International Journal of Industrial Engineering, 30(2), 298-311, 2023

A NOVEL SPLIT SELECTION OF A LOGISTIC REGRESSION TREE FOR THE CLASSIFICATION OF DATA WITH HETEROGENEOUS SUBGROUPS

Sudong Lee[1] and Chi-Hyuck Jun[2, *]

[1]Department of Industrial Engineering
University of Ulsan
Ulsan, Korea

[2]Department of Industrial and Management Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, Korea
[*]Corresponding author's e-mail: chjun@postech.ac.kr

A logistic regression tree (LRT) is a hybrid machine learning method that combines a decision tree model and logistic regression models. An LRT recursively partitions the input data space through splitting and learns multiple logistic regression models optimized for each subpopulation. The split selection is a critical procedure for improving the predictive performance of the LRT. In this paper, we present a novel separability-based split selection method for the construction of an LRT. The separability measure, defined on the feature space of logistic regression models, evaluates the performance of potential child models without fitting, and the optimal split is selected based on the results. Heterogeneous subgroups that have different class-separating patterns can be identified in the split process when they exist in the data. In addition, we compare the performance of our proposed method with the benchmark algorithms through experiments on both synthetic and real-world datasets. The experimental results indicate the effectiveness and generality of our proposed method.

Keywords: Model Tree; Logistic Regression Tree; Subgroup Identification; Class Separability

(Received on November 24, 2022; Accepted on March 16, 2023)

1. INTRODUCTION

A model tree is a predictive machine learning method that incorporates parametric models into a decision tree. The tree structure iteratively partitions the input data space, and a set of parametric models at leaf nodes make predictions in each subspace. The model tree approach has two main advantages. First, a model tree can effectively handle a tradeoff between prediction accuracy and interpretability. In general, the best prediction is achieved by "black-box models," such as neural networks that are not easily interpretable. However, model interpretability is often the key to success in numerous research areas that require comprehensive human understanding. On the contrary, relatively interpretable models, such as decision trees and generalized linear models, have a limitation in the sufficient representation of complex relationships between the input and target variables. A model tree divides the input data into more homogeneous subgroups by the input variable values, and the model tree enables simple parametric models to adequately explain the partitioned data in each subgroup. In this way, the prediction performance of the parametric models can be improved while maintaining the ease of interpretation. The second advantage of model trees is subgroup identification. The real-world data often consist of subgroups with heterogeneous patterns (Liang *et al.*, 2020). For example, in the healthcare field, a dosing effect of a medicine on a particular disease may differ depending on the patients' gender or age group. As another example, in the marketing field, the optimal promotions can be different depending on the heterogeneous characteristics of the customer segmentation. Thus, it is essential to accurately identify the heterogeneous subgroups and reflect them in the prediction model. Model trees can address this challenge of accurately identifying the heterogeneous subgroups by partitioning the training data into subgroups according to the different patterns and learning the prediction models optimized for each. Due to such advantages of the model tree approach, model trees have been successfully used in various research areas, such as business (Bright *et al.*, 2017; Kuruzovich and Lu, 2017; Sankaranarayanan *et al.*, 2016), finance (Ben-David and Frank, 2009; Gerlein *et al.*, 2016), medical and health sciences (Anil *et al.*, 2017; Choi and Zeng, 2020; Di Leo *et al.*, 2017; Jo and Jun, 2021; Osmanovic *et al.*, 2017; Trincado *et al.*, 2016), social science (Cappelli *et al.*, 2019), bioinformatics (Chen *et al.*, 2016), natural language processing (Espina and Figueroa, 2017; Nozza *et al.*, 2016; Ravi and Ravi, 2017), and geosciences (Chen *et al.*, 2017; Heung *et al.*, 2017).

Logistic regression is one of the most popular models for classification. Logistic regression trees (LRTs) are a hybrid machine learning method that combines a decision tree model and logistic regression models. LRTs have the advantages of simplicity, interpretability, and accuracy compared with other model trees for classification using Naïve Bayes classifier (Kohavi, 1996), discriminant analysis (Loh and Shih, 1997; Kim and Loh, 2001; López-Chau *et al.*, 2013; Wickramarachchi *et al.*, 2016; Kim *et al.*, 2018), linear regression model (Frank *et al.*, 1998), and support vector machines (SVM) (Menkovski *et al.*, 2008; Madzarov *et al.*, 2009; Kumar and Gopal, 2010). Like other model trees, the performance of an LRT mainly depends on the model fitting and split selection. The model-fitting methods have been evolving through several algorithms. The early LRTs, such as the Smoothed and Unsmoothed Piecewise Polynomial Regression Trees (SUPPORT) algorithm (Chaudhuri *et al.*, 1995) and the Logistic Tree with Unbiased Selection (LOTUS) algorithm (Chan and Loh, 2004), build the logistic regression models in a tree only using a fraction of observations that belong to the corresponding node. This local learning scheme is not only unstable but also vulnerable to overfitting due to the small number of training samples. Landwehr *et al.* (2005) proposed an epoch-making algorithm called Logistic Model Tree (LMT), which employs boosting for incremental learning of LRTs. Lee and Jun (2018) improved the computational efficiency of LMT by applying least-angle regression (LAR) (Efron *et al.*, 2004) in the boosting process.

Unlike the advances in model fitting, there has been no such improvement in the split selection of LRTs, especially in capturing the heterogeneous class-separating patterns. The ideal split for a model tree is the partition that maximally improves the predictive performance of consequent models at the child nodes. The simplest approach is exhaustively fitting the child models for all possible candidate splits and selecting the best one. However, the computational cost for the repetitive model fitting is impractical. Thus, a split criterion that seeks an improvement in the split selection of LRTs without the exhaustive search is necessary. Technically, in the classification of model trees, we can improve the predictive performance of child models in two senses: class impurity minimization and subgroup identification. The traditional decision trees for classification, which have a constant prediction at each leaf node, calculate the class impurity of candidate splits and select the split that generates the most class-homogeneous child nodes. The Gini impurity of CART (Classification And Regression Tree) (Breiman *et al.*, 1984) and the information gain ratio of C4.5 (Quinlan, 2014) are typical examples of this approach. Most of the previous model trees use such class impurity measures for their split selection. The splits simply cut off a group of data that have the same class, regardless of how the parametric models separate the classes. Therefore, the splits do not consider how the model classifies the data.

