IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Review of Decision Tree Based Classification Algorithms in Medical Data

Diksha[1*], D. Gupta[2]

[1,2]Department of Computer Science and Engineering, IKG Punjab Technical University, Jalandhar, India

Corresponding Author: dikshaaggarwal100@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.230234 | Available online at: www.ijcseonline.org

Accepted: 05/May/2019, Published: 31/May/2019

***Abstract*—** Classification problem in data mining is widely used to discover the potential information hidden in the data. Clinical, microarray data or image data related to medical field consists of high dimensions which pose difficulties for biomedical researchers in acquiring and analyzing data. Three principal challenges related to high dimensional data are Volume, Velocity and Variety. Various dimensionality reduction techniques are been used to remove irrelevant features to make the task easier and efficient. Also, using dimensionality techniques result in improved classification performance of the classifiers. This paper presents a review on the supervised machine learning algorithms for classification and prediction of various diseases. It also discusses various splitting criterion to determine the best attributes. Decision Tree algorithms are easy to understand and easy to use among all the classifiers.

***Keywords*—** Classification, CART, C4.5, C5.0, Decision tree , Dimensionality Reduction, ID3.

I. INTRODUCTION

Data mining is the process of extracting hidden and potentially useful information from data stored in databases or collected from various sources. This process is also known as Knowledge Discovery in Databases (KDD). Classification is one of the techniques that can be applied to extract information from the data. Classification is supervised machine learning in which the goal is to accurately predict the target class for unknown data item in the data. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Classification constructs the classification model by using training data set and the model can be tested by performing classification on the testing dataset. Decision trees are the most widely used classification method in data mining [1].

Decision tree is a supervised machine learning algorithm that is used to visualize the data in graphical form. It is one of the easiest and popular classification algorithms to understand and interpret. It implements a top-down greedy approach by partitioning the dataset recursively [2]. It can perform both classification and regression tasks on the dataset. Decision tree produces a set of rules that can be used to classify the data into the target class. Decision tree holds the advantage of being simple to understand [3], can handle both numerical and categorical data and requires little data preparation.

Rest of the paper is organized as follows, Section II contain the introduction of dimensionality reduction, Section III-VI describes the various decision tree algorithms, Section VII shows the implementation of these algorithms in classification of various diseases and Section VIII concludes research work with future directions.

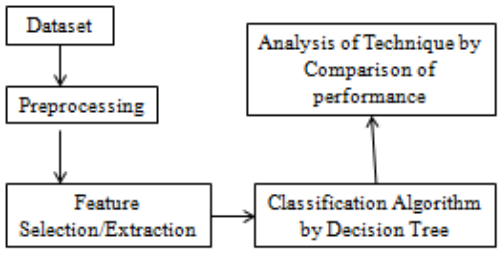

Figure 1: Classification with Decision Tree

II. DIMENSIONALITY REDUCTION

Dimensionality reduction, nowadays, is regarded as an important and compulsory pre-processing step before doing any analysis. It is a process through which a high dimensional data is converted into data having lesser number of dimensions that conveys the similar information as the original data [4]. Advancement in technologies and cost minimization of storing the data has lead to the accumulation of high dimensional data in all experiments. During accumulation of the data, generally irrelevant features are also aggregated along with the necessary and relevant features which does not play role in drawing any conclusion but increases the computational complexity and storage

space required to store the data. In order to handle the high dimensional data effectively, various techniques have been used to relieve the data analysts with the overhead of the irrelevant features. The dimensionality reduction techniques can be categorized into two categories according to the criteria they use to reduce the dimensionality. These are feature selection and feature extraction [5]. The former tends to find the subset of relevant features from the original features making them intact and the selected features do not lose their meaning. The latter technique extracts the relevant features as a combination of the original features. Features which define the maximum covariance in the original dataset are combined. Doing this may lead to the loss of the meaning of the actual attribute in the dataset.

Due the numerous advantages of the dimensionality reduction, this is included as a pre-processing step in the analysis of data in various fields. Some of the numerous fields are business analysis, medical science, image and video processing, gene analysis etc.

III. ID3 ALGORITHM

The basic algorithm developed for building decision tree is called **ID3** (Iterative Dichotomiser 3) by J. R. Quinlan first presented in 1975 in a book, Machine Learning, vol. 1, no. 1. Dichotomisation means dividing into two completely opposite things. They can work on nominal attributes but the numeric also needs to be transformed into nominal data. The ID3 follows the Occam's razor principle which roughly explains that more things should not be used than necessary. ID3 is the successor of Hunt's Concept Learning System (CLS) algorithm. It improves on CLS by adding a feature selection heuristic. It does not support backtracking since it is greedy search algorithm.

