Decision Tree for Classification and Regression: A State-of-the Art Review

Monalisa Jena and Satchidananda Dehuri
P.G. Department of Information and Communication Technology
Fakir Mohan University, Balasore, Odisha, India
E-mail: bmonalisa.26@gmail.com, satchi.lapa@gmail.com

Overview paper

Keywords: data mining, classification, regression, decision tree, prediction

Received: December 8, 2019

Classification and regression are defined under the umbrella of the prediction task of data mining. Discrete values are predicted using classification techniques, whereas regression techniques are most suitable for predicting continuous values. Analysts from different research areas like data mining, statistics, machine learning, pattern recognition, and big data analytics preferred decision trees over other classifiers as it is simple, effective, efficient, and its performance is competitive with others in a few cases. In this paper, we have extensively reviewed many popularly used state-of-the-art decision tree-based techniques for classification. Additionally, this work also reviews some of the decision tree based techniques for regression. We have presented a review of more than forty years of research that has been emphasized on the application of decision tree in both classification and regression. This review could be a potential resource for all the researchers who are keenly interested to apply the decision tree based classification/regression in their research work.

Povzetek: V preglednem članku je podana analiza raznovrstnih metod in tehnik odločitvenih in regresijskih dreves za namene rudarjanja podatkov.

1 Introduction

With the advancement of technologies, the process of data generation and collection is increasing at an exponential rate. The embedded sensors, IoTs, ubiquitous devices like scanners, bar code readers, and smartphones generate a huge amount of data at an exponential rate, which contributes to the expansion of data size and volume [1] [2] [3]. Intuitively, the valuable hidden knowledge and information in this huge amount of accumulated data could be the potential source to enhance the decision-making capability of the decision-makers of an organization or society [4] [5] [6]. Some of the classification techniques like decision tree (DT), support vector machine (SVM), and random forest [7] [8] have been proven to be effective models for extracting knowledge, that is valid, potential, novel, and finally useful. In a decision tree, interpretable rules together with the constraints can be extracted by the decision-maker without compromising the performance of the model [9] [10]. A decision tree is an acyclic graphical structure $G(V, E)$, where, $V \in \{V_1, V_2\}$ represents a finite, non-empty set of nodes; V_1 represents a set of leaf nodes containing the class values and V_2 is the set of intermediate nodes corresponding to one of the attributes. Similarly, the set of edges, E represents distinct attribute values. DT is one of the popularly used classifiers because of its intelligible nature that takes after the human thinking [11]. DT induction algorithms are preferred over other learning algorithms due to their flexibility, robustness to noise, the low computational cost for model construction, and the ability to handle redundant attributes. They are quite simple and easy to understand by human beings and their performance is comparable with others [12] in certain cases. Decision trees can handle both classification and regression tasks. In classification, a discrete value is predicted, whereas a continuous value is predicted through regression [13]. Decision trees are also competent in handling unseen samples having multiple class labels [14].

A sample DT is depicted in Figure 1. In this Figure, the DT is used to identify the types of contact lenses suitable for an individual having a set of features. It employs the *lenses* data set, one of the popular datasets collected from the University of California, Irvine (UCI) Machine Learning repository [15]. In Figure 1, the internal nodes and class labels are represented in the form of ovals and rectangles, respectively. Four different features such as tear production rate, age, spectacle prescription, astigmatic and three class labels namely hard, soft, and none are considered in this example. A path $\{v_1, v_2, ..., v_n\}$ drawn from v_1 to v_n represents the class prediction for a tuple, where v_1 is the root node, v_2 to v_{n-1} are the intermediate nodes, and v_n is the leaf node of that particular path. For example, for the tuple (age: presbyopic, astigmatic: yes, spectacle prescription: myope, tear production rate: reduced), the class label is "none". In this way, several rules can be extracted from the decision tree and using those rules, the class label of

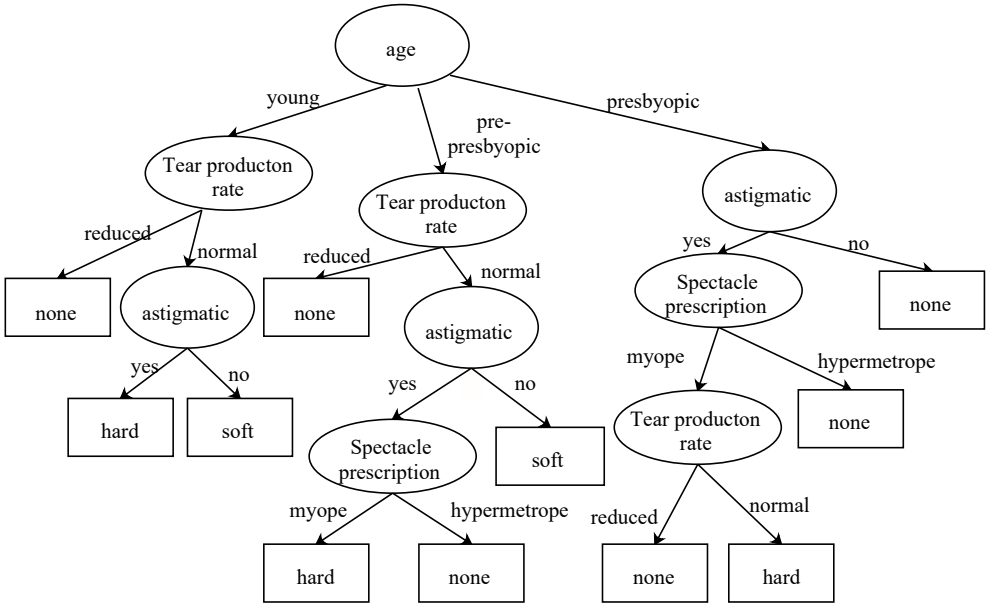

Figure 1: Decision tree to ascertain type of contact lenses to be used by a person.

an unseen sample can be predicted [16]. In Figure 1, the attribute "age" is taken as the root node. The root is selected using several attribute selection measures [17], and the splitting attribute is chosen at a particular node as per the well-defined splitting criterion. For example, the DT in Figure 1 is generated by applying entropy as the attribute selection measure. Hence, age becomes the root node as it is selected as the splitting attribute.

In the past few decades, a number of classification as well as regression tree algorithms have been proposed by several pioneers. Figure 2 gives an overall idea of the number of research papers published in the domain of DT for classification and regression from 1971 to date. An effort has been made to make an extensive review of the different classification and regression tree techniques which would be helpful for the beginners and enthusiastic researchers in this specific field of research. From Figure 2, it can be observed that over the years, research in this particular field has increased spectacularly because of the efficiency, performance, and effectiveness of DT in several application domains. Many researchers in the literature have presented reviews on the classification and regression tree algorithms. Some of the works have missed few parameters while some of them have provided just a brief overview, and some are outdated. Even though we intend to give a balanced discussion, some of the remarks certainly reflect the viewpoints of the authors.

Lim et al. [18], have compared twenty-two decision tree algorithms based on performance parameters like accuracy and computation speed. Classification accuracy is measured by the mean error rate and the mean rank of error rate. Along with the decision tree algorithms, they have also presented nine statistical and two neural network algorithms. They have experimented on these algorithms using thirty-two datasets, out of which fourteen are from real life domains, five are from the STATLOG project, two are synthetic, and the rest are from the UCI repository. Among the decision tree algorithms, QUEST with linear splits is found to have the highest accuracy, and logistic regression is the second best among the thirty three statistical algorithms. Podgorelec et al. [19], have limited their review work on decision trees specific to the field of medicine. They have presented alternatives to the few traditional induction approaches while emphasizing the existing and future applications of medicine. Perlich et al. [20], came up with a large scale comparison between two famous classification models of that time, tree induction and logistic regression. Based on the class membership probabilities, they had estimated classification accuracy and quality of rankings. They have observed that logistic regression performed well for smaller training sets while tree induction methods for comparatively larger datasets.