On the contrary, some model trees employ statistical tests as an explicit split criterion for subgroup identification. The algorithms used for subgroup identification aim to detect the instability of the parametric models that occurs due to the different patterns in the subgroups. This criterion solely focuses on subgroup identification while ignoring the distribution of class labels. Therefore, if the statistical test fails to detect a significant model change by a split, the model tree will stop growing, even though the model tree can still improve its predictive performance by more splits in a direction to increase class purity. Thus, a model tree must consider both aspects to achieve the best predictive performance.

In this study, we present a novel split selection method for constructing an LRT. We define the class "separability," which indicates how well the classifier separates the data into different classes. The class separability measure is defined on the feature space of a logistic regression model, and it looks ahead the classification performance at the child nodes without model fitting. The proposed split selection process partitions the input data space into subpopulations with heterogeneous class-separating patterns in terms of both class impurity and parametric models. Thus, every intermediate node aims for the best split that maximally improves the consequent logistic regression models by efficient comparison of the candidate splits without an exhaustive search for all possible candidate splits. Experimental results on both synthetic and real-world datasets indicate that our proposed method effectively detects the heterogeneous subpopulations of the training data, and the method consequently yields a simple and accurate LRT compared with the benchmarking methods.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 introduces our proposed methods, including the class separability measure and the split selection rule. Section 4 describes our experimental study, followed by a discussion of the results. Finally, Section 5 concludes the paper with a summary and final remarks.

2. RELATED WORK

Here, we briefly review the existing LRTs and their split selection methods. The first algorithm that uses logistic regression models in a model tree is SUPPORT (Chaudhuri *et al.*, 1995). This algorithm learns a logistic regression model at each node and calculates pseudo-residuals based on the nearest-neighbor averaging. Then, the observations are grouped by the sign of the pseudo-residuals, and the input variable with the largest difference between the groups is chosen as the split variable. Another LRT learner that focuses on unbiased split variable selection based on a modified chi-squared test is LOTUS (Chan and Loh, 2004). Like SUPPORT, this algorithm adopts a local learning scheme that fits the node models only based on the isolated observations at each node. LMT (Landwehr *et al.*, 2005) has made a breakthrough in the model fitting process by

proposing a boosting method that incrementally learns logistic regression models. This approach enables the leaf models to reflect the global effect of surrounding the tree structure by incrementally updating models inherited from ancestor nodes. This process is called global learning. Furthermore, LMT implements a variable selection, whereas the previous algorithms only build full models using all the input variables. Lee and Jun (2018) proposed an LMT-L model that improves the computational efficiency of LMT by adopting LAR in the boosting process, called the LAR-Logistic algorithm.

The split selection methods of the LRTs can be divided into two groups. The first group uses a class impurity measure for the split criterion. The split criterion recursively partitions the input data space to be as class-homogeneous as possible. This approach is the most common approach that is widely used by not only traditional decision trees but also LRTs. LMT selects the splits using the information gain ratio criterion (Quinlan, 2014), which is one of the most popular class impurity measures. LOTUS selects the splits using a modified chi-squared test, which is another example of the class impurity measures. Similarly, SUPPORT divides the observations in a node into two groups based on the signs of their pseudo-residuals, which are identical to the class labels. Although these class impurity-based splits are intuitive and simple, they do not involve the classification models, as the measures are calculated based on the class label of training samples. The second group includes the methods of subgroup identification. The model-based recursive partitioning (MOB) algorithm (Zeileis *et al.*, 2008) is a representative subgroup identification algorithm for a parametric model that can be fitted using M-type estimators (e.g., the least-squares and maximum likelihood estimators). MOB examines the change in model parameters considering each split variable using a parameter instability test. Although MOB suggests a unified framework that embeds heterogeneous subgroup identification in model trees, it has several limitations. First, it only detects structural changes or parameter instability in parametric models. As such, if the statistical test fails to find a significant model change, the model tree will stop growing. However, as aforementioned, the model tree can still improve its predictive performance by more splits in a direction to increase class purity. Second, the proposed model tree by MOB divides the input variables into predictors and covariates. The predictors are used only for the parametric models, whereas the covariates are used only for splitting the tree. Because we have no prior knowledge of the input variables, it is advantageous to allow them for both the model and split.

3. PROPOSED METHOD

In this section, we propose a novel split selection method for constructing an LRT. First, we explain the class separability measure of logistic regression for a split evaluation. Second, we describe a split selection rule based on the proposed class separability measure. Finally, we propose an LRT algorithm, called as an LMT, using fast incremental learning and separability-based split selection (FS-LMT), which applies the proposed split rule to the LMT-L model in Lee and Jun (2018).

3.1 Class Separability Measure of Logistic Regression for a Split Evaluation

Logistic regression learns a linear separating hyperplane by formulating the odds ratio of class probability as a linear regression model. Given an observation of p input variables, $\boldsymbol{x} = \left[x_1, x_2, \dots, x_p\right]^T$, and the binary class label, $y \in \{-1, +1\}$, the predicted class probability $\hat{p} = P(y = 1)$ is calculated using (1):

$$\hat{p} = \frac{exp(x\hat{\boldsymbol{\beta}})}{1+exp(x\hat{\boldsymbol{\beta}})}, \tag{1}$$

where $\hat{\beta}$ denotes the estimated coefficient vector of the logistic regression model (the intercept term, β_0, is omitted for the sake of simplicity). Figure 1 presents the fitting curve of $\hat{p}$ versus $X\hat{\beta}$. The markers at the top and bottom indicate the training data points whose true label is +1 and -1, respectively. The estimated value of $\hat{p}$ is equal to 0.5 at the point where $X\hat{\beta} = 0$. The data points that satisfy $X\hat{\beta} < 0$ are classified as class -1 as their $\hat{p}$'s are less than 0.5. Conversely, the data points that lie on the right from the origin point, i.e., $X\hat{\beta} \geq 0$, are classified as class 1 as their $\hat{p}$'s are equal to or greater than 0.5. Figure 2 demonstrates how the logistic regression predicts the class of training data. The "x" markers indicate the observations that are incorrectly classified by the logistic regression model, whereas the "o" markers stand for the correctly classified observations. The incorrect predictions occur in the overlapped region of the markers at the top and bottom. Therefore, a split that removes the overlap can improve the predictive accuracy of the consequent logistic regression models. For example, the classification accuracy of a logistic regression model is significantly improved when the observations are split into two groups, namely, correctly predicted and incorrectly predicted, as presented in Figure 3. The "overlapping area" is the region that ranges from the leftmost observation at the top to the rightmost observation at the bottom. Intuitively, the class separability by logistic regression can be measured via the number of observations that are included in the overlapping area. The logistic regression model is unable to separate the observations of different classes in this region. As shown in the

example of Figure 3, a split can greatly improve the classification accuracy of the child models if it finds the partitions in which there are no overlapping areas on the feature space of logistic regression. Note that in defining the overlapping area, careful attention must be paid, as the area can possibly be overestimated by outliers. To address this problem, the confidence interval is defined as follows:

