We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Classic and Bayesian Tree-Based Methods

Amal Saki Malehi and Mina Jahangiri

Abstract

Tree-based methods are nonparametric techniques and machine-learning methods for data prediction and exploratory modeling. These models are one of valuable and powerful tools among data mining methods and can be used for predicting different types of outcome (dependent) variable: (e.g., quantitative, qualitative, and time until an event occurs (survival data)). Tree model is called classification tree/regression tree/survival tree based on the type of outcome variable. These methods have some advantages over against traditional statistical methods such as generalized linear models (GLMs), discriminant analysis, and survival analysis. Some of these advantages are: without requiring to determine assumptions about the functional form between outcome variable and predictor (independent) variables, invariant to monotone transformations of predictor variables, useful for dealing with nonlinear relationships and high-order interactions, deal with different types of predictor variable, ease of interpretation and understanding results without requiring to have statistical experience, robust to missing values, outliers, and multicollinearity. Several classic and Bayesian tree algorithms are proposed for classification and regression trees, and in this chapter, we provide a review of these algorithms and appropriate criteria for determining the predictive performance of them.

Keywords: classic classification trees, Bayesian classification trees, classic regression trees, Bayesian regression trees

1. Introduction

Different parametric traditional models are proposed for predicting different types of outcome variable (e.g., (quantitative, qualitative, and survival data)) and exploratory modeling. These parametric models are: generalized linear models (GLMs) [1], discriminant analysis [2], and survival analysis [3]. Also, different nonparametric methods are proposed for data prediction and some of these methods are: classic and Bayesian tree-based methods, support vector machines [4], artificial neural networks [5], multivariate adaptive regression splines [6], K-nearest neighbor [7], Bayesian networks [8], and generalized additive models (GAMs) [9].

Classic and Bayesian tree-based methods are defined as machine-learning methods for data prediction and exploratory modeling. These methods are supervised methods and are one of powerful and most popular tools for classification and prediction. These methods have some good advantages over traditional statistical methods and these advantages are [10–12]:

- easy to interpret due to display result as graphically;
- understanding result without requiring to have statistical experience;
- deal with high-dimensional dataset and large dataset;
- without requiring to determine assumptions about the functional form of the data;
- deal with nonlinear relationships and high-order interactions;
- invariant to monotone transformations of predictor variables;
- robust to missing values;
- robust to outliers;
- robust to multicollinearity;
- extract homogeneous subgroups of observations.

Tree-based methods have been used in different sciences such as medical studies and epidemiologic studies [13–17]. In these studies, tree models are used for determining risk factors of diseases and identifying high-risk and low-risk subgroups of patients. Tree methods can determine subgroups of patients that need to different diagnostic tests or treatment strategies, indeed these methods are useful for subgroup analysis [18, 19].

Several classic and Bayesian tree algorithms are proposed for classification trees, regression trees, and survival trees. These tree algorithms classify observations into a finite homogeneous subgroups based on predictor variables. Tree model is called classification tree, regression tree, and survival tree, if the outcome variable is a quantitative variable, qualitative variable, and survival data, respectively. Tree-based methods extract homogeneous subgroups of data by a recursively partitioning process and then fit a constant model or a parametric model such as linear regression, Poisson regression, and logistic regression for data prediction within these subgroups. Finally, this process is displayed graphically like a tree structure and this advantage is one of the attractive properties of tree models [20].

In this chapter, we review classic and Bayesian classification and regression tree approaches. Owing to space limitation, Bayesian approaches are discussed more, because this chapter provides the first comprehensive review of Bayesian classification and regression trees.

We begin with a discussion of the steps for tree generating of classic classification and regression trees in Section 2. We mention classic classification trees on Section 3. Section 4 provides a review on classic regression trees. Section 5 contains a discussion of treed generalized linear models. A review of Bayesian classification and regression trees is provided in Section 6. Appropriate criteria for determining the predictive performance of tree-based methods are mentioned in Section 7, and Section 8 presents the conclusion.

2. Classic classification and regression trees

In a dataset with an outcome variable Y and P-vector of predictor variables as $X = \{x_1, \ldots, x_p\}$, recursive partitioning process of tree generating for classic tree

algorithms has several main steps and these steps are: tree growing step, stopping the tree growth step, and tree pruning step. Some of the tree algorithms use two steps (tree growing and stopping the tree growth) for tree generating. These steps are as follows:

2.1 Tree growing

Tree growing step is the first step for tree generating and this step is performed using a binary recursive partitioning process based on a splitting function that this binary tree subdivides the predictor variable space. Tree growth begins at the root node and this node is the top-most node in the tree and includes all observations in the learning dataset. Tree grows by either splitting or not splitting each node of tree (each node contains a subset of learning dataset) into two child nodes or left and right daughter nodes using splitting rules for classifying observations into homogeneous subgroups in terms of outcome variable. Splitting rules for classifying observations are selected using some splitting functions. Binary recursive partitioning process continues until none of the nodes can split or stopping rule of tree growth is reached. We will mention these stopping rules. Binary recursive partitioning process splits each node of tree into only two nodes, but some of tree algorithms can generate multiway splits [20].

In tree growing process, nodes that split are called internal node and otherwise are called terminal node. Each internal node includes a subset of dataset and all internal nodes in tree are parent of their subnodes. Each sample of learning dataset is placed in one of the terminal nodes of tree, and the tree size is equal to the number of terminal nodes of tree. Each node of tree is splitted based on a splitting rule for classifying observations into left and right daughter nodes. If chosen splitting rule is based on a quantitative predictor variable, then observations divide based on $\{x_i \leq s\}$ or $\{x_i > s\}$ into left and right nodes, respectively (s: an observed value of quantitative predictor variable). If chosen splitting rule is based on a qualitative predictor variable, then observations divide based on $\{x_i \in C\}$ or $\{x_i \notin C\}$ into left and right nodes, respectively (C: a category subset of qualitative predictor variable). Many splitting rules can be in each node and all possible splitting rules must be checked for determining best splitting rule using a goodness of fit criterion. This criterion shows the degree of homogeneity in the daughter nodes, and homogeneity is computed using a splitting function and best splitting rule has the highest goodness of fit criterion [20].

Several splitting functions are proposed for classification trees and some of them are [21]: Entropy, Information Gain, Gini Index, Error Classification, Gain Ratio, Marshal Correction, Chi-square, Twoing, Distance Measure [22], Kolmogorov-Smirnov [23, 24], and AUC-splitting [25]. Also, several studies compared the performance of splitting functions [21, 26, 27].

In tree growing process, a predicted value is assigned to each node. Data prediction in classification trees such as C4.5 [28], CART [29], CHAID [30], FACT [31], QUEST [32], CRUISE [33], and GUIDE [34] is based on fitting a constant model like the proportion of the categories of outcome variable at each node of tree. CRUISE algorithm also can fit bivariate linear discriminant models [35] and GUIDE algorithm also can fit kernel density model and nearest neighbor model at each node of tree [34]. All mentioned classification trees except C4.5 tree algorithm accept user-defined misclassification cost, and all except CHAID and C4.5 methods accept user-defined class prior probabilities.

Data prediction in regression trees such as AID [36], M5 [37], CART [29], and GUIDE [38] is based on fitting a constant model like the mean of outcome variable at each node of tree. M5 also can fit linear regression model and GUIDE can fit models such as linear regression model and polynomial model.

2.2 Stopping the tree growth step

Stopping the tree growth step is the second step for tree generating. Tree growth is continued until it is possible, and several rules are proposed for stopping the tree growth and we mention some of them [29, 39]:

- There is only one observation in the terminal nodes.
- All observations in the terminal nodes are belong to a category of outcome variable.
- Node splitting is impossible, because all observations in each of terminal nodes have the same distribution of predictor variables.
- Determining a user-specified minimum threshold for goodness-of-fit criterion of splitting rules.
- There is the number of observations less than a user-specified minimum threshold in the terminal nodes.
- Determining a user-specified maximum for depth of tree.

2.3 Tree pruning step

Tree pruning step is the third step for tree generating and this step is one of the main steps for tree generating. Tree algorithm produces a large maximal tree or saturated tree (the nodes of this tree cannot split any further, because terminal nodes have one observation or observations are belong to a category of outcome variable within each terminal node) and then prunes it to avoid overfitting. In this step, a sequence of trees is generated and each tree in this sequence is an extension of previous trees. Finally, an optimal tree is selected among the trees of sequence based on having lowest cost of misclassification (for classification tree) and lowest estimated prediction error (for regression tree) [29].