Rokach and Maimon [21] have presented an updated survey on the induction of decision tree algorithms of that time in a top-down manner. Besides, they suggested a unified algorithmic framework for presenting the decision tree induction algorithms and provided profound descriptions of the various pruning technologies and splitting criteria. They have observed that most of the algorithms fitted the framework with different stopping criteria and pruning methods. Barros et al. [22], have provided a review, which mainly focused on decision tree and evolutionary algorithms. They have presented a taxonomy that designs the decision tree components using evolutionary algorithms. They have also discussed various applications of evolutionary algorithms on decision tree induction in several domains. Loh [23] has presented a brief review of both classification and regression tree algorithms. In his paper, a brief comparison of the classification tree algorithms C4.5, RPART, QUEST, CRUISE, and GUIDE is presented using

prediction accuracy as the performance measure. The author has applied these algorithms on cars dataset for the 1993 model year, and GUIDE appeared to have the highest prediction accuracy. For comparing regression tree models, he has collected data from 654 children aged between 3 and 19 and applied those models on these datasets. GUIDE linear regression tree model was found to have higher prediction accuracy than piecewise constant models. For classification trees, prediction error was measured by misclassification cost, and in the case of regression trees, it was measured by the squared difference between predicted and actual values. Loh [24] again performed a comprehensive review on classification and regression tree algorithms which have been adopted in the last fifty years. In his paper, he focused on the majority of the algorithms that performed consistently well for a long period and for which software was widely available. The review work also provided the developments and key ideas supporting these algorithms. He has also presented a comparative analysis of the classification tree models and their partitions given by all the classification tree models using iris data from the UCI repository. A Similar procedure has been followed for regression tree models using baseball data from Statlib.

In contrast to others, we have presented a survey of all the classification and regression tree algorithms in a technical yet easy to understand manner. We have provided an extensive review of DT algorithms that have consistently better performance and stood the test of the time in the last forty years. We have also discussed the application details of the techniques in various domains under DT for classification as well as regression. This paper would be a potential resource for future researchers and enthusiast readers to get an overall idea about which algorithm works best in what domain, and accordingly, they can use as per their requirements. Additionally, we have given a comparative view of the algorithms, which highlights the suitability of each algorithm in the respective domains. It also presents the advantages and disadvantages of each algorithm in several domains.

The rest of the sections are set out as follows: In Section 2, the DT induction algorithm is discussed and the classification tree techniques are explained in detail. Section 3 highlights the application details of the techniques explained in Section 2. A comparative analysis of various classification tree algorithms is presented in Section 4. In Section 5, the DT algorithms used for regression are explained in a simplified manner. Sections 6 and 7 incorporate application details and comparative analysis of the techniques reviewed under DT for regression, respectively. Sections 3-7 will help the beginners in deciding which algorithms to choose for their experimental works as they will get a broad perspective of the different techniques. Finally, in Section 8, the paper is concluded along with future works.

2 DT as a classifier

Classification is a way of fitting objects to a category which best suits its characteristics. Classification is a two-step process in which the first one constructs the classifier by examining vividly the training set containing the attributes and their associated class labels [25]. This step is called the training or learning phase [26] [27]. The second step is known as the classification phase where the performance of the classifier is measured for the testing dataset. If performance is found up to the mark, then those rules are applied to unknown data tuples to predict their class labels [28]. Classification intends to distinguish the discrete category of a new sample by contemplating a training dataset. Mathematically, the classification process can be presented as a function as follows [29]:

$$C = f(X, \theta), C \in L \tag{1}$$

where X is the feature vector, C is the class label of the new sample, $f(.)$ is the classification function, θ is the parameter set of the classification function and L, the set of class labels. The main objective of DT is to represent maximum possible training datasets correctly with the better performance [30]. The decision tree is constructed by observing the behavior of the training tuples. This procedure is known as decision tree induction [31]. The attribute values for a tuple whose corresponding class label is unknown are tested against the decision tree. In that way, the path traced from the root node to the leaf is used to obtain several possible intelligent classification rules.

The entire DT induction procedure is explained in Algorithm 1. The algorithm starts with a training set and an empty tree. In step 1, a single node N is generated. If instances are of the same class, then a node is appended to the tree containing that class (step 2). Step 3 illustrates the terminating condition. It says when the attribute list becomes empty, the leaf node of the DT contains the class label whose occurrence is highest. This is called the majority-voting approach [32]. Otherwise, the attribute that splits the dataset into best partitions is perceived using attribute selection methods (step 4). Steps 5 to 22 focus on the splitting criterion and possible subsets as a result of partitioning tuples as per the splitting criterion. While inducing a decision tree, the splitting criterion is the most important factor to be considered [33]. The splitting criterion helps us in choosing the attribute that divides the tuples in the dataset into partitions containing individual classes by making a test at node N. Hence, the split-point or the splitting subsets are determined according to the decision tree induction algorithm [34].

The dataset is partitioned with the aim that each of the partitions should be as pure as possible. If all samples in a partition of the dataset are linked to the same class, the partition is said to be pure. For an attribute A, having x number of values $a_1, a_2, ..., a_x$, if it is discrete-valued, a set of branches are created corresponding to each attribute value. If it is continuous, then possible splits are in the

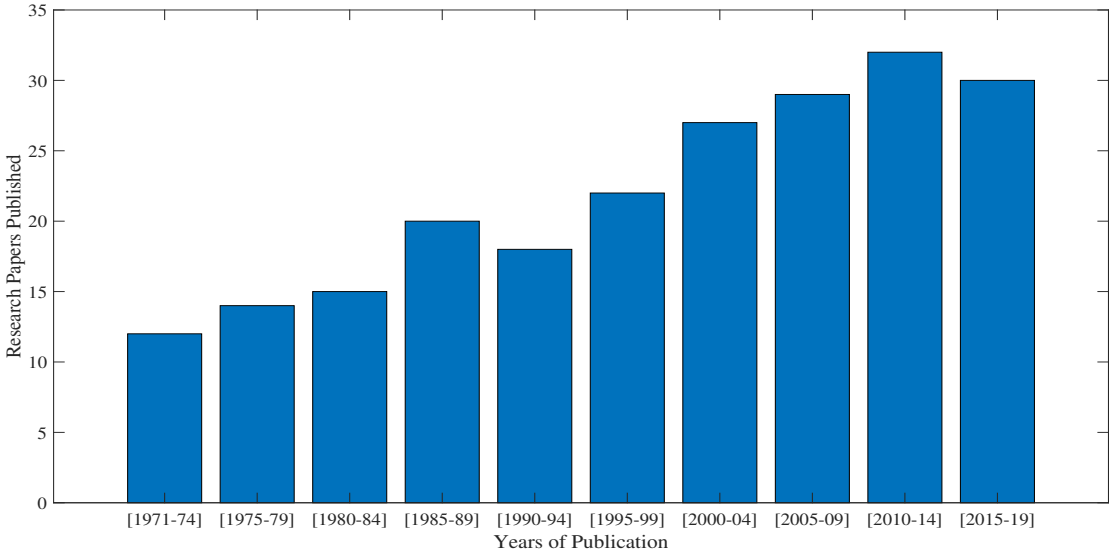

Figure 2: No. of research papers published over the years in the field of DT for Classification & Regression [Paper Sources: SCI, DBLP, Scopus indexed journals and conferences]

form of $a \leq c$ for one partition and $a > c$ for the other, where c is the splitting point. If the attribute is discrete and binary trees are to be generated only, then the splitting is in the form of $a \in S_a$, where S_a is the splitting subset for attribute A. The scenario is depicted in Figure 3. Several decision tree algorithms have been proposed for the classification task of data mining by many pioneers in the field of machine learning and data mining. In this paper, we have discussed some of the popularly used algorithms and their working patterns.