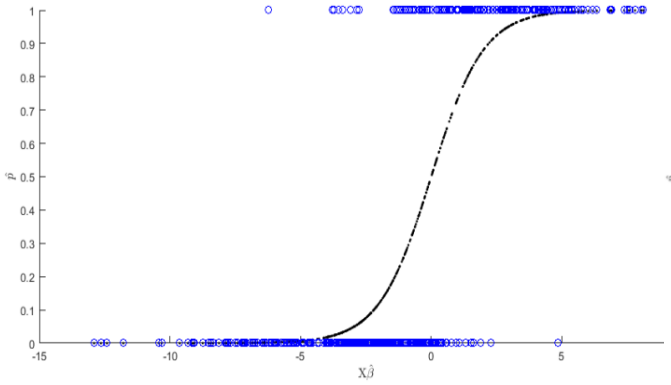


Figure 1. Feature space induced by logistic regression.

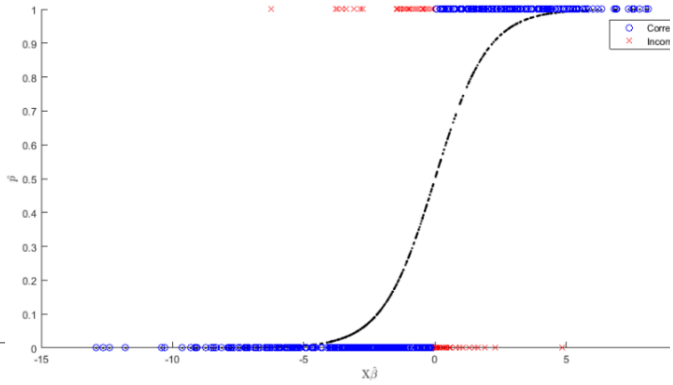


Figure 2. Prediction of class via logistic regression.

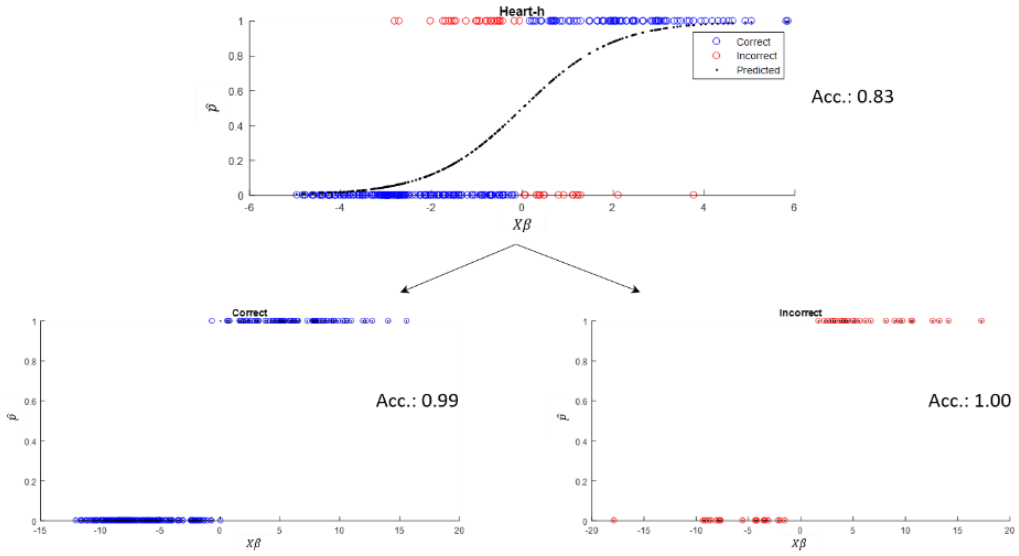


Figure 3. The optimal data split for logistic regression.

Definition 1. *Let $c \in \{-1, +1\}$ be the class for the true label, y. μ_c and σ_c denote the mean and the standard deviation of $\boldsymbol{X\beta}$ values for class c, respectively. The confidence interval of class c, CI_c is defined as*

$$CI_c = [l_c, u_c] = [\mu_c - k\sigma_c, \mu_c + k\sigma_c] \,. \tag{2}$$

The value of k can be determined by using Chebyshev's inequality, which is defined as

$$P(|X - \mu| < k\sigma) > 1 - \frac{1}{k^2} \,, \tag{3}$$

where X is a random variable of an arbitrary probability distribution, and μ, σ denote the mean and the standard deviation of X, respectively. Chebyshev's inequality holds with any probability distributions. In this research, the value of k is set to 2.236 for the confidence interval containing at least 80% of the observations. We define the overlapping area, OA, as the overlapped range of CI_c's:

Definition 2. *The overlapping area, OA, is defined as*

$$OA = [l_{+1}, u_{-1}] , \tag{4}$$

when $u_{-1} > l_{+1}$.

Figure 4 presents CI and OA for the "Australian Credit Approval Data Set" from the UCI machine learning repository (Dua and Graff, 2019). The observations included in OA are indicated in purple, whereas the outliers located out of *CI* are indicated in gray. If the outliers were included, the OA would be overestimated.

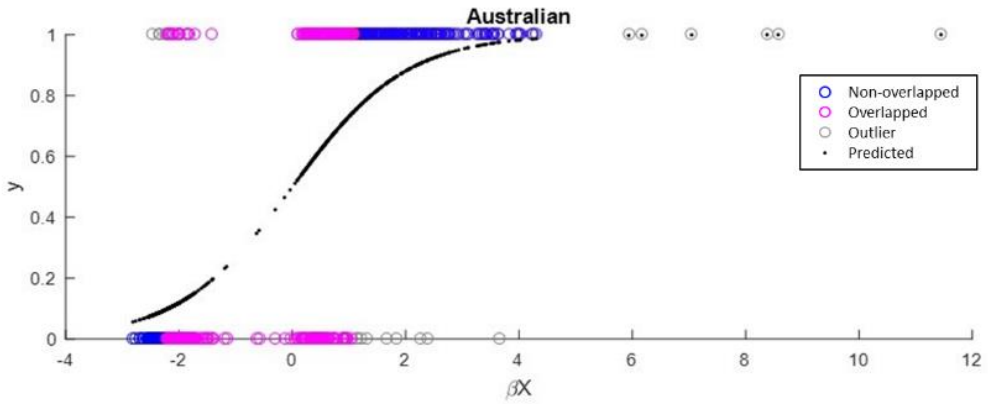


Figure 4. Illustration of CI and OA.