Several methods are proposed for tree pruning and some of these methods are [39, 40]: cost-complexity pruning, reduced error pruning, pessimistic error pruning, minimum error pruning, error-based pruning, critical value pruning, and minimum description length pruning [41]. Also, several studies compared the performance of pruning methods [39, 40].

3. Classic classification trees

Several classic classification tree approaches are proposed to classify observations, and data prediction in a dataset contains a qualitative outcome variable Y with K categories or classes and P-vector of predictor variables as $X = \{x_1, \ldots, x_p\}$. We review some of these classification tree algorithms and these algorithms are: THAID, CHAID, CART, ID3, FACT, C4.5, QUEST, CRUISE, and GUIDE. Also, we only checked software programs such as SPSS, STATISTICA, TANAGRA, WEKA, CART, and R for being these tree methods and available software programs are mentioned for each model. Owing to space limitation, we only mention the name of other classification tree algorithms and these algorithms are: SLIQ [42], SPRINT [43], RainForest [44], OC1 [45], T1 [46], CAL5 [47, 48], and CTREE [49].

3.1 THAID (theta automatic interaction detector)

THAID classification tree algorithm is developed by Messenger and Mandell in 1972 and is the first published classification tree algorithm [50]. This tree algorithm only deals with qualitative predictor variables and uses a greedy search approach for tree generating. Splitting function in THAID algorithm is based on the number of cases in categories of outcome variable, and splitting rule for node splitting is selected based on minimizing the total impurity of new two daughter nodes. THAID method does not use any pruning method, and tree growth is continued until decrease in impurity is higher than a minimum user-specified limit.

3.2 CHAID (chi-square automatic interaction detector) and exhaustive CHAID

CHAID classification tree algorithm is developed by Kass in 1980 and this algorithm is a descendant of THAID tree algorithm [30]. This algorithm can generate multiway splits and tree-growing process including three steps: merging, splitting, and stopping. Also, continuous predictor variables must be categorized, because CHAID only accepts qualitative predictor variables in tree generating process. CHAID algorithm uses significance tests with a Bonferroni correction as splitting function, and best splitting rule is selected based on having lowest significance probability. This tree algorithm generates biased splits and deals with missing values. CHAID algorithm is implemented in these software programs: SPSS, STATISTICA, and R (CHAID package).

Exhaustive CHAID algorithm is proposed by Biggs et al. in 1991 and this algorithm is an improved CHAID method. The splitting and stopping steps of this algorithm are the same as the CHAID algorithm, and it just changed to improve merging [51].

3.3 CART (classification and regression trees)

The classic CART model was developed by Breiman et al. in 1984 and this model is a binary tree algorithm [29]. CART algorithm is one of the best known classic classification and regression trees for data mining. CART algorithm generates a classification tree using a binary recursive partitioning, and tree generating process in this algorithm contains four steps: (1) tree growing: tree growth is based on a greedy search algorithm that CART algorithm grows tree by sequentially choosing splitting rules. This classification tree algorithm provides three splitting functions for choosing splitting rules, and these splitting functions are: entropy, Gini index, and twoing. (2) tree growing process continues until none of the nodes can split, and a large maximal tree is generated. (3) tree pruning: CART uses cost-complexity pruning method for tree pruning to avoid overfitting and to obtain "right-sized" trees. This pruning method generates several subtrees or a sequence of pruned trees, and each tree in this sequence is an extension of the previous trees. (4) best tree selection: CART uses independent test dataset or cross-validation to estimate the prediction error (misclassification cost) of each tree and then selects the best tree from sequence of trees with lowest estimated prediction error.

CART can generate linear combination splits and uses surrogate splits for dealing with missing values, and also, these surrogate splits are used to measure an importance score for predictor variables. This best known classic tree algorithm suffers from some problems such as greediness, instability, and bias in split rule selection [52]. CART is available at these software programs: CART, R (rpart package), SPSS, STATISTICA, WEKA, and TANAGRA.

3.4 ID3 (Iterative Dichotomiser 3)

ID3 classification tree algorithm is proposed by Quinlan in 1986 [53]. This algorithm uses a greedy algorithm using information gain as splitting function and this splitting function is based on entropy splitting criterion and best splitting rule has highest information gain. ID3 does not use any pruning methods, and tree growth process is continued until all observations in the terminal nodes are belong to a category of outcome variable and/or best information gain is near to zero. This algorithm only deals with qualitative predictor variables (if dataset contains quantitative predictor variables, they must be categorized). Also, ID3 algorithm cannot impute missing values, and this method like CART model suffers from selection bias, because ID3 algorithm favors the predictor variables with more values for node splitting of tree. ID3 is implemented in these software programs: WEKA and TANAGRA.

3.5 FACT (Fast and Accurate Classification Tree)

FACT classification tree algorithm was introduced by Loh and Vanichsetakul in 1988 [31]. In this algorithm, variable selection for node splitting based on quantitative predictor variable is based on having the largest F-statistics of analysis of variance (ANOVA), and then, linear discriminant analysis is used to determine split point for this variable. FACT model transforms qualitative predictor variables into ordered variables in two steps (first step: these variables are transformed into dummy vectors, second step: these vectors are projected onto the largest discriminant coordinate). FACT generates unbiased splits when dataset contains only quantitative predictor variables. Also, it, unlike other classification tree methods (C4.5, CART, QUEST, GUIDE and CRUISE), does not use any pruning methods, and tree growing is stopped when stopping rule is reached. FACT can deal with missing values and missing values of quantitative and qualitative predictor variables are imputed at each node by the means and modes of the non-missing values, respectively.

3.6 C4.5

C4.5 classification tree algorithm is developed by Quinlan in 1993 and this algorithm is an extension of ID3 tree algorithm [28]. This algorithm uses a greedy algorithm using gain ratio as splitting function and generates biased splits. C4.5, unlike ID3 method, deals with quantitative and qualitative predictor variables and also, deals with missing values. In this tree method, split of quantitative predictor variable is binary split and split of qualitative predictor variable is multiway split (a branch is created for each category of qualitative predictor variable). Pruning method used in this algorithm is error-based pruning method. C4.5 is available at these software programs: R (Rweka package), WEKA, TANAGRA, and also can obtain from: http://www.rulequest.com/Personal/. Also, J4.8 tree algorithm is Java implementation of the C4.5 algorithm in WEKA software.

3.7 QUEST (Quick, Unbiased, and Efficient Statistical Tree)

Quest classification tree algorithm is developed by Loh and Shih in 1997, and this model generates binary splits [32]. This method, unlike other classification algorithms such as CART and THAID, does not use exhaustive search algorithm (because these algorithms suffer from variable selection bias) and so improves computational cost and variable selection bias. Quest tree method uses statistical

test for selecting variable splitting and then variable with smallest significance probability is selected to split node of tree. This method uses F-statistics of analysis of variance (ANOVA) for quantitative predictor variables and chi-square test for qualitative predictor variables. After determining variable, an exhaustive search is implemented to find the best split point and QUEST method uses quadratic discriminant analysis for selecting split point. For determining split point of a qualitative variable, values of this variable must be transforming like method used in FACT algorithm.

Quest like CART can generate linear combination splits and uses cost-complexity pruning method for tree pruning. Missing values of quantitative and qualitative predictor variable are imputed at each node by the means and modes of the non-missing values, respectively. Software for QUEST algorithm can be obtained from: www.stat.wisc.edu/~loh/.

3.8 CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation)

CRUISE tree algorithm was introduced by Kim and Loh in 2001, and this algorithm, unlike other classification tree algorithms (CART and QUEST), generates multiway splits [33]. CRUISE method is free of selection bias and can detect local interactions. Two methods of variable selection are used in this tree model and these methods are: 1D (similar to the method used in QUEST method) and 2D. CRUISE method like CART and QUEST can generate linear combination splits and uses cost-complexity pruning method for tree pruning. Also, a bivariate linear discriminant model can fit instead of constant model in each node of tree [35]. CRUISE uses several methods for imputing missing values in the learning dataset and dataset used for tree pruning. Software for CRUISE algorithm can be obtained from: www.stat.wisc.edu/~loh/.