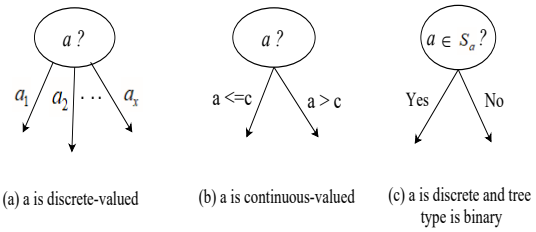

Figure 3: Different ways of partitioning tuples based on the splitting criterion

2.1 THeta automatic interaction detection (THAID)

This is the first published classification tree algorithm proposed by Messenger and Mandell [35]. It follows the concept of Automatic Interaction Detection(AID). AID is discussed in detail in section 5.1. THAID uses datasets having categorical variables. The node impurity at each node is measured based on the statistical distribution of the dependent variables over the mean. THAID searches the overall attributes of X extensively and finds a set S, which reduces the node impurity of its children, then splits a node for the split $\{X \in S\}$. If X is ordered, then $S \in (-\infty, c]$. Otherwise, $S \subseteq D(X)$, where $D(X)$ represents a set of possible values of X (the domain of X). This procedure is repeated for the tuples in each child node and splitting halts when the relative decrease in node impurity becomes less than a predetermined threshold. THAID merges similar categories of the predictors for tree pruning.

2.2 CHi-squared automatic interaction detection (CHAID)

This algorithm is the extension of the AID approach, where the chi-square statistical test has been employed for finding the best split for each independent variable [36]. It was initially developed for classification and later extended to the task of regression. This algorithm can be applied to the samples having categorical, ordered with missing values, and ordered without missing values. CHAID performs better for categorical values in comparison to mixed mode data values. If the variables are continuous, they are converted to categorical before applying the CHAID algorithm. If the sample consists of ordered variables with n distinct values, the chi-square test can be used to select the best suitable split out of $n - 1$ possible splits. If it consists of categorical variables and each variable is having n categories, it can have n splits. However, the number of splits can be lessened by applying Bonferroni adjusted significance tests. The significance test for each predictor follows a sequential cross-tabulation approach, whose steps are put forwarded in Algorithm 2. The major advantage of CHAID

Algorithm 1 Decision Tree Induction Method

Input: Dataset S with attribute vector X $\{x_1, x_2, ..., x_n\}$ and each tuple in T$\{t_1, t_2, ..., t_p\}$ has associated class labels L$\{l_1, l_2, ..., l_m\}$
Output: Decision Tree
Procedure: DT_Induction

Generate a node N
if every t_i in S $\in$ C **then**
  return N labeled with C
end if
if X = ϕ **then**
  return N as leaf with L=$max\{count(L_i)\}$ in S, $1 \leq i \leq m$
else
  find the best splitting criterion by applying attribute selection methods
end if
Label node N with attribute 'a' (the splitting attribute obtained from step 4).
if a is discrete and non-binary **then**
  X=X-a
end if
for each distinct outcome i $\in$ a **do**
  divide the dataset into S_i partitions
  if $S_i = \phi$ **then**
    connect the leaf having $max\{count(L_i)\}$ to N, $1 \leq i \leq m$.
  else
    link the node returned by DT_Induction(S_i, X) to N.
  end if
end for
if a is discrete and binary **then**
  X=X-S_i
  for each a $\in S_a$ **do**
    split at node N in such a manner that one split contains the tuples satisfying the condition and the other contains the remaining tuples.
  end for
end if
if a is continous **then**
  two splits are formed at split-point c
  Split A =$\sum_{i=1}^{p} t_i$, if $a > c$
  Split B =$\sum_{i=1}^{p} t_i$, if $a <= c$
end if
if partition is not pure or spilitting is further Possible **then**
  goto ***Start***
end if
return N

is, it reduces the computational complexity by reducing the number of categories for each predictor using the merging procedure.

Algorithm 2 Sequential Cross-Tabulation Approach

1: Cross-tabulate n categories of independent variables with m categories of dependent variables.
2: Apply the chi-square test on the cross table and find the pair of categories of the independent variables which are least significantly different.
3: Merge the two categories which pass through step 2.
4: Repeat steps 2 and 3 until no non-significant chi-square test result is obtained.
5: Select the attribute whose chi-square result is largest, and split into k branches where, $k \leq l$, and l is the number of categories of the independent attributes obtained from the merging process.
6: Repeat step 5 until the stopping criteria is satisfied.

2.3 Iterative dichotomizer(ID3)

It employs entropy as a measure of node impurity [37] [38]. It uses ordered discrete attributes. The expected information or entropy relies on the probability of belongingness (P_i) of any tuple of a dataset D to a particular class. Entropy for n classes in a dataset can be computed as follows [17]:

$$En(D) = -\sum_{i=1}^{n} P_i \log_2(P_i), \tag{2}$$

where, $P_i = \frac{|S_i|}{|S|}$, in which the denominator denotes the number of tuples in D and the numerator contains the amount of samples with respect to class C_i. In addition, the entropy of the partitions is to be calculated based on the values of attribute t in the dataset (D). For s distinct values $\{t_1, t_2, t_3, ..., t_s\}$ of each attribute t, the entropy of the partition with respect to t is:

$$En_t(D) = \sum_{i=1}^{s} \frac{|D_i|}{|D|} \times En(D_i). \tag{3}$$

where D is partitioned into s subsets $\{D_1, D_2, D_3, ..., D_s\}$, and $En(D_i)$ is the entropy of the partition with respect to values of an attribute *t*. D_i consists of the tuples in D having outcome t_i of the attribute *t*. This is required to obtain the exact classification of the instances. The information gain, $G(t)$, is computed as follows:

$$G(t) = En(D) - En_t(D). \tag{4}$$

The attribute with highest $G(t)$ or minimum $En_t(D)$ is chosen as the splitting attribute. Originally, ID3 was proposed considering discrete data only, but later it experimented on continuous data in several works. Some have

considered the midpoint between each pair of adjacent values as a possible split-point and some have used discretization to convert continuous data to discrete and then applied ID3 on that data. In case of midpoint procedure, possible splits for an attribute t, are of the form $t \leq c$ for one set of tuples, and $t > c$ for another set of tuples where, c is the split-point between two adjacent pair of attribute values t_i and t_{i+1}. The value of c can be calculated as: $(t_i + t_{i+1})/2$.

2.4 C4.5

C4.5 is a descendant of ID3, proposed by J. R. Quinlan [39]. The major limitation of ID3 is that it gives preference to the attributes having more values and more missing values. In order to overcome this problem, gain ratio was adopted as the attribute selection measure instead of entropy. For s subsets $D_1, D_2, ..., D_s$ of dataset D, instead of using the entropy, it uses the splitting information (SI_t) [17]:

$$SI_t(D) = \sum_{i=1}^{s} \frac{|D_i|}{|D|} \times log_2 \left(\frac{|D_i|}{|D|} \right). \tag{5}$$

The gain ratio (GR) is the ratio of entropy and $SI_t(D)$:

$$GR(t) = \frac{G(t)}{SI_t(D)} \tag{6}$$

The attribute having the highest $GR(t)$ value is chosen as the splitting attribute. The problem arises when $SI_t(D)$ becomes negligible or tends to zero. It leads to unbalanced ratio; hence one constraint needs to be imposed, that is, $G(t)$ value should be large enough when the gain ratio is applied.

2.5 Classification and regression trees (CART)

In contrast to ID3 and C4.5, it generates binary decision trees [40]. It works on both discrete and continuous data. It uses gini index (GI) as a measure of node impurity [41].

$$GI(D) = 1 - \sum_{i=1}^{n} P_i^2, \tag{7}$$

where, $P_i = \frac{|S_i|}{|S|}$ is the ratio of number of tuples present in the dataset with respect to a particular class to the total number of tuples present in D. For a binary split with respect to an attribute 't', GI can be calculated as:

$$GI_t(D) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} GI(D_i) \tag{8}$$

where, D_i is the gini index with respect to a partition. Due to the binary split on attribute t, the reduction in impurity is computed as:

$$GI_{red}(t) = GI(D) - GI_t(D) \tag{9}$$

For each attribute, every feasible binary splits are taken into consideration. The subset with minimum $GI_{red}(t)$ is chosen as the splitting subset [42]. For continuous-valued attributes, it uses the same midpoint procedure as ID3 to find a possible split-point.