Now, we define the separability measure, SM, based on OA and CI_c.

Definition 3. *The separability measure, SM, is defined as*

$$SM = 1 - \frac{N_{OA}}{N_{CI_{+1}} + N_{CI_{-1}}} , \tag{5}$$

where N_{OA} and N_{CI_c} denote the number of observations located in OA and CI_c, respectively.

As aforementioned, the SM value measures how accurately the logistic regression model classifies the data. A higher value of SM implies that the logistic regression model more accurately separates the data of different classes. This implies that a split improves the predictive performance of the child models as much as the SM values are increased.

Proposition. SM is a normalized measure that satisfies $0 \leq SM \leq 1$.
Proof. Because OA is defined in the range of CA, then $\sum_c N_{CI_c} \geq N_{OA}$.

Therefore, $0 \leq SM = 1 - \frac{N_{OA}}{\sum_c N_{CI_c}} \leq 1$.

SM is a normalized measure that ranges from 0 to 1. When the SM value is 0, the logistic regression model is unable to classify the observations at all, whereas when the SM value is 1, the observations are perfectly classified. It should be noted that SM is equal to 1 when all samples in the current node have the same class. This property indicates the split by which the class of observations at each subspace is maximally homogeneous when no significant heterogeneity of class-separating patterns exists in the data. In other words, the separability-based split measure is not only able to detect the model heterogeneity when it exists but can also be used regardless of the existence of heterogeneous subpopulations in the given data, like the class-impurity-based measures. This property allows the proposed split selection method far more generality, and in Section 4, the property will be validated via a numerical experiment. While not discussed in this paper, SM could also be extended to multiclass classification through one-vs-one or one-vs-rest binary logistic regression models.

3.2 Split Selection Method

The proposed separability-based split selection method calculates SM for every candidate split of the current node and selects the best split with the largest SM value. First, a split rule to define candidate split rules is to be determined. Different approaches for split rules are divided to orthogonal, oblique, and nonlinear split. The orthogonal split approach divides the data space via an axis-parallel split. This approach is simple and intuitively interpretable. The oblique split approach statistically learns a hyperplane for a split, which is not necessarily orthogonal to the axis of the split variable. This flexibility of the oblique split rule leads to a smaller tree than the orthogonal split; however, the interpretability decreases because the split rule is represented by a linear function of multiple input variables. The nonlinear split approach is the most flexible approach, which requires the least assumptions for the split rule. A nonlinear split is represented by a linear combination of arbitrary basis functions. These functions require high computational complexity, and they are prone to noise in the training data. Our proposed method uses the orthogonal split approach to achieve interpretability. In the traditional decision trees, the simple orthogonal splits yield a large tree that is not only hard to interpret but also prone to overfitting. However, it does not necessarily happen in the proposed method as the logistic regression models at the leaves are much more complex than a constant prediction of the traditional decision trees.

At each intermediate node, the candidate split rules for the input variables are generated, and the best split is selected by the evaluation of SM. The orthogonal split rule for a numerical variable is given as a threshold value to the split variable. It groups the observations into two subsets: the value of the split variable of one subset is less than the threshold, whereas that of the other subset is greater than the threshold. The candidate threshold values, $\theta_{j,k}$ for a numerical variable X_j are calculated as an average of each adjacent pair of l distinct values observed in the training data as

$$\theta_{j,k} = \frac{v_{j,k}+v_{j,k+1}}{2}, \quad k = 1, \dots, l-1 , \tag{6}$$

where $v_{j,k}$ is the k^{th} smallest observed values of X_j. For a categorical split variable, the observations at the current node are partitioned according to the distinct values of the split variable.

The second step of split selection is the evaluation of the candidate splits. It is impractical to compare all candidate splits by fitting the child models because the split rule generates numerous candidates. Fortunately, our proposed separability measure, SM, can look ahead to the predictive performance of the logistic regression models in the child models without requiring model fitting. The weighted average of the SM values for the child nodes generated by the candidate split, θ, at the current node, t, is computed as

$$SM_\theta(t) = \frac{1}{N(t)} \sum_{t' \in S_\theta(t)} N(t')\, SM(t') , \tag{7}$$

where $N(t)$ is the number of samples at t, and $S_\theta(t)$ is the set of child nodes generated by θ. $SM_\theta(t)$ estimates the predictive performance of logistic regression models at the child nodes. The split with the largest $SM_\theta(t)$ is selected as the optimal split of t:

$$\theta^* = \arg\max_\theta SM_\theta(t) . \tag{8}$$

The node t splits only when $SM_{\theta^*}(t)$ is greater than $SM(t)$, otherwise, the tree stops growing.

3.3 The FS-LMT Algorithm

Here, we propose the FS-LMT algorithm. FS-LMT selects the splits using the separability-based measure and learns the parameters of logistic regression models via LAR-Logistic proposed by Lee and Jun (2018). Figure 5 presents the procedure of FS-LMT, which is described as follows:

- LAR-Logistic builds a logistic regression model at the root node. The number of boosting iterations is determined using the Akaike information criterion (Akaike, 1998) with maximum iterations of 200. The number of iterations identically applies to all the nodes in the tree.
- The root node finds a split using the proposed separability-based split selection algorithm. The child nodes fit the logistic regression models on the corresponding subsets of the data using LAR-Logistic. The boosting starts with the parameters inherited from the parent node.
- The splitting and model fitting continue in the same fashion until a node satisfies at least one of the stopping criteria.

- The CART cross-validation-based pruning algorithm (Breiman *et al.*, 1984) prunes the grown tree. As a result, the algorithm provides the optimal subtree.