Splitting Criteria: The splitting criteria implemented by the ID3 algorithm is entropy or information gain. The attribute for which the entropy is minimum or the information gain is maximum is used to split the data. Entropy measures the expected gain in information. The entropy has the value zero when the distribution contains data items belonging to the single class and has the value one when the distribution of the classes is even.

Stopping Criteria: The decision tree stops when neither attribute is left to classify with nor instance is left to be classified. Pruning technique is used to avoid overfitting of the data. It removes the extra branches which do not participate in the classification task.

Nevertheless, ID3 also has some disadvantages, for example: (1) the algorithm is biased towards attributes with multiple values [3], but the attribution that has more values is not always optimal; (2) calculating information entropy with logarithmic algorithms is very time consuming [6-7]; and (3) the tree size is difficult to control [8], and the tree with a big size requires many long classification rules.

IV. CART ALGORITHM

The next decision tree algorithm very widely used is known as CART (Classification and Regression Tree) which is used for classification and regression predictive modelling tasks. The CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone. CART algorithm partitions the decision tree recursively where each input node is split into two child nodes, thus forming Binary tree. CART decision trees can also be seen as a set of rules or questions for each example to reach the leaf node. The model predicts the value of a target (or dependent variable) based on the values of several input (or independent variables).

Splitting Criterion: CART algorithm uses Ginni Index as the splitting criteria to decide the best attribute. A Gini score gives an idea of how good a split is by how mixed the classes are in the two groups created by the split. A perfect separation results in a Gini score of 0, whereas the worst case split that result in 50/50 classes.

Splitting Criterion: The stopping criteria is to have the minimum count on number of training instances required at each node for splitting to be nonstop. If the number of instances is less than the minimum count then the node is not splitted further and is considered as leaf node. If the value of minimum count is set to be extremely low (eg. Count of 1) then the tree tends to overfit the data and it will affect the performance on the test dataset.

The disadvantages that the CART algorithm presents that it makes decision based on only one variable and the second is that it can lead to unstable decision trees. If the training dataset changes then the decision changes causing tree complexity to increase or decrease [9].

V. C4.5 ALGORITHM

Another decision tree algorithm is the **C4.5** algorithm which is the successor of the ID3 algorithm, developed by Quinlan in 1993. The algorithm provides many improvements to the existing ID3 algorithm. These are (1) uses information gain ratio as the splitting criteria instead of information gain to reduce the bias; (2) can handle continous values along with the discrete values; (3) handling incomplete training data with missing values; (4) prune during the construction of trees to avoid over-fitting [10].

Splitting Criteria: C4.5 generates a decision tree where each node splits the classes based on the gain of information. The attribute with the highest normalized information gain is used

as the splitting criteria [3]. Once the splitting attribute is determined, the instance space is partitioned into several parts. Within each partition, if all training instances belong to one single class, the algorithm terminates. Otherwise, the splitting process will be recursively performed until the whole partition is assigned to the same class. Gain ratio takes into account the intrinsic information (number and size of the branches) of the split while choosing the attribute.

Stopping Criteria: When the all instances that covered by a specific branch are pure, OR, the number of instances fall below a certain threshold, the tree stops to grow.

VI. C5.0 ALGORITHM

The most recent advancement in the decision tree C4.5 algorithm is C5.0 algorithm. This classification algorithm is best suited for big data set. It is improved than C4.5 on the speed, memory and the efficiency. Also this algorithm is very efficient for handling missing values and the continous attributes. Faisal et al. [11] proved that C5.0 algorithm performs better than C4.5 algorithm in terms of memory, computational time and error rates. Decision trees can sometimes be quite difficult to understand. An important feature of C5.0 is its ability to generate classifiers called rulesets that consist of unordered collections of (relatively) simple if-then rules. C5.0 also offers the powerful boosting method to increase accuracy of classification.

Splitting Criteria: The splitting criteria of C5.0 algorithm is same as that of the C4.5 decision tree algorithm i.e. information gain ratio.

VII. RELATED WORK

In medical field, the decisions (classification, prediction) made must be reliable and accurate. Decision trees are such techniques that can provide reliable and effective decisions with high accuracy and a simple representation. Decision support systems are becoming an integral part of decision making in medical area providing a great help to the physicians. Decision trees are a suitable candidate for conceptual decision making models with automatic learning.