3.9 GUIDE (Generalized, Unbiased, Interaction Detection, and Estimation)

GUIDE tree algorithm was introduced by Loh in 2009, and this method is an evolution of FACT, QUEST, and CRUISE algorithms and improves the weaknesses of these algorithms [34]. It like QUEST and CRUISE generates unbiased binary splits and can perform splits on combinations of two predictor variables at a time. Also, GUIDE like QUEST and CRUISE methods performs the two-step approach based on significance tests for splitting each node. GUIDE uses a chi-squared test of independence of each independent variable versus dependent variable on the data in the node and computes its significance probability. It chooses the variable associated with the smallest significance probability and finds a split point that minimizes the sum of Gini indexes and uses it to split the node into two daughter nodes.

GUIDE method uses cost-complexity pruning method for tree pruning (this method is used in other tree algorithms such CART, QUEST, and CRUISE). It deals with missing values and assigns them as a separate category. Also, this tree method can compute importance score for predictor variables and can use nearest neighbor model and bivariate kernel density instead of constant model in the nodes of tree. Software for GUIDE algorithm can be obtained from: www.stat.wisc.edu/~loh/.

3.10 Classification tree algorithms for ordinal outcome variable

Several tree methods are proposed for predicting an ordinal outcome variable. Twoing splitting function is extended by Breiman et al. for using classification tree for ordinal outcome variable [29] and also Piccarreta extended Gini-Simpson

criterion for this case [54]. Archer proposed a package in R software (rpartOrdinal package) and this package contains some splitting functions for tree generating for predicting an ordinal outcome variable [55]. Also, Galimberti et al. developed a package in R software (rpartScore package) that overcomes some problems of rpartOrdinal package [56]. Tutz and Hechenbichler extended ensemble tree methods such as bagging and boosting for analyzing an ordinal outcome variable [57]. For study about other approaches, refer to Refs. [49, 57–60].

3.11 Classification tree algorithms for imbalanced datasets

In an imbalanced dataset, one of the classes of outcome variable has fewer samples than other classes and this class is rare. In real applications such as medical diagnosis studies, this rare class is the interest for analyzing. Due to the skew distribution of classes, most classification tree algorithms predict all samples of rare class as a class with more samples. Indeed, these models are not robust to unbalance between classes and have good diagnostic performances only on the class with more samples. Several remedies have been proposed to solve this problem for using classification tree algorithms on the imbalanced datasets. Some of these remedies are: sampling methods (undersampling, oversampling, and synthetic minority oversampling technique (SMOTE)), cost-sensitive learning, class confidence proportion decision tree [61], and Hellinger distance decision trees [62]. Ganganwar in 2012 provides a review of classification algorithms for imbalanced datasets [63].

4. Classic regression trees

Several classic regression trees are proposed to classify observations, and data prediction in a dataset contains a quantitative outcome variable Y and P-vector of predictor variables as $X = {}_{\{}x_1, \ldots, x_{p\}}$. We review some of these regression tree algorithms and these algorithms are AID, CART, M5, and GUIDE. Also, we only checked software programs such as SPSS, STATISTICA, TANAGRA, WEKA, CART, and R for being these tree algorithms, and available software programs are mentioned for each model. Owing to space limitation, we only mention the reference of other classification tree algorithms and refer to references for study about other regression tree approaches [49, 64, 65]. Also, for Poisson regression trees, refer to Refs. [66–69].

4.1 AID (automatic interaction detector)

AID regression tree algorithm is proposed by Morgan and Sonquist in 1963, and this algorithm is the first published regression tree algorithm [36]. It generates binary splits and piecewise constant models. This algorithm uses a greedy search for tree generating and a splitting rule is selected based on minimizing the total sum of the square errors. AID suffers from bias in variable selection and this method does not use any pruning method and tree growing is stopped when the reduction in total sum of the square errors is less than a predetermined value.

4.2 CART

CART algorithm considers both classification and regression trees, and tree-generating process in CART algorithm for generating a regression tree is like classification tree [29]. But another splitting function is used to choosing splitting rules of regression tree, and this function is least squares deviation. Also, CART algorithm for selecting best regression subtree uses independent test dataset or

cross-validation to estimate the prediction error (sum of squared differences between the observations and predictions) of each tree to select the best tree from sequence of trees with lowest estimated prediction error. CART algorithm for regression tree generating like classification tree uses surrogate splits for imputing missing values and can generate linear combination splits. CART is available at these software programs: CART, R (rpart package), STATISTICA, WEKA, and TANAGRA.

4.3 M5

M5 tree algorithm is proposed by Quinlan in 1992 and this algorithm like AID and CART methods generates a piecewise constant model and then fits a linear regression model in nodes of tree [37]. M5 improves the prediction accuracy of tree algorithm using linear regression model at nodes and deals with missing values. Also, this method like CART algorithm uses least-squares deviation as splitting function and can generate multiway splits. In M5, smoothing technique is used after tree pruning, and this technique improves the accuracy predictions. Wang and Witten in 1996 proposed M5′ tree algorithm, and this method is based on M5 method [70]. This method is available at these software programs: WEKA and R (RWeka package).

4.4 GUIDE

GUIDE method is introduced by Loh in 2002, and it generates unbiased binary splits [38]. This method uses a regression model at each node of tree and calculates the residuals. Then, residuals are transformed to a binary variable based on the sign of them (positive or negative), and algorithm is followed like algorithm used for classification tree. This tree method like method used for classification tree uses missing value category and can compute importance score for predictor variables. GUIDE can fit models such as linear regression model and polynomial model instead of constant model in the nodes of tree. Software for GUIDE method can be obtained from: www.stat.wisc.edu/~loh/.

5. Treed generalized linear models

Some of tree-based methods such as CART, QUEST, C4.5, and CHAID fit a constant model in the nodes of tree, thus a large tree is generated, and this tree has hard interpretation. Treed models, unlike conventional tree models, partition data into subsets and then fit a parametric model such as linear regression, Poisson regression, and logistic regression instead of using constant models (mean or proportion) for data prediction. Treed models generate smaller trees in comparison to tree models. Also, treed models can be a good alternative for traditional parametric models such as GLMs, when these parametric models cannot estimate relationship between outcome variable and predictor variables across a dataset. Several tree algorithms are developed that fit parametric models into terminal nodes, and to study these algorithms, refer to Refs. [71–77].

6. Bayesian classification and regression trees

The classic CART algorithm was developed by Breiman et al. in 1984, and this model is one of the best known classic classification and regression trees for data mining. But this algorithm suffers from some problems such as greediness, instability, and

bias in split rule selection. CART generates a tree by using a greedy search algorithm, and this search algorithm has disadvantages such as: limit the exploration of tree space, dependence future splits to previous splits, generate optimistic error rates, and the inability of the search to find a global optimum [78]. CART has instability problem, because by resampling or drawing bootstrap samples from dataset may generate tree with different splits [79]. The splitting method in CART model is biased toward predictor variables with many distinct values and more missing values [80, 81].

Several tree models are suggested to solve these problems and these remedial models are ensemble of trees such as Random Forests [82], Bagging [83], Boosting [84], Multiboost [85], and LogitBoost [86] (for solving instability problem), tree algorithms such as CRUISE [33, 35], QUEST [32], GUIDE [34], CTREE [49], and LOTUS [71] (for solving bias in split rule selection problem), and Bayesian tree approaches and evtree algorithm [78] are suggested to solve greediness problem of CART. Also, Bayesian tree approaches can quantify uncertainty, and these approaches explore the tree space more than classic approaches.

Several Bayesian approaches are proposed for tree-based methods [87–98]. In these Bayesian tree approaches like classic tree approaches, a model is called Bayesian classification trees if the outcome variable is a qualitative variable. Also, a model is called Bayesian regression trees if the outcome variable is a quantitative variable. The method of data prediction in these Bayesian approaches is like classic approaches. The method of data prediction for Bayesian classification trees is based on fitting a constant model like the proportion of the outcome variable in the terminal nodes. Data prediction in Bayesian regression tree is based on fitting a constant model like the mean of the outcome variable in the terminal nodes.

Classic tree approaches use only observations for data analysis, but Bayesian approaches combine prior information with observations. Bayesian tree approaches define prior distributions on the components of classic tree approaches and then utilize stochastic search algorithms through Markov chain Monte Carlo (MCMC) algorithms or deterministic search algorithms for exploring tree space [87–98].