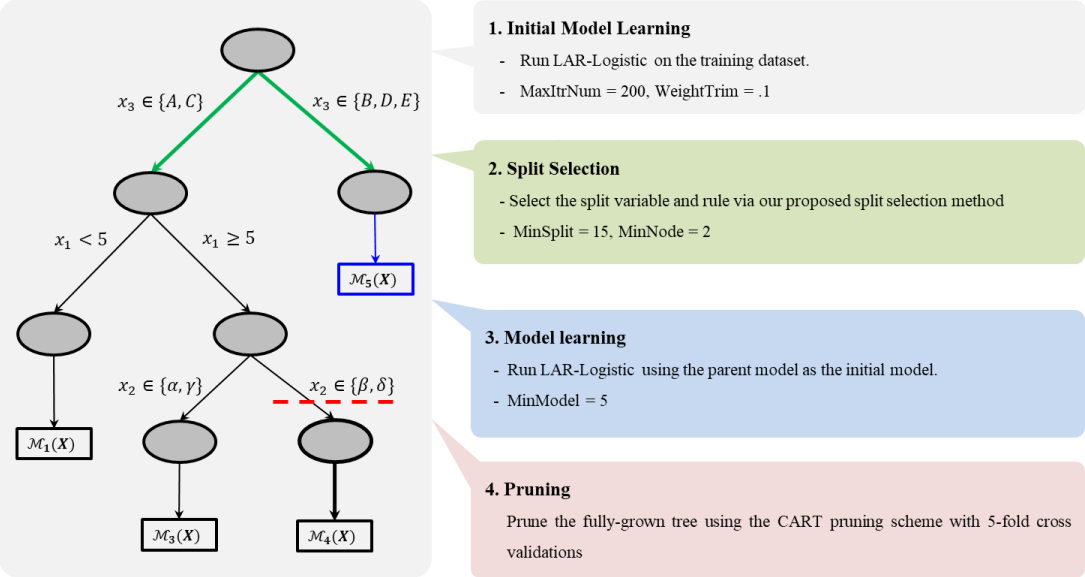


Figure 5. The procedure of the FS-LMT algorithm.

For the FS-LMT algorithm, several user-defined parameters need to be determined. The stopping criteria for pre-pruning were set to be the same as those used by Landwehr, Hall, and Frank (2005). Specifically, a node does not split if it satisfies either of the two conditions: 1) it contains fewer than 15 observations or 2) it does not have more than two subsets that contain more than two instances each. A logistic regression model is built at a node only if it contains at least five examples. For efficient model learning, we adopt the heuristics from Landwehr, Hall, and Frank (2005) and Sumner, Frank, and Hall (2005). The maximum number of boosting iterations is set to 200, and each iteration uses training samples with 90% of the total weight mass.

4. EXPERIMENTS

Here, we describe the experiments and discuss the results. First, we conducted experiments using simulated datasets. The experimental results indicate how the proposed method finds heterogeneous subgroups. Second, we performed experiments using real-world datasets. The experimental results evaluate the proposed method by comparing it with other benchmark classification methods.

4.1 Experiments Using Simulated Data

Figure 6 presents a two-dimensional (2D) XOR dataset. The shapes of the data points indicate two different classes: class 1 and class 2. The samples for class 1 were generated via random sampling of 100 samples from two different bivariate Gaussian distributions of $[X_1, X_2]$, $N([2.5, 2.5], \boldsymbol{I})$ and $N([7.5, 7.5], \boldsymbol{I})$, where $\boldsymbol{I}$ denotes the corresponding identity matrix. The samples for class 2 were generated in the same manner from $N([2.5, 7.5], \boldsymbol{I})$ and $N([7.5, 2.5], \boldsymbol{I})$. The XOR dataset is a typical example of heterogeneous data with different class-separating patterns in the subpopulations. The entire observations cannot be classified by a linear function; however, they are linearly separable when divided by the vertical (or horizontal) centerline. For this reason, linear classifiers, such as logistic regression, linear discriminant function, and linear SVM, fail to classify the XOR dataset. On the contrary, nonlinear classifiers, such as neural networks, kernel SVM, and Gaussian processes, may successfully classify the observations; however, they provide no meaningful interpretation of the heterogeneity for reasoning. To solve this problem, one can use the model tree approach that builds stratified predictive models in the subspaces. The key is to detect the heterogeneity and find proper partitions of the data.

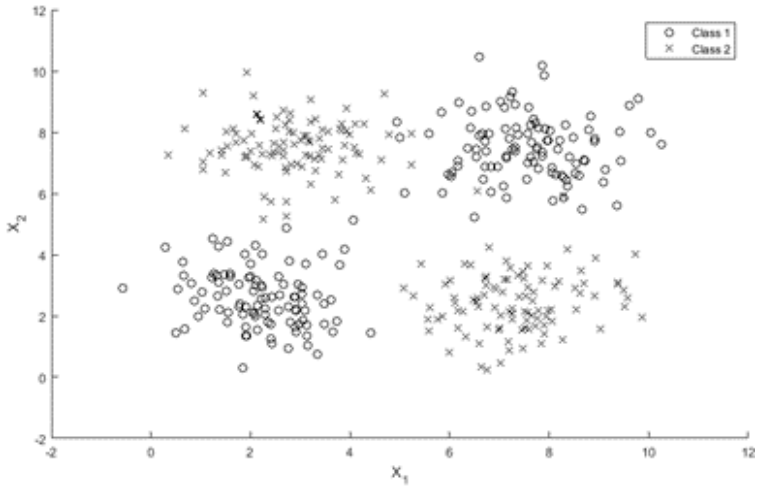


Figure 6. The 2D XOR dataset.

Figure 7 presents the trees constructed by LMT-L and FS-LMT for the 2D XOR dataset. LMT-L, which uses identical procedures with FS-LMT, except for the split selection, was included to validate the effectiveness of the separability-based split selection algorithm of FS-LMT. The plots for individual nodes illustrate the distribution of observations and the separating hyperplane (dashed line) of logistic regression. The prediction accuracy of logistic regression models in each group is given below the plots. Figure 7(a) demonstrates how the class-impurity-based split method partitions the 2D XOR data. The LMT-L algorithm, which uses the information gain ratio as the split criterion, takes off a small part of observations around the rim when it splits as it looks for a group of a homogeneous class, as presented in Figure 7(a). The final model fails to find the proper split for heterogeneity and leads to poor prediction accuracy. On the contrary, the FS-LMT algorithm, which employs the separability-based split selection algorithm, effectively selects the optimal splits, as presented in Figure 7(b). The SM value is maximized when the observations are divided according to whether the X_2 value is less than 5 or not. After a single split, the prediction accuracy reaches 0.97 and 1.00. These experimental results support the effectiveness of the separability-based split selection algorithm in the detection of heterogeneous class-separating patterns.