2.6 Fast and accurate classification trees (FACT)

The FACT algorithm for decision tree used for classification is similar to the recursive Linear Discriminant Analysis (LDA) procedure in which the tree is constructed with linear splits [43]. The number of children for each predictor is the same as the number of classes for that variable. In this algorithm, the predictors are being ranked based on the Analysis of Variance (ANOVA) and F-test, and the splitting procedure is performed on the selected predictor based on the LDA method [44]. Initially, all the categorical independent variables are transformed into ordered variables using an intermediate binary vector. One of the specialties of this algorithm is the procedure of handling the missing values. It estimates the means and modes of non-missing data values of ordered and categorical predictors respectively, and replaces those values in place of missing values. The size of the tree is identified based on the stopping criteria of the ANOVA test [45]. The major advantage of this algorithm is, it is unbiased towards the selection of predictors at each level. However, it is biased towards the predictors, which are categorical as LDA is employed to convert it into an ordered one. This limitation is addressed by the Quick Unbiased Efficient Statistical Tree (QUEST) algorithm, which removes the bias for the splitting of ordered variables.

2.7 Quick unbiased efficient statistical tree (QUEST)

It is an efficient decision tree classifier that addresses the FACT algorithm's limitation, which is biased towards the selection of categorical variables. QUEST uses the cross-tabulation approach of chi-squared tests, and F-tests to handle categorical and ordered predictors, respectively to give a fair chance of selection [46]. When a binary split is required at a node with more than one class, it merges the classes into two superclasses before the significance test is applied. If the variable is ordered, the split-point is chosen by quadratic discriminant analysis or the exhaustive search. Apart from that, if the variable is categorical, the point of splitting is chosen after transforming it into a larger discriminant coordinate. The major advantage of QUEST is, it improves the computational time over CART when variables with many categories exist.

2.8 Classification rules with unbiased interaction selection and estimation (CRUISE)

In CRUISE, each node is split into multiple branches, which depends on the number of class labels associated with the independent variables [47]. It is an extension of QUEST. The variable selection at each level is based on the cross-tabulation approach used in CHAID, where the columns and rows of the cross table contain the predictors and class labels, respectively. Unlike QUEST, it performs the significance tests between two independent variables, say X_i and X_j instead of performing pairwise significance tests between two categories of X variables [48]. If the significance test between X_i and X_j are found to be best, X_i is chosen for splitting instead of X_j. The split-point is then identified by the LDA approach after the independent variable goes through a Box-Cox transformation. The major advantage of CRUISE is, it allows splitting of all the variables linearly that can fit the LDA model at each leaf node. Another advantage is, it is unbiased towards the selection of variables that have more missing values.

2.9 Generalized unbiased interaction detection and estimation (GUIDE)

GUIDE is the improvised version of QUEST and CRUISE. It models the decision tree classifier by leveraging the strengths of both algorithms. It also reduces the limitations of CRUISE by minimizing the number of interaction tests among the categorical variables. The amount of computation is drastically reduced as it restricts the frequency of tests. The multi-level searching technique is employed for splitting at each node when the significant difference between two variables X_i and X_j is noticed. The first level splitting of a node is performed based on X_i and the second level splitting based on X_j in order to reduce the amount of impurity. This process is repeated in a reverse manner, i.e, X_j is considered for splitting at first level and X_i in second. The one whose reduction in impurity is greater is chosen to split the node. One of the advantages of GUIDE is, it can perform bivariate splits of two independent variables at a time along with univariate splits. Bivariate linear split is preferable over univariate if the number of observations at each node is found to be lesser than the number of independent variables.

2.10 Conditional inference tree (CTREE)

It can handle ordered, nominal, continuous, censored as well as multivariate attributes. It uses the combination of recursive binary partitioning and theory of permutation to select split variables [49]. Based on Bonferroni adjusted p-values, it derives stopping rules to regulate the tree size instead of applying tree pruning to reduce the tree size. Like CART, it also uses surrogate splits to deal with missing values, and the number of surrogate splits can be regulated by defining maximum surrogate splits using a function.

3 Application details of the techniques reviewed under DT for classification

This section exemplifies a brief illustration of the splitting criterion used, application areas, dataset details, and performances of the different algorithms reviewed under DT for classification. THAID was used in finance and health care for various purposes. CHAID was applied in many application areas like marketing, health care, coal mining, etc and its performance is comparable with several algorithms of its time. CHAID was also used in the public vocational rehabilitation program to predict the employment outcomes and acceptance rates of rehabilitation clients with orthopedic disabilities. ID3 is adopted in many application areas like price prediction in stock markets, in health care for medical diagnosis and it is having better classification accuracy than neural networks and rough sets classifiers. The extended version of ID3, i.e., C4.5 was employed in several sectors like health care for liver disease diagnosis, detection of cancer disease with the help micro-array datasets, and tumor classification [50]. It is also used in land cover mapping and change assessment in remote sensing, etc. Its performance is comparable with k-Nearest Neighbor (kNN), Naive-Bayes, and Support Vector Machine (SVM) classifiers.

Similarly, CART is used in various fields like intrusion detection, bankruptcy prediction in companies [51]; diagnosis of diabetes and prediction of heart disease in health care [52]; landslide hazard, etc and have shown better performance than ID3. Likewise, FACT is also used in many areas like waveform recognition, digit recognition, and normal discrimination. Researchers employed QUEST in educational institutions for evaluating teachers' performance [33], in health care for predicting mortality rate because of head injury, financial firms for measuring firm performance, etc. Likewise, GUIDE, CRUISE, and CTREE are used in several research areas and are efficient and effective for the researchers. The details are mentioned in Table 1.

4 Comparative analysis of various classification tree algorithms

In this section, different classification tree techniques, as discussed, are compared based on various parameters, as listed in Table 2. The parameters considered for the comparison are different types of splits (univariate or linear), the maximum number of splits, the way they handle missing valued attributes, node models, etc. CHAID and C4.5 algorithms do not support linear splits. However, most of the algorithms support both linear and univariate splits. The prediction accuracy of THAID is not up to the mark.

Table 1: Application details of the techniques under DT for Classification

Sl No.	Method	Splitting Criterion	Application Area	Dataset Details	Remarks
1	THAID	Sum of Squared Deviation	Finance, Health care	Car dataset from 1970 survey of Consumer finances, IRIS dataset from UCI repository,	Low predictive accuracy, Biaseness in variable selection
2	CHAID	Chi-squared Statistical test	Marketing modelling. Healthcare, Coal mining	IRIS dataset, Breast cancer patients' data, Coal mines data from Coal Industry Promotion Board, Rehabilitation Service Administration (RSA)-911 dataset	Performance comparable and in many cases outperforms other algorithms, restricted to categorical variables
3	ID 3	Entropy	Product entry decision, Weather forecasting, Medical diagnosis, Marketing, Stock market trend mining	Heart disease data from UCI rep., Weather data, Buys_computer data	Predictive accuracy is directly proportional to the size of the training set; Better classification accuracy than rough sets and neural networks
4	C 4.5	Gain Ratio	Finance, Health care, Land cover change assessment	Car dataset from Journal of Statistics Education Data Archive, IRIS dataset from UCI repository, Liver Disorders datasets from UCI repository , data sets Landsat 5 (TM) for 1986 and Landsat 7 (ETM+) for 2001 located on the satellite path; Leukemia, Colon tumour and Diffuse Large B-cell Lymphoma data from Kent Ridge Bio-Medical Data Set Repository	Performance comprable with classifiers SVM, k-NN, Naive Bayes'
5	CART	Gini Index	Medicine and Health care, Landslide hazard, Intrusion Detection	Car dataset, Birth dataset, Type 2 Diabetic outpatient data, Survey data of malaria in central vietnam during 2008, KDD Cup 1999 dataset from UCI rep., Landslide data set of 137570 samples from Penang Island in Malaysia	Great flexibility and accuracy but splitting is biased towards variables having more distinct values
6	FACT	ANOVA and f-test	Normal discrimination, Digit recognition, Waveform Recognition, Spherical distribution problem	IRIS dataset, Boston housing dataset	Classification accuracy and interpretative capability is comparable with CART, but FACT runs many times faster
7	QUEST	Chi-squared & f-test	Financial firms, Health care, Landslide hazard, Coal mine	Car dataset from Journal of Statistics Education Data Archive, IRIS dataset, Financial data of Turkish firms from FINNET, Breast cancer patients' data, Coal mines data from Coal Industry Promotion Board	Unbiased splits, ranked fourth best overall for linear splits, Improved computational time over CART for variables of many categories
8	CRUISE	LDA, Contingency Table Chi-squared tests	Biomedicine, Education, Healthcare	IRIS dataset, Biomedical data, Cylinder bands, Credit approval, Echo-cardiogram, Fish catch, Horse colic, Hepatitis, Heart disease, Auto imports from UCI rep.; Demography data from Rouncefield (1995), Head injury from Hawkins(1997), College data from StatLib	Accuracy as high as CART and QUEST, fast computation speed, produces more intelligent splits and shorter trees, keeps track of local interactions
9	GUIDE	Bonferroni test, Chi-squared test	Education, Sports, Healthcare Region prediction	IRIS dataset from UCI, Cars dataset from the Journal of Statistics Education Data Archive for 2004 model year	Performance better than CRUISE and QUEST, Unbiased variable selection
10	C-TREE	Bonferroni p-test	Healthcare, Sports, Space Physics, Mammography, Biology	Breast cancer, Credit, Heart, Hepatitis, Ionosphere, Sonar, Liver, TicTacToe, Titanic House votes 84 from UCI repository	Performance comparable and in some cases better than GUIDE, Unbiased, uses permutation tests