Bayesian tree approaches have materials such as prior distribution function, posterior distribution function, data likelihood function, marginal likelihood function, stochastic search algorithm or deterministic search algorithm for exploring tree space, stopping rule of simulation algorithm (if stochastic search algorithms are used to simulate from posterior distribution and explore tree space) and criteria for identify good trees (if model produces several trees). In this section, we review Bayesian tree approaches and also mention the results of published papers based on using these Bayesian algorithms for data analysis.

6.1 BUNTINE's Bayesian classification tree approach

The first Bayesian tree approach for classification tree model was proposed by Buntine in 1992. This proposed approach offers a full Bayesian analysis for classification tree model by using a deterministic search algorithm instead of using a stochastic search algorithm [87]. This model like other classic tree models uses a splitting function for tree growth using Bayesian statistics with similar performance to splitting methods such as Information Gain and Gini. Buntine also like traditional tree models, in order to prevent overfitting model, used Bayesian smoothing and averaging techniques instead of pruning the tree.

In this Bayesian approach, prior distributions are defined on the tree space and data distribution in the terminal nodes of tree (similar priors distributions use for data distribution in the terminal nodes unlike prior distributions considered on the tree space). Buntine showed the superior performance of Bayesian approach in comparison to classic tree algorithms such as CART model of Breiman et al. and C4

model of Quinlan et al. [99] on several datasets [87]. This Bayesian approach may be obtained from: http://ksvanhorn.com/bayes/free-bayes-software.html.

6.2 CGM's Bayesian CART approach

CGM (Chipman, George, McCulloch) proposed a Bayesian approach for CART model by defining prior distributions on the two components of CART model (Θ, T) in 1998, and these components are a binary tree T with $\mathcal{K}$ terminal nodes and parameter set $\Theta = (\theta_1, \theta_2, \ldots, \theta_{\mathcal{K}})$ [89, 91–93]. Indeed, they define prior distributions on tree structure and parameters in terminal nodes. In this approach, following equation is established for joint posterior distribution of components according to Bayes' theorem:

$$P(\Theta, T) = p(\Theta|T)p(T) \tag{1}$$

where $p(T)$ and $p(\Theta \mid T)$ show the prior distribution for the tree and parameters in terminal nodes given the tree, respectively. In this approach, a similar tree-generating stochastic process is used for $p(T)$ of both classification and regression tree models [89], and this recursive stochastic process for tree growth includes the following steps:

- Start from T that includes only a root node (terminal node η).
- Slit terminal node η with probability $P_{SPLIT} = \alpha\left(1 + d_\eta\right)^{-\beta}$ (d_η shows the depth of the node η. α parameter is the base probability of tree growth by splitting a current node, and β parameter determines the rate at which the propensity to split decreases as the tree gets larger). α and β parameters control the shape and size of the tree and these parameters provide a penalty to avoid overfitting tree.
- If terminal node η splits, then a splitting rule ρ is assigned to this node according to the distribution P_{RULE} (discrete uniform distribution is used for selecting predictor variable to split the terminal node η and splitting threshold for this selected variable)
- Let T as newly created tree from step 3 and run steps 2 and 3 on this tree with η equal to the newly created child nodes.

In this approach, the posterior distribution function $p(T|X, y)$ is computed with combining the marginal likelihood function $p(Y|X, T)$ and tree prior $p(T)$ as follows:

$$p(T|X, y) \propto p(y|X, T)\, p(T) \tag{2}$$

$$p(y|X, T) = \int p(y|X, \Theta, T)\, p(\Theta|T) d\Theta \tag{3}$$

$p(y|X, \Theta, T)$ in Eq. (3) shows the data likelihood function.

A stochastic search algorithm is used for finding good models and simulating from relation (2) by using a MCMC algorithm such as Metropolis-Hastings algorithm. This Metropolis-Hastings algorithm simulates a Markov chain sequence of trees namely T^0, T^1, T^2, ..., and this algorithm starts with an initial tree T^0, then iteratively simulates the transitions from T^i to T^{i+1} by two steps as shown below:

1. Generate a candidate value T^* with probability distribution $q(T^i, T^*)$.

2. Set $T^{i+1} = T^*$ with probability below:

$$\alpha(T^i, T^*) = \min\left\{\frac{q(T^*, T^i)\, p(Y|X, T^*)p(T^*)}{q(T^i, T^*)\, p(Y|X, T^i)p(T^i)}, 1\right\} \tag{4}$$

Else, set $T^{i+1} = T^i$.

In this simulation algorithm, $q(T, T^*)$ generates T^* from T by randomly selecting among four steps. These steps are GROW step, PRUNE step, CHANGE step, and SWAP step. This simulation algorithm is run with multiple restarts instead of a single long chain for reasons such as convergence of the posterior distribution or simulation chain, to avoid wasting long time waiting in areas of trees with high posterior distribution function, and generate a wide variety of different trees. Also, the stopping criterion of simulation algorithm is based on that the chain became trapped in a local posterior model.

This Bayesian approach unlike classic CART model does not generate a single tree, thus good trees for classification tree are selected based on criteria such as having lowest misclassification and largest marginal likelihood function. Also, good trees for regression tree are determined based on having the largest marginal likelihood function and lowest residual sums of squares. CGM by using simulation showed that stochastic search algorithm can find better trees than a greedy tree algorithm. They indicated that the Bayesian classification approach has lower misclassification rate than CART model and they also used Bayesian model averaging for improving prediction accuracy of Bayesian classification trees [89].

6.3 DMS'S Bayesian CART approach

DMS (Denison, Mallick, Smith) in 1998 proposed a Bayesian approach for the CART model, and this approach is quite similar to Bayesian approach of CGM with just minor differences [88]. In this approach, prior distributions are defined over the splitting node (S), splitting variable (V), splitting rule (R), tree size ($\mathcal{K}$), and parameters of data distribution in terminal nodes (ψ).

In this Bayesian approach, joint distribution of model parameters is defined as follows ($p(\mathcal{K})$: prior distribution for size of tree, $p(\theta_k|\mathcal{K})$: prior distribution for parameter set $\theta_k = \{R_k, S_k, V_k, \psi_k\}$ given $\mathcal{K}$ (tree size), $p(y|\mathcal{K}, \theta_k)$: data likelihood function):

$$p(\mathcal{K}, \theta_k, y) = p(\mathcal{K})\, p(\theta_k|\mathcal{K})\, p(y|\mathcal{K}, \theta_k) \tag{5}$$

This Bayesian approach puts a prior distribution over the tree size to avoid overfitting data and uses a truncated Poisson distribution with parameter λ (λ shows the expected number of nodes in the tree and a weakly informative is used prior for tree size by setting λ equal to 10) for $p(\mathcal{K})$ as follows:

$$p(\mathcal{K}) \propto \frac{\lambda^{\mathcal{K}}}{(e^{\lambda} - 1)\mathcal{K}!} \tag{6}$$

Also, $p(\theta_k|\mathcal{K})$ in Eq. (5) is defined as follows:

$$p(\theta_k|\mathcal{K}) = p(R_k|V_k, S_k, \mathcal{K})p(V_k|S_k, \mathcal{K})\, p(S_k|\mathcal{K})\, p(\psi_k|V, S, \mathcal{K}) \tag{7}$$

So, prior for this Bayesian approach is defined as follows:

$$p(\theta_k|\mathcal{K})\, p(\mathcal{K}) = p(R_k|V_k, S_k, \mathcal{K}) p(V_k|S_k, \mathcal{K}) \\ p(S_k|\mathcal{K})\, p(\psi_k|V, S, \mathcal{K})\, p(\mathcal{K}) \quad (8)$$

In this approach, Bayesian analysis of tree size $\mathcal{K}$ and parameter set θ_k is as follows:

$$p(\theta_k, \mathcal{K}|y) = p(\mathcal{K}|y)\, p(\theta_k|\mathcal{K}, y) \quad (9)$$

