credit, etc | Improve the classification accuracy | Fault diagnosis problem is presented in this approach | Dataset = ALLAML<br>Classification accuracy = 0.9251<br>Tree size = 3.1<br>Time complexity = 1 |
| [23] | Leukemia, Duke breast cancer, Colon, WPBC, WDBC, etc | Speed up the training time of SVM, high classification accuracy | - | Dataset = Leukemia<br>Time = 0.21s |
| [24] | Adult and mushroom | Improve the accuracy and stableness | Still, improvement in classification accuracy is required | Dataset = Adult<br>Privacy budget = 0.01<br>MaxForest = 76.31(±1.96)<br>MaxTree = 18.88 (±1.89) |
| [25] | Adult and Ultra Short Stay Unit, K-R vs. K-P, Hypothyroid, etc | Fast and scalable | - | Dataset = K-R vs. K-P<br>Accuracy = 97.72%<br>Tree size = 57 |

## IV. CONCLUSION AND FUTURE SCOPE

In this article, a detailed comparative study on different decision tree techniques in data mining is presented. From this comparative analysis, it is clearly noticed that the decision tree techniques are widely used to classify the data with satisfied performance. Among those methods, CDTs based data classification has better performance. Even though, few limitations are addressed in CDTs based data classification Mantel-Haenszel test of CDTs has the problem with respect to stratified variables and its limitation to binary variables. Therefore, the future extension of this study could be focused on using different technique instead of the Mantel-Haenszel test that further increases the classification accuracy of CDTs.

## REFERENCES


[1] R. Hettiarachchi, J. F. Peters, "*Multi-manifold LLE learning in pattern recognition*", Pattern Recognition, Vol.**48**, Issue.**9**, pp.**2947-2960**, **2015**.

[2] A. Rosenfeld, H. Wechsler, "*Pattern recognition: Historical perspective and future directions*", International Journal of Imaging Systems and Technology, Vol.**11**, Issue.**2**, pp.**101-116**, **2000**.

[3] Marie Fernandes, "*Data Mining: A Comparative Study of its Various Techniques and its Process",* International Journal of Scientific Research in Computer Science and Engineering, Vol.**5**, Issue.**1**, pp.**19-23**, **2017**.

[4] P.N. Tan, "*Introduction to data mining*", Pearson Education India, **2007**.

[5] F. Saqib, A. Dutta, J. Plusquellic, P. Ortiz, M. S. Pattichis, "*Pipelined Decision Tree Classification Accelerator Implementation in FPGA (DT-CAIF)*", IEEE Trans. Computers, Vol.**64**, Issue.**1**, pp.**280-285**, **2015**.

[6] P. Breheny, "Classification and regression trees", **1984**.

[7] J.R. Quinlan, "Induction of decision trees", Machine learning,Vol.**1**,Issue.**1**, pp.**81-106**, **1986.**

[8] G.V. Kass, "*An exploratory technique for investigating large quantities of categorical data*", Applied statistics, pp.**119-127**, **1980**.

[9] T. Hothorn, K. Hornik, A. Zeileis, "*Unbiased recursive partitioning: A conditional inference framework*", Journal of Computational and Graphical statistics, Vol.**15**, Issue.**3**, pp.**651-674**, **2006**.

[10] Himanshi, Komal Kumar Bhatia, "*Prediction Model for Under-Graduate Student's Salary Uisng Data Mining Techniques",* International Journal of Scientific Research in Network Security and Communication, Vol.**6**, Issue.**2**, pp. **50-53**, **2018**.

[11] G.L. Agrawal, H. Gupta, "*Optimization of C4. 5 decision tree algorithm for data mining application*", International Journal of Emerging Technology and Advanced Engineering, Vol.**3**, Issue.**3**, pp.**341-345**, **2013**.

[12] N. Patel, D. Singh, "*An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor*", International Journal of Computer Applications, Vol.**111**,Issue.**10**, **2015**.