CHAID favors categorical variables, and it allows multiple splits at a node. CART has great flexibility and accuracy, but splitting is biased towards variables having more distinct values. The classification accuracy and interpretative capability of FACT are comparable with CART, and it runs many times faster than CART. CART, CHAID, and QUEST are the most popular techniques used for modeling decision trees for classification. The QUEST algorithm is a little bit faster as compared to CART and CHAID. However, it is not suitable for processing bigger datasets as it requires high storage space to store the intermediate results obtained at each level of the tree.

QUEST, CRUISE, GUIDE, and CTREE are the advanced approaches to model the classification tree. They were found effective in terms of both time and space complexity. They also provide unbiased splits during the construction of the classification tree. Accuracy of CRUISE is as good as CART and QUEST; it has fast computational speed, generates shorter trees, more intelligent splits, and also keeps track of local interactions that makes it distinguishable from other algorithms proposed before it [18]. GUIDE is having better accuracy than CRUISE and

Table 2: Comparision of classification tree algorithms

Author Name	Year	Algorithm	Split type	Unbiased Split	No. of splits	Missing values Method	Interaction Test	Node Model
R. Messenger & L. Mandell	1972	THAID	U	No	2	–	Yes	C
G. V. Kass	1980	CHAID	U	No	>=2	B	Yes	C
J R Quinlan	1986	ID3	U	No	>=2	–	No	C
J R Quinlan	1993	C 4.5	U	No	>=2	W	No	C
L Breiman et al.	1984	CART	U,L	No	2	S	No	C
W Y Loh & N. Vanichsetakul	1988	FACT	U,L	No	>=2	I	No	C
W Y Loh & Y S Shin	1997	QUEST	U,L	Yes	2	I	No	C
H Kim & W Y Loh	2001	CRUISE	U, L	Yes	>=2	I, S	Yes	C, D
W Y Loh	2002	GUIDE	U, L	Yes	2	M	Yes	C, K, N
T Hothorn et al.	2006	C-TREE	U,L	Yes	>=2	I,S	No	C

Description: U- univariate splits, L- Linear splits, B-Missing value branch, W- Probability weights, S- Surrogate splits, I- Missing value imputation, C- Constant model, M- missing value category, D- Discriminant model, K- kernel density model, N- Nearest neighbour model. Blank entries indicate 'no missing values' .

QUEST and is having an unbiased variable selection. In contrast to others, CTREE uses permutation tests. Its performance is comparable, and in some cases, it is better than GUIDE.

5 DT for regression

Regression aims to predict a continuous value for an unseen tuple by studying a training sample of data [29]:

$$O = f(x, \theta), O \in R \tag{10}$$

where, x is the new observation, O is the output, $f(.)$ is the regression function and θ is the regression function's parameter set. DT for regression is similar to classification trees with the difference that it contains values or piecewise models at leaves rather than class labels [53]. The values may be the result of any test or the outcome of any operation. Some of the popularly used regression tree algorithms are discussed in this section.

5.1 Automatic interaction detection (AID)

It is the first regression tree algorithm, introduced by Morgan and Sonquist in the year 1963. This algorithm starts with a large dataset. The large dataset is then successively divided into several subgroups after applying binary divisions. At every step, the binary divisions of the groups are defined by one of the independent variables. It uses the sum of squared deviations as a measure of node impurity [54]. For each independent variable, all possible splits are considered. Each binary split divides the whole dataset into two parts. The one having least sum of squared deviations is chosen. The node impurity measure ($I(d)$) is computed as follows [24]:

$$I(d) = \sum_{i=1}^{n} (y_i - \bar{y_d})^2 \tag{11}$$

where, $\bar{y_d}$ is the sample mean of dependent variables with respect to the partition. The attribute with the least sum of squared deviations is taken as the splitting attribute. The splitting process continues till very few tuples remain in the dataset or when $I(d)$ becomes less than a predefined value. The task of deciding the predefined value is a matter of concern, as it might lead to the problem of over-fitting or under-fitting if the number is either too large or too small, respectively. Inter-correlation among attributes leads to spurious results. A biased value is considered during the model building process.

5.2 CART for regression

It uses the same approach as AID for splitting and computing the node impurity measure. It solves the over-fitting problem of AID by using the tree pruning procedure. The yield of CART is piecewise constant models. CART uses *surrogate splitting* approach to handle datasets with missing values [55]. If splitting needs to be performed on an attribute with missing values, then it finds an attribute that is highly correlated to the original attribute and replaces that attribute with the original one.

5.3 Multivariate adaptive regression splines (MARS)

MARS is suitable for handling datasets of higher dimensions. It follows the recursive, divide and conquer approach as regression and generates continuous models with continuous derivatives [56]. It splits the range of independent

attribute values into $n+1$ disjoint intervals partitioned by n knots, which results in the construction of functions, called spline functions [57]. MARS comprises of a series of connected straight line segments. The general form of MARS model is defined as [58]:

$$y = f(x) = z_0 + \sum_{i=1}^{n} z_i B_{kn}(x_{v(k,i)}) \tag{12}$$

where, y is the output function, n is the number of basis functions, z_0 is a constant value, k is the order of interactions, $x_{v(k,i)}$ is the independent attribute in the k^{th} of the i^{th} product, $B_{k,n}(x_{v(k,i)})$ is the i^{th} basis function and z_i is its corresponding coefficient. The basis function can be defined as: $B_{kn} = \prod_{i=1}^{k} b_{in}$. The value of k is one if the model is additive, and it is two, for the pairwise interactive model. In the first step, a significant quantity of basis functions are constructed which overfit the data. The permitted data values are categorical, continuous, and/or ordinal and they are selected as per the intervals defined. The different variables may have direct interaction with each other or some constraints may be imposed on them. In the second phase, a generalized cross validation technique is applied on the basis functions and the functions having the least contribution are eliminated. The variables having better cross-validation results are chosen. In this way, an optimal MARS model is selected. MARS successfully handles missing values by employing dummy variables. By using the above-mentioned procedures, MARS also keeps track of complex data structures hidden in high dimensional datasets.

5.4 GUIDE for regression

It employs the chi-square test to detect the interrelationships between the signed residuals and groups of independent variables [59]. It can handle datasets having both discrete and continuous-valued attributes. In GUIDE, two tests are performed, curvature test and interaction test. In the curvature test, for each continuous-valued attribute, a 2 x 4 table, called contingency table is created using the dataset, whose rows indicate signs of the residuals and columns stipulate groups. Based on the number of observations in each cell, the p-value is obtained from the chi-square distribution. In the Interaction test, to find interaction among two continuous variables, the sample median is computed, and based on the result, the range of each variable is divided into two equal partitions. A 2×4 contingency table is generated, whose rows represent residual signs and columns denote quadrants. The chi-square distribution and p-value are also computed in this algorithm. If the acquired p-value is a consequence of the curvature test, the corresponding independent variable is chosen as the splitting attribute; and if it is from the interaction test, then one of the interacting variables is chosen as the splitting attribute. The sum of squared error is computed for each sub-node, and the variable having the least sum of squared error is selected. In case one of the variables is categorical, the one having a smaller p-value as a result of the curvature test is selected. The major advantage of GUIDE is that it is unbiased towards the splitting process.