Also, simulation from the above equation is done by using MCMC algorithms to find good trees and Reversible Jump MCMC algorithm is used to simulate from this equation [100]. This simulation algorithm is performed for a single long chain with a burn-in period to explore the tree space. In this simulation algorithm, trees cannot have sample size less than 5 in the terminal nodes and also cannot have size higher than 6 in during burn-in period of simulation chain of posterior distribution. Reversible Jump MCMC algorithm used by DMS to simulate from Eq. (9) includes four steps: BIRTH (GROW), DEATH (PRUNE), VARIABLE, and SPLITTING RULE. In this simulation algorithm, BIRTH step, DEATH step, VARIABLE step, and Splitting RULE step are randomly chosen with probability $b_{\mathcal{K}}$, $d_{\mathcal{K}}$, $V_{\mathcal{K}}$, and $R_{\mathcal{K}}$, respectively, and algorithm is as follows:

1. Stating with an initial tree.

2. Set $\mathcal{K}$ to the tree size in the present tree.

3. Generate u ~U[0, 1]

4. Go to step type determined by u (a step type is determined based on following conditions):
   if ($u \le B_{\mathcal{K}}$), then go to BIRTH step
   else if ($b_{\mathcal{K}} \le u \le b_{\mathcal{K}} + d_{\mathcal{K}}$), then go to DEATH step
   else if ($b_{\mathcal{K}} + d_{\mathcal{K}} \le u \le b_{\mathcal{K}} + d_{\mathcal{K}} + V_{\mathcal{K}}$), then go to VARIABLE step
   else, go to RULE step

Then, acceptance probability (α) of each step that changes tree $(\mathcal{K}, \theta)$ to tree $(\mathcal{K}^*, \theta^*)$ as follows ($\mathcal{K}_{\text{die}}$ shows the number of possible locations for a death in the current tree):

$$\text{BIRTH step:}\alpha = \min\left\{1, (\text{likelihood ratio}) \times \frac{(\mathcal{K}_{\text{die}} + 1)}{\mathcal{K}}\right\} \quad (10)$$

$$\text{DEATH step:}\alpha = \min\left\{1, (\text{likelihood ratio}) \times \frac{\mathcal{K}}{(\mathcal{K}_{\text{die}} + 1)}\right\} \quad (11)$$

$$\text{VARIABLE and RULE steps:}\alpha = \min\{1, (\text{likelihood ratio})\} \quad (12)$$

if ($u \le \alpha$), then proposed tree to accept, else reject.

The stopping criterion of the above simulation algorithm is based on the stability of the posterior distribution and it can be assessed by drawing a plot of iterations of chain against sampled parameter values. This Bayesian approach, unlike CART, does not produce a tree using stochastic search algorithm. Thus, good classification trees are selected based on criteria such as misclassification rate, deviance ($-2\log p(y|\mathcal{K}, \theta_k)$), and posterior probability, and good classification trees have lowest misclassification

rate, deviance, and largest posterior probability. Also, good regression trees have largest posterior probability and lowest residual sum of squares. DMS indicated that Bayesian approach provides richer output and superior performance than classic CART model [88].

6.4 CGM's hierarchical priors for Bayesian regression tree shrinkage approach

CGM, 2000 proposed a Bayesian approach for regression tree with mean-shift model based on computational strategy of CGM's Bayesian approach in 1998. Unlike the Bayesian approach (1998), it can assume dependence of parameters in the terminal nodes. Indeed, hierarchical priors are used for these parameters and therefore shrunk trees are generated [90]. Hierarchical priors have some advantages such as: shrinkage is used in the stochastic search algorithm unlike proposed methods for tree shrinkage (because these methods use shrinkage after searching tree), fitting a larger tree to the dataset without overfitting and improve predictions. CGM by using simulation showed the superior performance of new Bayesian approach for regression tree with mean-shift model in comparison to Bayesian approach of CGM in 1998, CART model, and tree shrinkage methods of Hastie and Pregibon [90, 101].

6.5 WTW'S Bayesian CART approach

WTW (Wu, Tjelmeland, West), 2007 proposed a Bayesian approach for CART model based on the computational strategy of Bayesian approach of CGM (1998) [95]. In this approach, prior distributions define on the tree, splitting variables, splitting thresholds, and parameters in the terminal nodes. This Bayesian approach like approaches of CGM [89, 90, 92, 93] simulates from the posterior distribution by using the Metropolis-Hastings algorithm. The steps used in simulation algorithm of WTW include GROW step and PRUNE step, CHANGE step, SWAP step, and RESTRUCTURE (RADICAL) step (first three steps are similar to steps of simulation algorithm in Bayesian approaches of CGM). RESTRUCTURE step creates large changes in the structure of tree, but tree size is unchanged. There are some advantages by adding this step to simulation algorithm of posterior distribution such as: improving the convergence of the MCMC algorithm, elimination of the need for restarts of the simulation algorithm unlike Bayesian approaches of CGM, and large changes in the structure of tree without change in tree size.

In this approach, convergence diagnostics of simulation algorithm are based on plots such as: plots of iteration number against log posterior distribution, log marginal likelihood function, number of terminal nodes, and number of times that a particular predictor variable is shown as a splitting variable in the tree. WTW showed the superior performance of Bayesian approach in comparison to CART model and that the Bayesian approach had a lower misclassification rate than the CART model [95].

6.6 OML'S Bayesian CART approach

OML (O'Leary, Mengersen, Low Choy), 2008, proposed a Bayesian approach for CART model by extending the Bayesian approach of DMS. These two Bayesian approaches have differences such as the stopping rule of the simulation algorithm or convergence diagnostic plots, criteria for identifying good trees and prior distributions considered for parameters in the terminal nodes [88, 96, 98].

The stopping criterion of simulation chain in OML'S Bayesian classification trees approach has two steps. The first step includes the plot of iterations against

accuracy measures (false and positive negative rate and misclassification rate), log posterior, log likelihood, and tree size. If these plots show stability in mentioned items, then in second step, structure of component trees (variables and splitting rules at each splitting node) examines in the set of good trees and if this structure was stabilized and/or the same trees were in this set, then convergence has occurred for this simulation chain; otherwise, iterations must be increased until convergence.

The set of good trees in this Bayesian classification tree approach is determined based on the accuracy measures computed from the confusion matrix of Fielding and Bell [102]. Good trees have lowest misclassification rate and false positive and negative rate (or using highest sensitivity and specificity instead of lowest false positive and negative rate) [96, 98, 103]. After convergence of simulation chain, two or three trees are selected as the best trees among set of good trees based on criteria such as modal structure of tree (same size tree with the same variables and splitting rules), lowest misclassification rate, false negative and positive rate and deviance, highest posterior probability and likelihood, using expert judgment and biological interpretability [96, 98, 103].

The stopping rule of simulation algorithm for regression tree like classification tree includes two steps. In the first step, plot of iterations are drawn against posterior probability, residual sum of squares, and deviance. If these abovementioned items are stable, then structure of component trees examines in the set of good trees and if this structure was stabilized, convergence has been occurred for this simulation chain. Also, set of good trees for regression tree is selected based on having the highest posterior probability and likelihood, lowest residual sum of squares, and deviance [98].

OML compared the Bayesian classification trees with the classic CART model on an ecological dataset and concluded that Bayesian approach has smaller false positive rate, misclassification rate, and deviance than CART model, while the CART model has lower false negative rate, but this model had higher false positive rate [96]. They, in 2008, indicated that this Bayesian approach had a lower false negative rate in comparison to Bayesian approach of DMS, but approach of DMS had a lower false positive rate and misclassification rate [96].

OML in 2009 compared predictive performance of random forests with the Bayesian classification trees on the three datasets and they concluded that the best tree selected with Bayesian classification trees has higher sensitivity and better accuracy in comparison to random forests. They expressed that the Bayesian approach may have better performance than random forests in determining important predictor variables in datasets with a large number of noise predictor variables. OML also indicated that the Bayesian classification tree approach unlike random forests is not biased toward assignment of observations to the largest class of outcome variable in predicting data [103].

OML and Hu in 2011 compared the performance of Bayesian classification trees with the CART of Breiman et al., and they concluded that the Bayesian approach has higher sensitivity and specificity in comparison to CART. They also investigated overfitting of the Bayesian approach by using cross-validation method, and this approach did not show any evidence of overfitting [98].