[13] L. Deng, Y. Hu, J.P.Y. Cheung, K.D.K. Luk, "*A data-driven decision support system for scoliosis prognosis",*IEEE Access, Vol.**5**, pp.**7874-7884**, **2017**.

[14] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, D. Lorenzi, "*A random decision tree framework for privacy-preserving data mining*", IEEE transactions on dependable and secure computing, Vol.**11**, Issue.**5**, pp.**399-411**, **2014**.

[15] P. Melillo, N. De Luca, M. Bracale, L. Pecchia, "*Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability*", IEEE journal of biomedical and health informatics, Vol.**17**, Issue.**3**, pp.**727-733**, **2013**.

[16] H. Xu, W. Wang, Y. Qian, "*Fusing complete monotonic decision trees*", IEEE Transactions on Knowledge and Data Engineering, Vol.**29**, Issue.**10**, pp.**2223-2235**, **2017**.

[17] Y.C. Cheng, P.C. Wang, "*Packet classification using dynamically generated decision trees*", IEEE Transactions on Computers, Vol.**64**, Issue.**2**, pp.**582-586**, **2015**.

[18] Z. Jiang, S. Shekhar, X. Zhou, J. Knight, J. Corcoran, "*Focal-test-based spatial decision tree learning*", IEEE Transactions on Knowledge and Data Engineering, Vol.**27**, Issue.**6**, pp.**1547-1559**, **2015**.

[19] Y. Qian, H. Xu, J. Liang, B. Liu, J. Wang, "*Fusing monotonic decision trees*", IEEE Transactions on Knowledge and Data Engineering, Vol.**27**, Issue.**10**, pp.**2717-2728**, **2015**.

[20] Y. Song, S. Yao, D. Yu, Y. Shen, Y. Hu, "*A New K-Ary Crisp Decision Tree Induction with Continuous Valued Attributes*", Chinese Journal of Electronics, Vol.**26**, Issue.**5**, pp.**999-1007**, **2017**.

[21] A. Cherfi, K. Nouira, A. Ferchichi, "*Very Fast C4. 5 Decision Tree Algorithms",* Applied Artificial Intelligence, Vol.**32**, Issue.**2**, pp.**119-137**, **2018**.


 

[22] Y. Cai, H. Zhang, Q. He, S. Sun, "*New classification technique: fuzzy oblique decision tree*", Transactions of the Institute of Measurement and Control, 0142331218774614, **2018**.
[23] P. Arumugam, P. Jose, "*Efficient Decision Tree Based Data Selection and Support Vector Machine Classification*", Materials Today: Proceedings, Vol.**5**, Issue.**1**, pp.**1679-1685**, **2018**.
[24] X. Liu, Q. Li, T. Li, D. Chen, "*Differentially private classification with decision tree ensemble*", Applied Soft Computing, Vol.**62**, pp.**807-816**, **2018**.
[25] J. Li, S. Ma, T. Le, L. Liu, J. Liu, "*Causal decision trees",* IEEE Transactions on Knowledge and Data Engineering, Vol.**29**, Issue.**2**, pp.**257-271**, **2017**.

## Authors Profile

**Ms Nirmala.V** pursued Master of Computer Application from Anna University in 2010 and M.Phil.(Computer Science) from Bharathiar university in 2012. She is currently pursuing Ph.d. and currently working as Assistant Professor in RathnavelSubramaniam College of Arts and Science, Sulur , Coimbatore. Her Main Research work focuses on Data Mining, Decision Trees, and Machine Learning. She has 6 years of Teaching Experience and 3 years of Research Experience.

**Dr Nithya A** pursued MSc., MPhil., Ph.D.. She is Research Guide and currently working as Associate Professor in RathnavelSubramaniam College of Arts and Science, Sulur ,Coimbatore.She has cleared State Level Eligibility Test(SET).She had produced 2 M.Phil Scholars and 4 Ph.D Scholars.She had published more than 20 Journals and presented more than 10 papers in International Conference. Her Main Research work focuses on Data Mining, Decision Trees, and Machine Learning. She has 13 years of Teaching Experience and 8 years of Research Experience.

 