5.5 M5

It involves the construction of model trees rather than the rule based, recursive binary trees [60]. Model trees are smaller in structure than regression trees and have shown better performance than the later. They can handle datasets of large dimensions. In contrast to regression trees, which contain values at their terminal nodes, model trees employ linear functions. M5 can handle both discrete and continuous data. Its objective is to build a model that associates the target values of the dependent variables to other attributes' values [61]. The construction of model trees follows the divide and conquer approach. If the constructed model suffers from over-fitting, tree pruning is applied by substituting a subset with a leaf. M5 considers standard deviation as a measure of node impurity. At first, the standard deviation of the dependent variables is computed in the training sample of the dataset (Dt). Based on the outcomes of the test, the splitting process continues until there is no notable distinction between the values of the attributes. By ascertaining the subset of data tuples associated with each outcome, every potential test is evaluated. The expected reduction in error can be calculated as [60]:

$$\Delta err = \sigma(Dt) - \sum_{i=1}^{n} \frac{Dt_i}{Dt} \times \sigma(Dt_i) \tag{13}$$

where, Dt_i denotes the subset of data tuples having i^{th} outcome of the potential test, n denotes the number of outcomes of a test, and $\sigma(Dt)$ represents standard deviation of the training dataset. The test having maximum Δerr is chosen as the potential test to predict the target values of the unseen data tuples. The test set error (Δerr_t) can be computed as:

$$\Delta err_t = \Delta err \times (\frac{m+p}{m-p}) \tag{14}$$

where m represents the number of training set tuples at a particular node, and p refers to the number of parameters in the regression model of the node.

5.6 $M5'$

It is an extension of M5, designed to address some issues that arose during the construction of M5. It is a k+1 parameter model, where k attributes and one constant term w_0 are there. In M5, as the size of the tree becomes smaller, the standard deviation in Δerr lessens. Hence, to manage this, a pruning factor called α is used in $M5'$ while computing Δerr_t [62].

$$\Delta err_t = \Delta err \times (\frac{m+\alpha p}{m-p}) \tag{15}$$

As the α value increases, Δerr_t increases but the size of the tree decreases outstandingly. Hence, to get less error and better performance, a smaller value of α must be taken; but if preference will be given to generate smaller trees then α value must be increased a little bit. In addition to this, $M5'$ also successfully handles the datasets containing missing values [63]. To address missing values, some modifications to Δerr has been done as follows [62]:

$$\Delta err = \frac{k}{|Dt|} \times \beta(i) \times \left[\sigma(Dt) - \sum_{j \in \{A,B\}} \frac{|Dt_j|}{|Dt|} \times \sigma(Dt) \right] \tag{16}$$

where k refers to the number of tuples without missing values, Dt is the subset of the dataset containing tuples that are to be split based on a condition, Dt_A and Dt_B are the sets after partition, and $\beta(i)$ is the correction factor defined as [62]: $\beta = e^{7 \times \frac{2-a}{m}}$, where m is the number of tuples in the dataset, and a is the total number of values of original enumerated attributes. β is used for converting 'a' valued enumerated attributes to 'a-1' binary values. For continuous attributes, β is taken as 1. Hence, after splitting, all attributes in Dt_A and Dt_B become binary. The attribute with a maximum Δerr is chosen as the splitting attribute. The prediction accuracy of $M5'$ is comparable with techniques like Artificial Neural Networks(ANN) and is found to be better than that of regression trees like CART [64].

6 Application details of the techniques reviewed under DT for regression

This segment epitomizes a short depiction of the different algorithms reviewed under DT for regression based on few parameters as alluded in Table 3. It starts with AID. It has been adopted in several research areas like education for predicting the factors affecting the academic survival of students, in population studies for the adoption of family planning in Koyang [65], in marketing for exploratory analysis of market data, etc. CART is used in several fields like sports for analyzing the salary of baseball players [24], in public health for analyzing causes of morbidity and mortality from specific diseases, and in many areas for different tasks using several datasets [53]. MARS is adopted in several research areas like health care on heart attack survival data, in biology for prediction of species distributions, etc. In some applications like credit scoring, CART and MARS outperform traditional logistic regression, discriminant analysis, SVM, and neural network techniques in terms of predictive accuracy. GUIDE is employed in various sectors like education, sports, and automobiles. MARS and M5 were also used for groundwater level forecasting, solar radiation, in the construction industry for evaluating mechanical properties of concretes containing coarse recycled concrete aggregates [66] and they have outperformed other algorithms in many scenarios. $M5'$ is used in various application areas like coastal engineering for prediction of wave height in Lake Superior [67], for scour depth prediction [68], in construction industry for predicting modulus elasticity of recycled concrete [63], etc. For more details, Table 3 may be referred.

7 Comparative analysis of regression tree algorithms

Comparative analysis of various regression tree algorithms based on different parameters is presented in Table 4. Some of the parameters considered for comparison are the same as the parameters used for comparing classification tree algorithms. However, few parameters like pruning, variable importance ranking, loss criteria, ensemble approach are added for an extensive comparison of regression tree algorithms. Regardless of its novelty, AID faced some problems and was criticized by several authors [24]. While splitting, it experiences over-fitting as well as under-fitting. It doesn't employ tree pruning to reduce the tree size, whereas CART and others do the same to reduce the complexity. The square of mean deviation is considered as the node impurity adopted in the AID and CART algorithm. MARS employs a spline basis function and incorporates a generalized cross-validation approach that increases the prediction accuracy of the model. GUIDE employs ensemble and bagging techniques in contrast to others. $M5'$ is the best regression model which constructs the piecewise constant tree by fitting the linear regression model at each leaf node whereas, the GUIDE algorithm fits linear regression models at each node in the constructed tree. For detailed analysis, Table 4 may be referred.

8 Conclusions and future work

The popularity of the classification and regression trees has been increasing exponentially, as they are easy to understand and implement. In the decision tree, the hidden rules along with the constraints can be extracted from the data and can be mapped with the nodes and branches of the tree, which makes it more convenient for understanding. However, the complexity of the model increases with the increase in the size of the datasets. To handle the complexity, a wide number of advanced algorithms have been adopted in the field of DT for classification and regression. In this paper, we have presented the list of the datasets and various applications in which these algorithms can be applied. This paper could be a potential resource for the researchers in searching and deciding the appropriate algorithms suitable for their area of research, which involve regression and classification task. The comparative analysis of numerous algorithms based on various parameters is also presented for both classification and regression tasks. In future, this work can be extended by including all the en-

Table 3: Application details of the techniques under DT for Regression

Sl. No.	Method	Splitting Criterion	Application Area	Dataset Details	Remarks
1	AID	Sum of Squared Deviations	Education, Population Studies, Market Research, Operational Research, Fishing Industry, Gasoline Consumption	Data from almost 6,000 gas stations from major oil companies in the Unites States during 1970; The shing log data from The White Fish Authority, Hull, England; Data taken from Ronald Freedman, P. Whelpton and Arthur Campbell; Family Planning, Sterility and Population Growth(New York, 1959); Data from NestleCompany, Edu- cational data from The National Survey of Health and Development	It is biased towards datasets of higher dimensions, Experiences overfitting and underfitting problems
2	CART	Gini Index	Medicine and Health Care, Landslide hazard	Baseball salary data from American Statistical Association Section(StatLib), Data from the 1999 Behavioral Risk Factor Surveillance System (BRFSS) (61), conducted annually by U.S. states' Departments of Health in collaboration with the Centers for Disease Control and Prevention	Great flexibility and accuracy but splitting is biased towards variables having more distinct values
3	MARS	Spline Basis Functions, Generalized Cross Validation	Health Care, Biology, Credit Scoring, Solar Radiation, Construction Industry	Heart attack survival data from Specialized Center of Research on Ischemic Heart Disease at the University of California, San Diego; Birds data of many countries and Plants data of Switzerland; Credit card data set provided by a local bank in Taipei, Taiwan, Solar data from Data obtained from Adana and Antakya stations, Turkey	Incorporation of Generalized Cross Validation increases the prediction accuracy, Handles Curse of dimensionality problem
4	GUIDE	Chi-squared test of interaction	Sports, Automobiles, Education	Baseball salary data from StatLib; Car dataset from the Journal of Statistics Education Data Archive for 2004 model year	Fast computation speed, Unbiased and keeps track of local interactions during split selection
5	M5	Standard deviation	Medicine and Health Care, Manufacturing, Automobiles, Hydrology, Solar Radiation, Evapotranspiration	CPU performance data; Car price data; Drug Activity data, LHRH data from Arris Pharmaceuticals, San Francisco; Data from a discharge measuring station Swarupganj on the river Bhagirathi, India; Sediment yield data from Nagwa watershed in India from 1993 to 2004; Solar data from Data obtained from Adana and Antakya stations, Turkey, Climatic data of Davis station maintained by California Irrigation Management Information System (CIMIS)	Better accuracy and smaller in structure than regression trees
6	M5'	Standard deviation	Marine & Coastal Engineering, Construction Industry, Coastal and Ocean engineering	Wind and Wave data gathered in Lake Superior from 6 April to 10 November 2000 and 19 April to 6 November 2001; wave run-up data of Van der Meer and Stam (1992)	The prediction accuracy is comparable with techniques like Artificial Neural Networks and is found to be higher than CART, Handles datasets with missing values