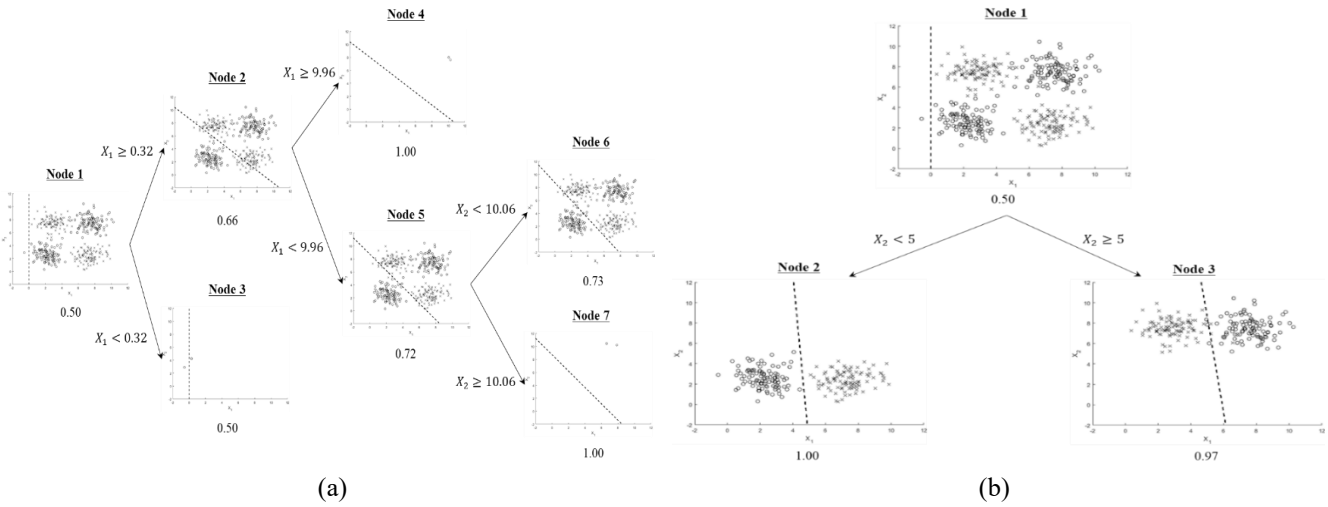


Figure 7. Classification results of the 2D XOR dataset:
The tree constructed by (a) the LMT-L algorithm and (b) the FS-LMT algorithm.

Figure 8 presents the experimental results on the 3D XOR datasets. The samples for class 1 were generated via random sampling of 100 samples from $[X_1, X_2] \sim N([7.5, 7.5], \boldsymbol{I})$ and $N([2.5, 2.5], \boldsymbol{I})$ while $X_3 = 1$ and $X_3 = 0$, respectively, where $\boldsymbol{I}$ denotes the corresponding identity matrix. The samples for class 2 were generated in the same manner from $[X_1, X_2] \sim N([2.5, 2.5], \boldsymbol{I})$ and $N([7.5, 7.5], \boldsymbol{I})$ while $X_3 = 1$ and $X_3 = 0$, respectively. As in the 2D XOR example, the

information gain ratio cannot successfully find a separating hyperplane of the 3D XOR data, whereas in the case of the proposed separability-based split selection, the accurate linear functions that perfectly classify classes 1 and 2 can be learned with only one split.

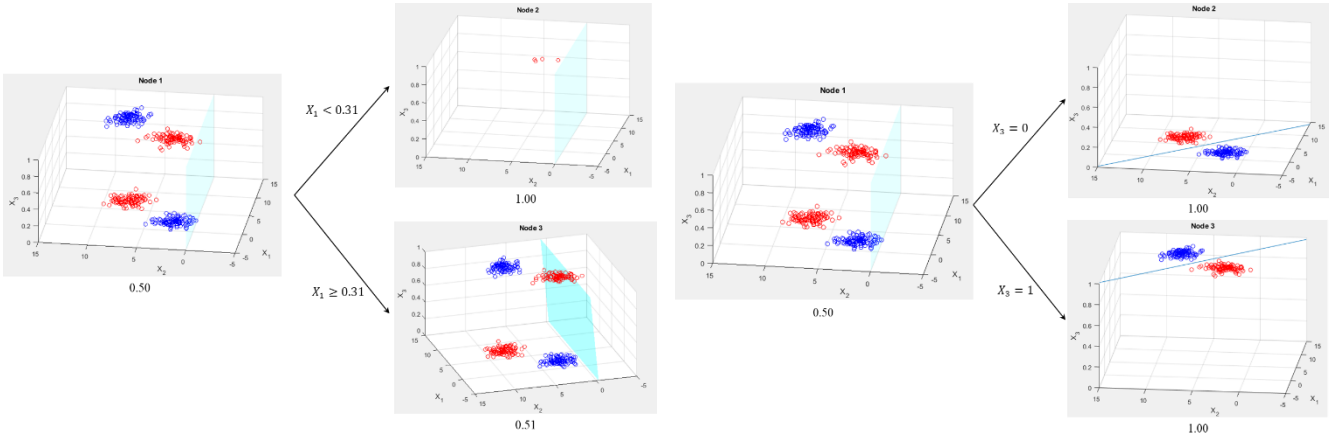


Figure 8. Classification results of the 3D XOR dataset:
Tree constructed by (a) the LMT-L algorithm and (b) the FS-LMT algorithm.