TABLE1: DECISION TREE ALGORITHMS USED IN CLASSIFICATION OF VARIOUS DISEASES

Sr. No.	Author	Algorithm		Dataset	Attributes (no. of instances / no. of features or attributes)	Remarks
		Dimensionality Reduction	Classification			
1.	Sarah A. Soliman (2005) [12]		Decision Tree C4.5 Algorithm	Thrombosis disease	407/58	Decision trees are easy to understand, easy to build and maps nicely to set of decision rules. C4.5 algorithm improved the accuracy but at the cost of large decision trees and memory usage.
2.	My Chau Tu (2009) [13]	-	Decision Tree C4.5 Algorithm and Naive Bayes	Heart Disease Database	920/13	C4.5 is an extension of ID3. It improves computing efficiency, deals with continuous values, handles attributes with missing values, avoids over fitting, and performs other functions.
3.	Mr. Chintan Shah (2013) [14]		Naive Bayes, Decision Tree and K- Nearest Neighbour Algorithm.	Wisconsin Breast Cancer data set	699/10	Random Forest algorithm of decision tree performed similar to the naive bayes algorithm but both of them performed better than the knn algorithm.
4.	Tzung-I Tang (2013) [15]		ID3, C4.5, CART, CHAID (Chi-Square Automatic Iteration Detection), and exhausted CHAID.	Coronary Heart Disease Dataset	1723/71	C4.5 algorithm has better accuracy than these algorithms with minimum number of leaves and depth second to CHAID.
5.	M Z F Nasution (2017) [16]	PCA algorithm	Decision Tree C4.5 Algorithm	Cervical cancer clinical dataset	858/36 reduced to 858/12	C4.5 can handle misclassification and overfitting. The accuracy improved with the decrease in number of features.

6.	S. Sathya (2017) [17]	-	Naive Bayes, K-Nearest Neighbour, Decision Tree, Baggings, Decision Tree Naive Bayes, J4.8 and Reduced Error Pruning Tree	Wisconsin breast cancer dataset	286/9	J48 (C4.5) resulted to be superior to all the other algorithms. AD Tree (Alternative decision tree) provides simple and compact rule set over ID3 and CART.
7.	Y. M. S. Al-Wesabi (2018) [18]	Sequential Forward Selection, Sequential Backward Selection	Gaussian Naive Bayes, K-nearest neighbour, Decision Tree, Logistic Regression and Support Vector Machine	Cervical Cancer Clinical dataset	858/36 reduced to 858/12	The decision tree outperformed all the algorithms used for classification. Use of dimensionality reduction techniques provides better results.
8.	Phonethep Douangnoulack (2018) [19]	PCA Algorithm	Decision Tree C4.5 (J48) Algorithm, Random Forest and Reduced Error Pruning Tree	Wisconsin Breast Cancer data set	699/11 reduced to 699/7	J48 has the ability to generate the simple rules. The Principal Component Analysis (PCA) is known as a lossless data reduction technique with good classification performance.
9.	M. I. Faisal (2018) [20]	-	Support Vector Machine, Decision Tree C4.5 Algorithm, Nearest Neighbour, and Naive Bayes, ensembles such as Random Forest and Majority Voting.	Lung Cancer dataset	32/57	Gradient boosted tree outperformed all the other classifiers even the ensemble techniques.
10.	Shweta Kharya (2018) [21]		Naive Bayes, Artificial Neural Network, Decision Tree C4.5 and C5.0 Algorithm	SEER dataset	433272/72 reduced to 202932/17	The decision tree algorithms (both C4.5 and C5.0) performed better than the other two techniques.

VIII. CONCLUSION AND FUTURE WORK

Decision trees have found a wide applicability in the medical field to predict various diseases. The various decision tree algorithms are able to classify and predict the disease to the appropriate target class. Decision tree being simple and easy to use have been used extensively by the analyst in various other fields along with medical data. Decision trees along with dimensionality reduction techniques are able to get more accurate results in less amount of time.

In the future, classification and prediction of cancer related data using advanced decision tree algorithms like C5.0 can be implemented and the result can be compared with previous decision tree algorithms like C4.5.

REFERENCES

[1] M. 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**.

[2] Himanshi, K.K. Bhatia, "*Prediction Model for UnderGraduate 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**.

[3] B. Hssina, A. Merbouha, H. Ezzikouri, M. Zrritali, "*A comparative study of decision tree ID3 and C4.5*", International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, pp.**13-19**, **2014.**

[4] M. Sabitha, M. Mayilvahanan, "A*pplication of dimensionality reduction techniques in real time dataset*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. **5**, Issue. **7**, pp.**2187-2189**, **2016**.

[5] R. Revathy, R. Lawrance, "*Comparative Analysis of C4.5 and C5.0 algorithms on crop pest data*", International Journal of Innovative Research in Computer and Communication Engineering, Vol. **5**, Issue. **1**, pp.**50-58**, **2017**.