Table 4: Comparison of regression tree algorithms

Author Name	Year	Algorithm	Split Type	Unbiased Split	Number of Splits	Pruning	Variable Importance Ranking	Node Models	Missing value methods	Loss Criteria	Bagging and Ensembles
J. N. Morgan and J. A. Sonquist	1963	AID	U	No	2	No	Yes	C	–	V	No
L Breiman et al.	1984	CART	U,L	No	2	Yes	Yes	C	S	V	No
J. H. Friedman	1991	MARS	L	Yes	>=2	Yes	Yes	C, M	A	V	No
W Y Loh	2002	GUIDE	U	Yes	2	Yes	Yes	C, M, P, R	A	V, W	Yes
J. R. Quinlan	1992	M5	U	No	>=2	Yes	No	C, R	–	V	No
Y. Wang and I. H. Witten	1996	$M5'$	U	No	>=2	Yes	No	C, R	G	V	No

Description: U-Univariate splits, L-Linear splits, C-Constant Model, M-Multiple linear model, R- Stepwise linear model, P- Polynomial Model, S- Surrogate splits, G- Global mean/mode imputation, A- missing value category, V- Least Square, W- Least Median square [Blank entries in the table indicate those algorithms do not handle datasets with missing values]

semble approaches and their comparison with the existing ones. We also aim to explore new techniques in the field of decision tree-based hierarchical multi-label classification, multi-output, and multi-objective regression trees, etc.

References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding (2013) Data mining with big data, *IEEE transactions on knowledge and data engineering*, 26(1), pp. 97–107. https://doi.org/10.1109/tkde.2013.109

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993) Database mining: A performance perspective. *IEEE transactions on knowledge and data engineering*, 5(6), pp. 914–925. https://doi.org/10.1109/69.250074

[3] Ranjan Kumar Behera, Santanu Kumar Rath, Sanjay Misra, Robertas Damaševičius, and Rytis Maskeliūnas (2017) Large scale community detection using a small world model. *Applied Sciences*, 7(11),

pp. 1173.
https://doi.org/10.3390/app7111173

[4] Satchidananda Dehuri and Ashish Ghosh (2013) Revisiting evolutionary algorithms in feature selection and nonfuzzy/fuzzy rule based classification, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(2), pp. 83–108. https://doi.org/10.1002/widm.1087

[5] Leszek Rutkowski (2004) Adaptive probabilistic neural networks for pattern classification in time-varying environment, *IEEE transactions on neural networks*, 15(4), pp. 811–827. https://doi.org/10.1109/tnn.2004.828757

[6] Ranjan Kumar Behera, Debadatta Naik, Dharavath Ramesh, and Santanu Kumar Rath (2020) Mr-ibc: Mapreduce-based incremental betweenness centrality in large-scale complex networks, *Social Network Analysis and Mining*, 10, pp. 1–13. https://doi.org/10.1007/s13278-020-00636-9

[7] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens (2011) Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert systems with applications*, 38(3), pp. 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

[8] Charu C Aggarwal (2014) *Data classification: algorithms and applications*, CRC press.

[9] Salvador García, Alberto Fernández, and Francisco Herrera (2009) Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems, *Applied Soft Computing*, 9(4), pp. 1304–1314. https://doi.org/10.1016/j.asoc.2009.04.004

[10] Shih-Wei Lin, Kuo-Ching Ying, Chou-Yuan Lee, and Zne-Jung Lee (2012) An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection, *Applied Soft Computing*, 12(10), pp. 3285–3290. https://doi.org/10.1016/j.asoc.2012.05.004

[11] Sreerama K Murthy (1998) Automatic construction of decision trees from data: A multi-disciplinary survey, *Data mining and knowledge discovery*, 2(4), pp. 345–389.

[12] Arno De Caigny, Kristof Coussement, and Koen W De Bock (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, *European Journal of Operational Research*, 269(2), pp. :760–772. https://doi.org/10.1016/j.ejor.2018.02.009

[13] Usama M Fayyad and Keki B Irani (1992) On the handling of continuous-valued attributes in decision tree generation, *Machine learning*, 8(1), pp. 87–102. https://doi.org/10.1007/bf00994007

[14] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski (2007) Ensembles of multi-objective decision trees, *European conference on machine learning*, Springer, pp. 624–631. https://doi.org/10.1007/978-3-540-74958-5_61.

[15] Dua Dheeru and Efi Karra Taniskidou (2017) UCI machine learning repository.

[16] Jieyue He, Hae-Jin Hu, Robert Harrison, Phang C Tai, and Yi Pan (2006) Transmembrane segments prediction and understanding using support vector machine and decision tree, *Expert Systems with Applications*, 30(1), pp. 64–72. https://doi.org/10.1016/j.eswa.2005.09.045

[17] Jiawei Han, Jian Pei, and Micheline Kamber (2011) *Data mining: concepts and techniques*, Elsevier.

[18] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine learning*, 40(3), pp. 203–228.

[19] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman (2002) Decision trees: an overview and their use in medicine, *Journal of medical systems*, 26(5), pp. 445–463. https://doi.org/10.1023/a:1016409317640

[20] Claudia Perlich, Foster Provost, and Jeffrey S Simonoff (2003) Tree induction vs. logistic regression: A learning-curve analysis, *Journal of Machine Learning Research*, 4(Jun), pp. 211–255.

[21] Lior Rokach and Oded Maimon (2005) Top-down induction of decision trees classifiers-a survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), pp. 476–487. https://doi.org/10.1109/tsmcc.2004.843247

[22] Rodrigo Coelho Barros, Márcio Porto Basgalupp, Andre CPLF De Carvalho, and Alex A Freitas (2012) A survey of evolutionary algorithms for decision-tree induction, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3), pp. 291–312. https://doi.org/10.1109/tsmcc.2011.2157494

[23] Wei-Yin Loh (2011) Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), pp. 14–23.

[24] Wei-Yin Loh (2014) Fifty years of classification and regression trees, *International Statistical Review*, 82(3), pp. 329–348. https://doi.org/10.1111/insr.12016

[25] Shlomo Geva and Joaquin Sitte (1991). Adaptive nearest neighbor pattern classification, *IEEE Transactions on Neural Networks*, 2(2), pp. 318–322. https://doi.org/10.1109/72.80344

[26] Se June Hong (1997) R-mini: An iterative approach for generating minimal rules from examples. *IEEE Transactions on Knowledge and Data Engineering*, 9(5), pp. 709–717. https://doi.org/10.1109/69.634750

[27] Eric WT Ngai, Li Xiu, and Dorothy CK Chau (2009) Application of data mining techniques in customer relationship management: A literature review and classification, *Expert systems with applications*, 36(2), pp. 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

[28] J Ross Quinlan (1987) Generating production rules from decision trees, In *ijcai*, Citeseer, 87, pp. 304–307.