4.1 Experiments Using Real-World Data

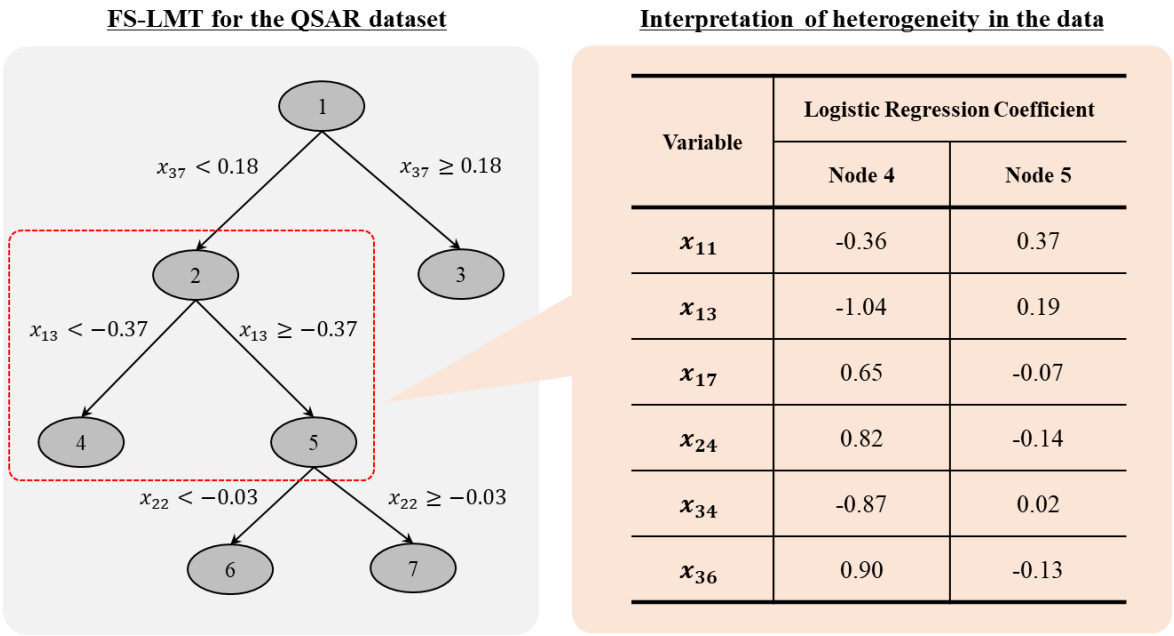
We apply the FS-LMT algorithm to the analysis of quantitative structure–activity relationships (QSAR). In chemical and biological sciences, QSAR models study the relationship between chemical structures and biological activities, and these models predict the characteristics of new chemicals (Huang, 2017). A predictor and the associated biological activity often have a different relationship according to the status of other variables, and it is crucial to reflect the complex interactions in the predictive model. Mansouri *et al.* (2013) used a given dataset to develop QSAR models for the study of the relationship between chemical structures and the biodegradation of molecules. In this study, we use the dataset employed by Mansouri *et al*. (2013). Biodegradation experimental values of 1,055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan. The predictors consist of 41 biodegradation experimental values, and the response class is whether the molecules are readily biodegradable or not.

We used several algorithms to the QSAR dataset for comparison. In addition to LMT-L, random forests (RF), SVM with Gaussian kernel (SVMg), and multilayer perceptron (MLP) were compared as the benchmark methods. The RF models were developed using 100 decision trees. The kernel parameter of SVMg was tuned via a grid search in a range of $[2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}]$. The MLP algorithm used two layers consisting of 10 hidden nodes of $tanh$ functions and one output node of the logistic function. The mean and standard deviation values for 10 repetitions of 10-fold cross-validations were calculated for these algorithms.

The experimental results are presented in Table 1. We observe that the FS-LMT algorithm obtains higher classification accuracy than the other methods. Figure 9 presents the tree constructed by the FS-LMT algorithm. The node models learned by FS-LMT are optimized in the corresponding subspaces. For example, the estimates for the logistic regression model coefficients at nodes 4 and 5 have different values; moreover, the coefficient estimates for $X_{11}, X_{13}, X_{17}, X_{24}, X_{34}$, and X_{36} in nodes 4 and 5 have opposite signs. This result clearly explains the different class-separating patterns. These final models and the competitive predictive accuracy of FS-LMT support that the FS-LMT algorithm successfully classifies the data with heterogeneous subgroups.

Table 1. Mean and standard deviation of the classification accuracy for FS-LMT, LMT-L, RF, SVMg, and MLP on the QSAR dataset.

Classifiers	FS-LMT	LMT-L	RF	SVMg	MLP
Avg. Acc	**0.8751**	0.8623	0.8653	0.8548	0.8423
Std.	0.0329	0.0313	0.0247	0.0274	0.0415



Variable	Logistic Regression Coefficient	
	Node 4	Node 5
x_{11}	-0.36	0.37
x_{13}	-1.04	0.19
x_{17}	0.65	-0.07
x_{24}	0.82	-0.14
x_{34}	-0.87	0.02
x_{36}	0.90	-0.13

Figure 9. The final predictive model constructed by FS-LMT for the QSAR dataset.

In addition, we applied FS-LMT to additional real-world datasets to evaluate the generality of the algorithm. The experiments compared the LMT, LMT-L, and FS-LMT algorithms to validate the effectiveness of the proposed split selection method. The two different versions of LMTs, namely, LMT using SimpleLogistic (LMT-S) and MultiLogistic (LMT-M)–by Landwehr, Hall, and Frank (2005), were used. Table 2 describes the 14 benchmark datasets collected from the UCI machine learning repository (Dua and Graff, 2019). The datasets have different characteristics with respect to the number of observations, proportion of missing values, and number of numerical, binary, and nominal (i.e., categorical variables with more than two distinct values) variables. We transformed each nominal variable by one-hot encoding when we built the logistic regression models.

Table 2. Description of the datasets used in the experiment.

Dataset	n	Numerical variables	Binary variables	Nominal variables	Sum of cardinality
Hepatitis	155	6	13	0	26
Sonar	208	60	0	0	0
Heart-statlog	270	6	3	4	19
Heart-h	294	6	3	4	19
Heart-c	303	6	3	4	19
Ionosphere	351	33	0	0	0
Horse-colic	368	7	3	12	54
Vote	435	0	16	0	32
Boston	506	12	1	0	2
Australian	690	6	4	4	37
Japan	690	6	4	5	40
Breast-w	699	9	0	0	0
Pima	768	8	0	0	0
German	1,000	7	2	11	54

Furthermore, we replaced missing values with the computed mean and mode for numerical and categorical variables, respectively. As aforementioned, the split selection method of FS-LMT seeks the partition by which the class of observations at each subspace is maximally homogeneous when there is no significant heterogeneity of class-separating patterns exists in

the data. In other words, the FS-LMT algorithm can be used regardless of the existence of heterogeneous subpopulations in the given data. Table 3 presents the comparison of the mean accuracy and standard deviation obtained via 10 repetitions of 10-fold cross-validations for LMT-S, LMT-M, LMT-L, and FS-LMT. The FS-LMT algorithm gives the highest predictive accuracy in a larger number of datasets than the LMT algorithms using the information gain ratio for the splits. These experimental results support that the FS-LMT algorithm is not only for data with explicit heterogeneity but also applicable to general datasets for classification.