6.7 OMML'S expert elicitation for Bayesian classification tree approach

OMML (O'Leary, Mengersen, Murray, Low Choy), 2008, proposed a Bayesian classification tree approach based on the computational strategy of Bayesian classification tree approach of OML and by using informative priors [96, 97]. In this Bayesian approach, informative priors are used to define Dirichlet distributions for splitting node, splitting variable, and splitting rule as follows:

$$p(S_k|\mathcal{K}) = \mathrm{Dir}(S_k|\alpha_{S_1}, \ldots, \alpha_{S_k}) \tag{13}$$

$$p(V_k|S_k, \mathcal{K}) = \mathrm{Dir}(V_k|\alpha_{V_1}, \ldots, \alpha_{V_k}) \tag{14}$$

$$p(R_k|V_k, S_k, \mathcal{K}) = \mathrm{Dir}(R_k|\alpha_{R_1}, \ldots, \alpha_{R_k}) \tag{15}$$

In Bayesian approach of OML, there was no prior information about splitting node, splitting variable, splitting rule, and hyperparameters in the Dirichlet distributions of above equations. So, these hyperparameters were set equal to 1 and uniform non-informative priors used for splitting node, splitting variable, and splitting rule [96, 98, 103]. In this new approach, an expert is subjected with three questions (ordering, grading, and weighting) about splitting node, splitting variable, splitting rule, and tree size for defining informative priors. Then, existing hyperparameters in the relations (13), (14) and (15) are determined by following the result of a question. Three questions are used for size of the tree to determine λ in relation (6). DMS and OML used a weakly informative prior for tree size by setting $\lambda = 10$ [88, 96, 98, 103]. But OMML unlike DMS and OML used an informative prior for size of the tree [96, 97].

O'Leary et al. in 2008 investigated sensitivity to the choice of the hyperparameters of informative priors for tree size, splitting nodes, splitting variables, and splitting rules in classification trees and they concluded that posterior distribution is relatively robust to these priors except for extreme choices of them [96, 97].

OMML by simulation indicated that the best tree of Bayesian classification trees based on the informative priors has lower false negative rate in comparison to the best tree of Bayesian classification trees based on the non-informative priors [96, 97]. They also indicated the superior performance of Bayesian classification trees based on the informative priors in comparison to proposed expert elicitation approaches for Bayesian logistic regression model [97, 104–107].

6.8 Other approaches for Bayesian classification and regression trees

Pratola like Wu et al. proposed new Metropolis-Hastings proposals for Bayesian regression trees for improving the convergence of the MCMC algorithm [108]. CGM, 2003, proposed Bayesian treed GLMs by extending CGM's Bayesian approach (1998) [91]. Gramacy and Lee developed Bayesian treed Gaussian process models for a continuous outcome by combining standard Gaussian processes with treed partitioning [109]. Other Bayesian approaches are also proposed for tree-based models that we mention in the references. Refer to the Refs. [110–112] for other Bayesian tree approaches of CGM. Also, Chipman et al. review advance models for Bayesian treed methods and refer to the Ref. [113]. For study about other tree-based Bayesian approach, refer to Refs. [114–118]. Also, Refs. [119, 120] are proposed Bayesian approaches for ensemble trees.

7. Criteria for determining the predictive performance of classification and regression trees

Predictive performance of classification tree models can compare using accuracy measures such as [17, 121]: sensitivity, specificity, false positive rate, false negative rate, positive predictive value, negative predictive value, positive likelihood ratio,

negative likelihood ratio, accuracy, Youden's index , diagnostic odds ratio (DOR), F-measure, and area under curve (AUC). Sensitivity, specificity, positive and negative predictive values, Youden's index, and accuracy have values between 0 and 1, and when these criteria are near to 1, then classification tree algorithm has better predictive performance. Also, false positive and false negative rates are between 0 and 1, and when these values are near to 0, then classification tree algorithm has better predictive performance. Classification tree models with positive likelihood ratio >10, negative likelihood ratio <0.1, and high diagnostic odds ratio have good predictive performance. AUC shows an overall performance measure and is between 0 and 1. Higher value shows an overall good performance measure, and a perfect diagnostic performance has an AUC equal to 1.

Predictive performance of regression tree algorithms can compare using criteria such as [122, 123]: Pearson correlation coefficient, root mean-squared error (RMSE), relative error (RE), mean error (ME), mean absolute errors (MAE), and bias.

8. Conclusion

Bayesian tree has some advantages in comparison to classic tree-based approaches. Classic CART model cannot explore the space of the tree fully and the result of tree is only locally optimal due to using greedy search algorithm. But Bayesian tree approaches investigate different tree structures with different splitting variables, splitting rules, and tree sizes, so these models can explore the tree space more than classic tree approaches. Indeed, Bayesian approaches are remedies for solving this problem of CART model. Also, CART is biased toward predictor variables with many distinct values, and Bayesian tree models can be a remedial for solving this problem. Because Bayesian approaches proposed by CGM, DMS, OML, and WTW utilize uniform distribution for selecting splitting node, splitting variables, and splitting rules, thus these approaches generate unbiased splits or have not any bias toward predictor variables with more splits. These approaches unlike classic tree approaches generate several trees that this advantage makes researchers to select the best tree based on study aim. Because in some studies, sensitivity is important for researcher and in other studies, specificity is important.

Some authors compared Bayesian approaches with classic tree approaches such as CART and random forests of Breiman and others models. Results of most papers indicated that Bayesian approach tends to present that the Bayesian method is superior to all other competitors. This can be for a variety of reasons: publication bias (methods that do not demonstrate superior performance typically do not get published), choice of examples that demonstrate superiority of their method, or more careful use of their method than the competing methods. Studies that may give more reliable comparisons would be ones in which there is no new method, and the paper is devoted to a comparison of existing approaches. For study about some of these papers, refer to Refs. [124–127].

According to empirical results, we can conclude that Bayesian approaches have better performance in comparison to classic CART model. Also, despite some advantages for Bayesian tree approaches in comparison with classic tree models, the number of published articles based on using Bayesian tree approaches for data analysis is low. One of the major reasons for this problem can be related to lack of user-friendly software and or need to have programming knowledge. On the other hand, the number of published papers based on employing CART model, random forests, and other classic tree models is many and one of the reasons for this frequency can be several software programs such as CART, SPSS, TANAGRA, STATISTICA, R, and WEKA.

Bayesian tree approaches need more research, because these approaches unlike CART and random forests cannot impute missing values. These approaches also cannot create linear combination splits like other tree algorithms (CART, QUEST, and CRUISE), even though interpretation of these splits is hard, but results indicated that tree methods with these splits have superior prediction accuracy in comparison to tree with univariate splits [128].

Author details

Amal Saki Malehi* and Mina Jahangiri
Faculty of Public Health, Department of Biostatistics and Epidemiology, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

*Address all correspondence to: amalsaki@gmail.com

References

[1] Agresti A. Categorical Data Analysis. Hoboken, NJ: John Wiley & Sons; 2003

[2] Huberty CJ, Olejnik S. Applied MANOVA and Discriminant Analysis. Hoboken, NJ: John Wiley & Sons; 2006

[3] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer Science & Business Media; 2006

[4] Moguerza JM, Muñoz A. Support vector machines with applications. Statistical Science. 2006;**21**(3):322-336

[5] Garson GD. Neural Networks: An Introductory Guide for Social Scientists. London: Sage; 1998

[6] Friedman JH, Roosen CB. An Introduction to Multivariate Adaptive Regression Splines. Thousand Oaks, CA: Sage Publications; 1995

[7] Duda RO, Hart PE, Stork DG. Pattern classification and scene analysis. New York: Wiley; 1973

[8] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning. 1997;**29**(2-3):131-163

[9] Hastie TJ. Generalized additive models. Statistical Models in S. Routledge; 2017. pp. 249-307

[10] De'ath G, Fabricius KE. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology. 2000;**81**(11):3178-3192

[11] Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Annals of Behavioral Medicine. 2003;**26**(3):172-181

[12] Speybroeck N, Berkvens D, Mfoukou-Ntsakala A, Aerts M, Hens N, Van Huylenbroeck G, et al. Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa. Agricultural Systems. 2004;**80**(2):133-149

[13] Marshall RJ. The use of classification and regression trees in clinical epidemiology. Journal of Clinical Epidemiology. 2001;**54**(6):603-609

[14] Nelson LM, Bloch DA, Longstreth W, Shi H. Recursive partitioning for the identification of disease risk subgroups: A case-control study of subarachnoid hemorrhage. Journal of Clinical Epidemiology. 1998;**51**(3):199-209