[29] Ye Ren, Le Zhang, and Ponnuthurai N Suganthan (2016) Ensemble classification and regression-recent developments, applications and future directions, *IEEE Computational Intelligence Magazine*, 11(1), pp. 41–53. https://doi.org/10.1109/mci.2015.2471235

[30] S Rasoul Safavian and David Landgrebe (1991) A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics*, 21(3), pp. 660–674. https://doi.org/10.1109/21.97458

[31] Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim (2002) A personalized recommender system based on web usage mining and decision tree induction, *Expert systems with Applications*, 23(3), pp. 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

[32] Michael Kearns and Yishay Mansour (1999) On the boosting ability of top-down decision tree learning algorithms, *Journal of Computer and System Sciences*, 58(1), pp. 109–128. https://doi.org/10.1006/jcss.1997.1543

[33] Wei-Yin Loh and Yu-Shan Shih (1997) Split selection methods for classification trees. *Statistica sinica*, pp. 815–840.

[34] Avrim L Blum and Pat Langley (1997) Selection of Relevant Features and Examples in Machine Learning, *Artificial intelligence*, 97(1-2). pp. 245–271. https://doi.org/10.1016/s0004-3702(97)00063-5

[35] Robert Messenger and Lewis Mandell (1972) A modal search technique for predictive nominal scale multivariate analysis, *Journal of the American statistical association*, 67(340), pp. 768–772. https://doi.org/10.1080/01621459.1972.10481290

[36] Gordon V Kass (1980) An exploratory technique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), pp. 119–127. https://doi.org/10.2307/2986296

[37] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, and Piotr Duda (2013) Decision trees for mining data streams based on the gaussian approximation, *IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp. 108–119. https://doi.org/10.1109/tkde.2013.34

[38] J. Ross Quinlan (1986) Induction of decision trees, *Machine learning*, 1(1), pp. 81–106.

[39] Salvatore Ruggieri (2002) Efficient C4.5 [classification algorithm], *IEEE transactions on knowledge and data engineering*, 14(2), pp. 438–444.

[40] Leo Breiman, 2017 *Classification and regression trees*, Routledge.

[41] Haidi Rao, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan, and Lichuan Gu (2019) Feature selection based on artificial bee colony and gradient boosting decision tree, *Applied Soft Computing*, 74, pp. 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

[42] B Chandra and P Paul Varghese (2009) Fuzzifying gini index based decision trees, *Expert Systems with Applications*, 36(4), pp. 8549–8559. https://doi.org/10.1016/j.eswa.2008.10.053

[43] Wei-Yin Loh and Nunta Vanichsetakul (1988) Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association*, 83(403), pp. 715–725. https://doi.org/10.1080/01621459.1988.10478652

[44] Xiao-Bai Li, James R Sweigart, James TC Teng, Joan M Donohue, Lori A Thombs, and S Michael Wang (2003) Multivariate decision trees using linear discriminants and tabu search, *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 33(2), pp. 194–205. https://doi.org/10.1109/tsmca.2002.806499

[45] Richard J Light and Barry H Margolin (1971) An analysis of variance for categorical data, *Journal of the American Statistical Association*, 66(335), pp. 534–544.

https://doi.org/10.1080/01621459.1971.10482297

[46] Dursun Delen, Cemil Kuzey, and Ali Uyar (2013) Measuring firm performance using financial ratios: A decision tree approach, *Expert Systems with Applications*, 40(10), pp. 3970–3983. https://doi.org/10.1016/j.eswa.2013.01.012

[47] Hyunjoong Kim and Wei-Yin Loh (2001) Classification trees with unbiased multiway splits, *Journal of the American Statistical Association*, 96(454), pp. 589–604. https://doi.org/10.1198/016214501753168271

[48] João Gama (2004) Functional trees, *Machine Learning*, 55(3), pp. 219–250.

[49] Torsten Hothorn, Kurt Hornik, and Achim Zeileis (2015) ctree: Conditional inference trees, *The Comprehensive R Archive Network*, pp. 1–34.

[50] Jianhua Dai and Qing Xu (2013) Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, *Applied Soft Computing*, 13(1), pp. 211–221. https://doi.org/10.1016/j.asoc.2012.07.029

[51] Vadlamani Ravi, H Kurniawan, Peter Nwee Kok Thai, and P Ravi Kumar (2008) Soft computing system for bank performance prediction, *Applied soft computing*, 8(1), pp. 305–315. https://doi.org/10.1016/j.asoc.2007.02.001

[52] Vikas Chaurasia and Saurabh Pal (2013) Early prediction of heart diseases using data mining techniques, *Caribbean Journal of Science and Technology*, 1, pp. 208–217.

[53] Hyunjoong Kim, Wei-Yin Loh, Yu-Shan Shih, and Probal Chaudhuri (2007) Visualizable and interpretable regression models with good prediction power, *IIE Transactions*, 39(6), pp. 565–579. https://doi.org/10.1080/07408170600897502

[54] Gordon V Kass (1975) Significance testing in automatic interaction detection (AID), *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2), pp. 178–189.

[55] Dan Steinberg and Phillip Colla (2009) Cart: classification and regression trees, *The top ten algorithms in data mining*, 9, pp. 179.

[56] Jerome H Friedman (1991) Multivariate Adaptive Regression Splines, *The annals of statistics*, 19(1), pp. 1–67.

[57] Tian-Shyug Lee and I-Fei Chen (2005) A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, *Expert Systems with Applications*, 28(4), pp. 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

[58] Mohammad Rezaie-balf, Sujay Raghavendra Naganna, Alireza Ghaemi, and Paresh Chandra Deka (2017) Wavelet coupled mars and M5 model tree approaches for groundwater level forecasting, *Journal of hydrology*, 553, pp. 356–373. https://doi.org/10.1016/j.jhydrol.2017.08.006

[59] Wei-Yin Loh (2002) Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, pp. 361–386.

[60] John R Quinlan et al (1992) Learning with continuous classes, In *5th Australian joint conference on artificial intelligence*, World Scientific, 92, pp. 343–348.

[61] Behrooz Keshtegar, Cihan Mert, and Ozgur Kisi (2018) Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree, *Renewable and Sustainable Energy Reviews*, 81, pp. 330–341. https://doi.org/10.1016/j.rser.2017.07.054

[62] Yong Wang and Ian H Witten (1996) Induction of model trees for predicting continuous classes.

[63] Ali Behnood, Jan Olek, and Michal A Glinicki (2015) Predicting modulus elasticity of recycled aggregate concrete using $M5'$ model tree algorithm, *Construction and Building Materials*, 94, pp. 137–147.https://doi.org/10.1016/j.conbuildmat.2015.06.055

[64] Lisham Bonakdar and Amir Etemad-Shahidi (2011) Predicting wave run-up on rubble-mound structures using M5 model tree, *Ocean Engineering*, 38(1), pp. 111–118. https://doi.org/10.1016/j.oceaneng.2010.09.015

[65] John A Ross and Sook Bang (1996) The AID computer programme, used to predict adoption of family planning in koyang, *Population studies*, 20(1), pp. 61–75. https://doi.org/10.1080/00324728.1966.10406084

[66] Aliakbar Gholampour, Iman Mansouri, Ozgur Kisi, and Togay Ozbakkaloglu (2018) Evaluation of mechanical properties of concretes containing coarse recycled concrete aggregates using multivariate adaptive regression splines (MARS), M5 model tree (M5tree), and least squares support vector regression (LSSVR) models, *Neural Computing and Applications*, pp. 1–14. https://doi.org/10.1007/s00521-018-3630-y

[67] A Etemad-Shahidi and Javad Mahjoobi (2009) Comparison between M5' model tree and neural networks for prediction of significant wave height in lake superior, *Ocean Engineering*, 36(15-16), pp. 1175–1181. https://doi.org/10.1016/j.oceaneng.2009.08.008

[68] Mehrshad Samadi, Ebrahim Jabbari, and H Md Azamathulla (2014) Assessment of M5' model tree and classification and regression trees for prediction of scour depth below free overfall spillways, *Neural Computing and applications*, 24(2), pp. 357–366. https://doi.org/10.1007/s00521-012-1230-9