[6] J. Liang, J. Shi, "*The information entropy, rough entropy and knowledge granulation in rough set theory*", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. **12,** pp.**37-46**, **2014**.

[7] T.P. Exarchos, M.G. Tsipouras, C.P. Exarchos, C. Papaloukas, D.I. Fotiadis, L.K. Michalis, "*A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythemic beat classification based on a set of rules obtained by a decision tree*", Artificial Intelligence in Medicine, Vol. **40**, pp.**187-200**, **2007**.

[8] J.R. Quinlan, "*Generating production rules from decision trees*", In Proceedings of the International Joint Conference on Artificial Intelligence, Milan, Italy, Vol. **1**, pp.**304-307**, **1987**.

[9] S. Singh, P. Gupta, "*Comparative study id3, cart and c4.5 decision tree algorithm: A Survey*", International Journal of Advanced Information Science and Technology (IJAIST), Vol. **27,** Issue. **27**, pp.**97-103**, **2014**.

[10] D. Ventura, T.R. Martinez, "*An empirical comparison of discretization methods*", In Proceedings of the Tenth International Symposium on Computer and Information Sciences, *pp.* **443-450**, **1995**.

[11] R. Revathy, R. Lawrance, "*Comparative analysis of c4.5 and c5.0 algorithms on crop pest data*", International Journal of Innovative Research in Computer and Communication Engineering, Vol. **5**, Issue. **1**, pp.**50-58**, **2017**.

[12] S.A. Soliman, A.S. Abbas, and A-B.M. Salem, "*Classification of thrombosis collagen diseases based on C4.5 algorithm*" , IEEE Seventh International Conference on Intelligent Computing and Information Systems, Vol. **3**, pp. **131-136**, **2015**.

[13] M.C. Tu, D. Shin, "*A comparative study of medical data classification methods based on decision tree and bagging*

algorithms", 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Washington, DC, USA, pp.**183-187**, **2009**.

[14] C. Shah, A.G. Jivani, "*Comparison of data mining classification algorithms for breast cancer prediction*", International Conference On Computing, Communication And Networking Technologies, Tiruchengode, Tamil Nadu, India, pp.**1-4**, **2013**.

[15] T-I. Tang, G. Zheng, Y. Huang, G. Shu, P. Wang, "*A comparative study of medical data classification methods based on decision tree and system reconstruction analysis*", Industrial Engineering & Management Systems (IEMS), Vol. **4**, Issue. **1**, pp.**102-108**, **2005**.

[16] M.Z.F. Nasution, O.S. Sitompul, M. Ramli, "*PCA based feature reduction to improve the accuracy of decision tree C4.5 classification*", 2nd International Conference on Computing and Applied Informatics Universitas Sumatera Utara (USU) Medan, Indonesia,pp.**1-6**, **2017**.

[17] S. Sathya, S. Joshi, S. Padmavathi, "*Classification of breast cancer dataset by different classification algorithms*", 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp.**1-4**, **2017**.

[18] Y.M.S. Al-Wesabi, A. Choudhury, D. Won, "*Classification of cervical cancer dataset*", Proceedings of the 2018 IISE Annual Conference, Loews Royal Pacific Resort, Orlando, Florida, pp.**1456-1461**, **2018**.

[19] P. Douangnoulack, V. Boonjing, "*Building minimal classification rules for breast cancer diagnosis*", 2018 10th International Conference on Knowledge and Smart Technology (KST), Thailand, pp.**278-281**, **2018**.

[20] M.I. Faisal, S. Bashir, Z.S. Khan, F.H. Khan, "*An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer*", 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), Karachi, Pakistan, pp.**1-4**, **2018**.

[21] S. Kharya, "*Using Data Mining Techniques for Diagnosis And Prognosis Of Cancer Disease*", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. **2**,Issue **2**, pp. **55-66**, **2012**.

Authors Profile

Ms. Diksha is the student of Masters of Technology in Computer Science and Engineering at IKG Pujab Technical University, Jalandhar, India. She has completed Bachelor in Technology in Computer Science and Engineering from Guru Nanak Dev University, Amritsar, India. Her main research work focuses on Data Mining and Big Data.

Mr. Dinesh Gupta, did his M.Tech in IT fromDepartment of CSE, GNDU, Amritsar, India. Currently he is pursuing his Ph.D in CSE from Desh Bhagat University. He has more than 8 years of experience in teaching. Currently he is working as Assistant Professor in department of CSE, IKG Punjab Technical Universiy, India. He has more than 8 publications in leading reasearch journal.

.