[15] Camp NJ, Slattery ML. Classification tree analysis: A statistical tool to investigate risk factor interactions with an example for colon cancer (United States). Cancer Causes & Control. 2002;**13**(9):813-823

[16] El-Solh AA, Sikka P, Ramadan F. Outcome of older patients with severe pneumonia predicted by recursive partitioning. Journal of the American Geriatrics Society. 2001;**49**(12):1614-1621

[17] Jahangiri M, Khodadi E, Rahim F, Saki N, Saki Malehi A. Decision-tree-based methods for differential diagnosis of β-thalassemia trait from iron deficiency anemia. Expert Systems. 2017;**34**(3):e12201

[18] Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Statistics in Medicine. 2015;**34**(11):1818-1833

[19] Li C, Glüer C-C, Eastell R, Felsenberg D, Reid DM, Roux C, et al. Tree-structured subgroup analysis of receiver operating characteristic

curves for diagnostic tests. Academic Radiology. 2012;**19**(12):1529-1536

[20] Zhang H, Singer BH. Recursive Partitioning and Applications. New York: Springer Science & Business Media; 2010

[21] Buntine W, Niblett T. A further comparison of splitting rules for decision-tree induction. Machine Learning. 1992;**8**(1):75-85

[22] De Mántaras RL. A distance-based attribute selection measure for decision tree induction. Machine Learning. 1991;**6**(1):81-92

[23] Friedman JH. A recursive partitioning decision rule for nonparametric classification. IEEE Transactions on Computers. 1977;**4**:404-408

[24] Rounds E. A combined nonparametric approach to feature selection and binary decision tree design. Pattern Recognition. 1980;**12**(5):313-317

[25] Ferri C, Flach P, Hernández-Orallo J, editors. Learning decision trees using the area under the ROC curve. ICML; 2002;**2**:139-146

[26] Mingers J. An empirical comparison of selection measures for decision-tree induction. Machine Learning. 1989;**3**(4):319-342

[27] Shih Y-S. Families of splitting criteria for classification trees. Statistics and Computing. 1999;**9**(4):309-315

[28] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993

[29] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. CRC press; 1984

[30] Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied Statistics. 1980;**29**(2):119-127

[31] Loh W-Y, Vanichsetakul N. Tree-structured classification via generalized discriminant analysis. Journal of the American Statistical Association. 1988;**83**(403):715-725

[32] Loh W-Y, Shih Y-S. Split selection methods for classification trees. Statistica Sinica. 1997:815-840

[33] Kim H, Loh W-Y. Classification trees with unbiased multiway splits. Journal of the American Statistical Association. 2001;**96**(454):589-604

[34] Loh W-Y. Improving the precision of classification trees. The Annals of Applied Statistics. 2009;**3**(4):1710-1737

[35] Kim H, Loh W-Y. Classification trees with bivariate linear discriminant node models. Journal of Computational and Graphical Statistics. 2003;**12**(3):512-530

[36] Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association. 1963;**58**(302):415-434

[37] Quinlan JR, editor. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence. World Scientific; 1992

[38] Loh W-Y. Regression tress with unbiased variable selection and interaction detection. Statistica Sinica. 2002;**12**:361-386

[39] Esposito F, Malerba D, Semeraro G, Kay J. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997;**19**(5):476-491

[40] Mingers J. An empirical comparison of pruning methods for decision

tree induction. Machine Learning. 1989;**4**(2):227-243

[41] Quinlan JR, Rivest RL. Inferring decision trees using the minimum description lenght principle. Information and Computation. 1989;**80**(3):227-248

[42] Mehta M, Agrawal R, Rissanen J, editors. SLIQ: A fast scalable classifier for data mining. In: International Conference on Extending Database Technology. Springer; 1996

[43] Shafer J, Agrawal R, Mehta M, editors. SPRINT: A scalable parallel classi er for data mining. In: Proc 1996 Int Conf Very Large Data Bases. Citeseer; 1996

[44] Gehrke J, Ramakrishnan R, Ganti V. RainForest—A framework for fast decision tree construction of large datasets. Data Mining and Knowledge Discovery. 2000;**4**(2-3):127-162

[45] Murthy SK, Kasif S, Salzberg S. A system for induction of oblique decision trees. Journal of Artificial Intelligence Research. 1994;**2**:1-32

[46] Holte RC. Very simple classification rules perform well on most commonly used datasets. Machine Learning. 1993;**11**(1):63-90

[47] Müller W, Wysotzki F. Automatic construction of decision trees for classification. Annals of Operations Research. 1994;**52**(4):231-247

[48] Müller W, Wysotzki F. The decision-tree algorithm CAL5 based on a statistical approach to its splitting algorithm. Machine Learning and Statistics: The Interface. 1997. pp. 45-65

[49] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics. 2006;**15**(3):651-674

[50] Messenger R, Mandell L. A modal search technique for predictive nominal scale multivariate analysis. Journal of the American Statistical Association. 1972;**67**(340):768-772

[51] Biggs D, De Ville B, Suen E. A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics. 1991;**18**(1):49-62

[52] Gray JB, Fan G. Classification tree analysis using TARGET. Computational Statistics and Data Analysis. 2008;**52**(3):1362-1372

[53] Quinlan JR. Induction of decision trees. Machine Learning. 1986;**1**(1):81-106

[54] Piccarreta R. Classification trees for ordinal variables. Computational Statistics. 2008;**23**(3):407-427

[55] Archer KJ. rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response. Journal of Statistical Software. 2010;**34**:7

[56] Galimberti G, Soffritti G, Maso MD. Classification trees for ordinal responses in R: The rpartScore package. Journal of Statistical Software. 2012;**47**(i10)

[57] Tutz G, Hechenbichler K. Aggregating classifiers with ordinal response structure. Journal of Statistical Computation and Simulation. 2005;**75**(5):391-408

[58] Kramer S, Widmer G, Pfahringer B, De Groeve M. Prediction of ordinal classes using regression trees. Fundamenta Informaticae. 2001;**47**(1-2):1-13

[59] Archer K, Mas V. Ordinal response prediction using bootstrap aggregation, with application to a high-throughput methylation data set. Statistics in Medicine. 2009;**28**(29):3597-3610

[60] Wheeler DC, Archer KJ, Burstyn I, Yu K, Stewart PA, Colt JS, et al. Comparison of ordinal and nominal classification trees to predict ordinal expert-based occupational exposure estimates in a case-control study. Annals of Occupational Hygiene. 2014;**59**(3):324-335

[61] Liu W, Chawla S, Cieslak DA, Chawla NV, editors. A robust decision tree algorithm for imbalanced data sets. In: Proceedings of the 2010 SIAM International Conference on Data Mining. SIAM; 2010

[62] Cieslak DA, Hoens TR, Chawla NV, Kegelmeyer WP. Hellinger distance decision trees are robust and skew-insensitive. Data Mining and Knowledge Discovery. 2012;**24**(1):136-158

[63] Ganganwar V. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering. 2012;**2**(4):42-47

[64] Yang L, Liu S, Tsoka S, Papageorgiou LG. A regression tree approach using mathematical programming. Expert Systems with Applications. 2017;**78**:347-357

[65] Su X, Wang M, Fan J. Maximum likelihood regression trees. Journal of Computational and Graphical Statistics. 2004;**13**(3):586-598

[66] Choi Y, Ahn H, Chen JJ. Regression trees for analysis of count data with extra Poisson variation. Computational Statistics and Data Analysis. 2005;**49**(3):893-915

[67] Therneau TM, Atkinson EJ. An Introduction to Recursive Partitioning using the RPART Routines. Technical Report 61. 1997. Available from: http://www.mayo.edu/hsr/techrpt/61.pdf

[68] Loh W-Y. Regression tree models for designed experiments. Institute of Mathematical Statistics. 2006. pp. 210-228

[69] Lee S-K, Jin S. Decision tree approaches for zero-inflated count data. Journal of Applied Statistics. 2006;**33**(8):853-865

[70] Wang Y, Witten IH. Induction of Model Trees for Predicting Continuous Classes; 1996

[71] Chan K-Y, Loh W-Y. LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. Journal of Computational and Graphical Statistics. 2004;**13**(4):826-852

[72] Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. Journal of Computational and Graphical Statistics. 2008;**17**(2):492-514

[73] Chaudhuri P, Lo W-D, Loh W-Y, Yang C-C. Generalized regression trees. Statistica Sinica. 1995;**5**:641-666

[74] Chaudhuri P, Huang M-C, Loh W-Y, Yao R. Piecewise-polynomial regression trees. Statistica Sinica. 1994;**4**:143-167

[75] Alexander WP, Grimshaw SD. Treed regression. Journal of Computational and Graphical Statistics. 1996;**5**(2):156-175

[76] Karalič A, editor. Employing linear regression in regression tree leaves. In: Proceedings of the 10th European Conference on Artificial Intelligence. John Wiley & Sons, Inc; 1992

[77] Landwehr N, Hall M, Frank E. Logistic model trees. Machine Learning. 2005;**59**(1-2):161-205

[78] Grubinger T, Zeileis A, Pfeiffer KP. evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. Working Papers in Economics and Statistics; 2011

[79] Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science. 2001;**16**(3):199-231

[80] Loh WY. Tree-structured classifiers. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;**2**(3):364-369

[81] Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011;**1**(1):14-23

[82] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[83] Breiman L. Bagging predictors. Machine Learning. 1996;**24**(2):123-140

[84] Freund Y, Schapire RE. Experiments with a New Boosting Algorithm. ICML. 1996;**96**:148-156

[85] Webb GI. Multiboosting: A technique for combining boosting and wagging. Machine Learning. 2000;**40**(2):159-196

[86] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics. 2000;**28**(2):337-407

[87] Buntine W. Learning classification trees. Statistics and Computing. 1992;**2**(2):63-73

[88] Denison DG, Mallick BK, Smith AF. A bayesian CART algorithm. Biometrika. 1998;**85**(2):363-377

[89] Chipman HA, George EI, McCulloch RE. Bayesian CART model search. Journal of the American Statistical Association. 1998;**93**(443):935-948

[90] Chipman H, McCulloch RE. Hierarchical priors for Bayesian CART shrinkage. Statistics and Computing. 2000;**10**(1):17-24

[91] Chipman H, George E, McCulloch R. Bayesian treed generalized linear models. Bayesian Statistics. 2003;**7**:323-349

[92] Chipman HA, George EI, McCulloch RE. Bayesian treed models. Machine Learning. 2002;**48**(1-3):299-320

[93] Moe WW, Chipman H, George EI, McCulloch RE. A Bayesian Treed Model of Online Purchasing Behavior Using in-Store Navigational Clickstream. Revising for 2nd Review at Journal of Marketing Research; 2002

[94] Pittman J, Huang E, Nevins J, Wang Q, West M. Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. Biostatistics. 2004;**5**(4):587-601

[95] Wu Y, Tjelmeland H, West M. Bayesian CART: Prior specification and posterior simulation. Journal of Computational and Graphical Statistics. 2007;**16**(1):44-66

[96] O'Leary RA. Informed Statistical Modelling of Habitat Suitability for Rare and Threatened Species (Doctoral dissertation, Queensland University of Technology). 2008

[97] O'Leary RA, Murray JV, Low Choy SJ, Mengersen KL. Expert elicitation for Bayesian classification trees. Journal of Applied Probability and Statistics. 2008;**3**(1):95-106

[98] Hu W, O'Leary RA, Mengersen K, Choy SL. Bayesian classification and regression trees for predicting incidence of cryptosporidiosis. PLoS One. 2011;**6**(8):e23903

[99] Quinlan JR, Compton PJ, Horn K, Lazarus L, editors. Inductive knowledge acquisition: A case study. In: Proceedings of the Second Australian

Conference on Applications of Expert Systems. Addison-Wesley Longman Publishing Co., Inc; 1987

[100] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;**82**(4):711-732

[101] Hastie T, Pregibon D. Shrinking trees. AT & T Bell Laboratories Technical Report; 1990

[102] Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation. 1997;**24**(01):38-49

[103] O'Leary R, Francis R, Carter K, Firth M, Kees U, de Klerk N. A Comparison of Bayesian Classification Trees and Random Forest to Identify Classifiers for Childhood Leukaemia. Proc. 18th World IMACS Congr. Int. Congr. Modell. Simul. Modell. Simul. Soc. Aust. NZ Int. Assoc. Math. Comput. Simul.(MODSIM09). 2009:4276-4282

[104] O'Leary RA, Choy SL, Murray JV, Kynn M, Denham R, Martin TG, et al. Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby *Petrogale penicillata*. Environmetrics. 2009;**20**(4):379-398

[105] Kynn M. Eliciting Expert Knowledge for Bayesian Logistic Regression in Species Habitat Modelling; 2005

[106] Denham R, Mengersen K. Geographically assisted expert elicitation for species' distribution models. Bayesian Analysis. 2007;**2**(1):99-136

[107] O'Leary R, Mengersen K, Murray J, Low Choy S, editors. Comparison of four expert elicitation methods: For Bayesian logistic regression and classification trees. In: 18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand; 2009

[108] Pratola MT. Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. Bayesian Analysis. 2016;**11**(3):885-911

[109] Gramacy RB, Lee HKH. Bayesian treed Gaussian process models with an application to computer modeling. Journal of the American Statistical Association. 2008;**103**(483):1119-1130

[110] Chipman HA, George EI, McCulloch RE. Bayesian ensemble learning. Advances in Neural Information Processing Systems. 2007;**19**:265

[111] Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. The Annals of Applied Statistics. 2010;**4**:266-298

[112] Pratola MT, Chipman HA, Gattiker JR, Higdon DM, McCulloch R, Rust WN. Parallel Bayesian additive regression trees. Journal of Computational and Graphical Statistics. 2014;**23**(3):830-852

[113] Chipman H, George EI, Gramacy RB, McCulloch R. Bayesian treed response surface models. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2013;**3**(4):298-305

[114] Angelopoulos N, Cussens J, editors. Tempering for Bayesian C&RT. In: Proceedings of the 22nd International Conference on Machine Learning. ACM; 2005

[115] Schetinin V, Fieldsend JE, Partridge D, Krzanowski WJ, Everson RM, Bailey TC, et al. The Bayesian Decision Tree

Technique with A Sweeping Strategy. arXiv preprint cs/0504042; 2005

[116] Lakshminarayanan B, Roy D, Teh YW, editors. Top-down particle filtering for Bayesian decision trees. In: International Conference on Machine Learning. 2013

[117] Lakshminarayanan B, Roy D, Teh YW, editors. Particle Gibbs for Bayesian additive regression trees. In: Artificial Intelligence and Statistics. 2015

[118] Taddy MA, Gramacy RB, Polson NG. Dynamic trees for learning and design. Journal of the American Statistical Association. 2011;**106**(493):109-123

[119] Duan LL, Clancy JP, Szczesniak RD. Bayesian ensemble trees (BET) for clustering and prediction in heterogeneous data. Journal of Computational and Graphical Statistics. 2016;**25**(3):748-761

[120] Quadrianto N, Ghahramani Z. A very simple safe-Bayesian random forest. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015;**37**(6):1297-1303

[121] Šimundić A-M. Measures of diagnostic accuracy: Basic definitions. Medical and biological Sciences. 2008;**22**(4):61-65

[122] Etemad-Shahidi A, Mahjoobi J. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Engineering. 2009;**36**(15-16):1175-1181

[123] Taghizadeh-Mehrjardi R, Dehghani S, Sahebjalal E. Comparison of multiple linear regression and regression tree for prediction of saturated hydraulic conductivity and macroscopic capillary length (Î±*). ProEnvironment/ ProMediu. 2013;**6**(13)

[124] Chen M, Cho J, Zhao H. Detecting epistatic SNPs associated with complex diseases via a Bayesian classification tree search method. Annals of Human Genetics. 2011;**75**(1):112-121

[125] Freeman AM, Lamon EC, Stow CA. Nutrient criteria for lakes, ponds, and reservoirs: A Bayesian TREED model approach. Ecological Modelling. 2009;**220**(5):630-639

[126] Lamon EC, Malve O, Pietiläinen O-P. Lake classification to enhance prediction of eutrophication endpoints in Finnish lakes. Environmental Modelling and Software. 2008;**23**(7):938-947

[127] Lamon EC, Stow CA. Bayesian methods for regional-scale eutrophication models. Water Research. 2004;**38**(11):2764-2774

[128] Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning. 2000;**40**(3):203-228