Table 3. Mean classification accuracy and standard deviation for LMT-S, LMT-M, LMT-L and FS-LMT.

Dataset	LMT-S	LMT-M	LMT-L	FS-LMT
Hepatitis	0.819 (0.096)	0.824 (0.090)	**0.836** (0.088)	0.830 (0.110)
Sonar	0.754 (0.086)	0.729 (0.089)	0.757 (0.086)	**0.764** (0.093)
Heart-statlog	0.825 (0.073)	0.829 (0.073)	0.840 (0.071)	**0.844** (0.064)
Heart-h	0.822 (0.066)	0.830 (0.062)	0.835 (0.063)	**0.860** (0.025)
Heart-c	0.825 (0.072)	0.827 (0.079)	0.831 (0.076)	**0.841** (0.082)
Ionosphere	**0.922** (0.041)	0.879 (0.046)	0.899 (0.052)	0.894 (0.075)
Horse-colic	**0.835** (0.056)	0.794 (0.063)	0.818 (0.062)	0.818 (0.083)
Vote	0.955 (0.029)	0.951 (0.031)	0.954 (0.031)	**0.961** (0.031)
Boston	0.868 (0.047)	0.869 (0.047)	0.873 (0.047)	**0.889** (0.042)
Australian	0.852 (0.044)	0.844 (0.046)	**0.854** (0.045)	0.852 (0.032)
Japan	0.856 (0.040)	0.847 (0.040)	**0.857** (0.041)	0.848 (0.036)
Breast-w	0.965 (0.022)	0.961 (0.022)	0.966 (0.022)	**0.967** (0.016)
Pima	0.768 (0.044)	0.768 (0.047)	0.768 (0.045)	**0.772** (0.033)
German	0.740 (0.046)	0.746 (0.044)	0.747 (0.043)	**0.758** (0.047)
Average accuracy	0.843	0.836	0.845	**0.850**

5. CONCLUSION

In this paper, we proposed a novel split selection method for constructing an LRT. The separability measure, defined on the feature space of logistic regression models, evaluates the performance of potential child models without fitting, and the optimal split is selected based on the results. The splits detect heterogeneous subgroups that have different class-separating patterns when they exist in data. Otherwise, the split continues to improve the predictive performance of the model by finding the subgroups that minimize class impurity. Several experimental results on the synthetic and real-world datasets indicate that our proposed method cannot only efficiently find proper splits for an LRT with few splits but also effectively builds LRTs that accurately predict the classes of data. Moreover, the splits directly explain different class-separating patterns. Thus, it is easier to interpret the final models with heterogeneous representations of the class distribution.

ACKNOWLEDGEMENT

This work was supported by the research fund of University of Ulsan. (Title: Split selection of logistic regression trees for heterogeneous data classification)

REFERENCES

Akaike, H. (1998). Information Theory and An Extension of The Maximum Likelihood Principle. In. *Springer Series in Statistics*. Springer: New York, NY, (199–213).

Anil, R., Khanna, H., Keshavamurthy, A. S., Khanna, R., and Haswarey, A. (2017). Autonomous Learning Approach to Characterizing Motion Behavior. in Wireless Sensors and Sensor Networks (Wisnet), 2017 IEEE Topical Conference On, 49–52.

Ben-David, A. and Frank, E. (2009). Accuracy of Machine Learning Models Versus "Hand Crafted" Expert Systems–A Credit Scoring Case Study. *Expert Systems with Applications*, *36*(3): 5264–5271.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Taylor & Francis.

Bright, P., Ahmad, W., and Zahra, U. (2017). An Investigation into The Relationship of Strategic Planning Practices and Organizational Performance Using Advanced Data Mining Techniques. In. *Lecture Notes in Computer Science* Asian Conference on Intelligent Information and Database Systems. Springer: Cham, 676–687.

Cappelli, C., Simone, R., and Di Iorio, F. (2019). Cubremot: A Tool for Building Model-Based Trees for Ordinal Responses. *Expert Systems with Applications*, 124: 39-49.

Chan, K. Y. and Loh, W. Y. (2004). LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees. *Journal of Computational and Graphical Statistics*, 13(4): 826–852.

Chaudhuri, P., Huang, M. C., Loh, W. Y., and Yao, R. (1994). Piecewise-Polynomial Regression Trees. *Statistica Sinica*, 4(1): 143–167.

Chen, D., Tian, X., Zhou, B., and Gao, J. (2016). Profold: Protein Fold Classification with Additional Structural Features and A Novel Ensemble Classifier. *Biomed Research International*, *2016*. DOI: http://dx.doi.org/10.1155/2016/6802832.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., and Ma, J. (2017). A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. *CATENA*, 151: 147–160.

Choi, D. and Zeng, L. (2020). Robust Logistic Regression Tree for Subgroup Identification in Healthcare Outcome Modeling. *IISE Transactions on Healthcare Systems Engineering*, 10(3): 184–199.

Di Leo, G., Liguori, C., Paciello, V., Pietrosanto, A., and Sommella, P. (2017). I3dermoscopyapp: Hacking Melanoma Thanks to IoT Technologies. In Proceedings of The 50th Hawaii International Conference on System Sciences.

Dua, D., and Graff, C. (2019). U.C.I. Machine Learning Repository. University of California, School of Information and Computer Science: Irvine, CA. Retrieved from: http://archive.ics.uci.edu/ml.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, *32*(2), 407–499.

Espina, A. and Figueroa, A. (2017). Why Was This Asked? Automatically Recognizing Multiple Motivations Behind Community Question-Answering Questions. *Expert Systems with Applications*, *80*: 126–135.

Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H. (1998). Using Model Trees for Classification. *Machine Learning*, *32*(1): 63–76.

Gerlein, E. A., Mcginnity, M., Belatreche, A., and Coleman, S. (2016). Evaluating Machine Learning Classification for Financial Trading: An Empirical Approach. *Expert Systems with Applications*, 54: 193–207.

Heung, B., Hodúl, M., and Schmidt, M. G. (2017). Comparing The Use of Training Data Derived from Legacy Soil Pits and Soil Survey Polygons for Mapping Soil Classes. *Geoderma*, 290: 51–68.

Huang, H. (2017). Regression in Heterogeneous Problems. *Statistica Sinica*, 27(1): 71-88.

Jo, S., and Jun, C. H. (2021). A Regression Model Tree Algorithm by Multi-Task Learning. *Industrial Engineering & Management Systems*, 20(2): 163-171.

Kim, D., Park, S. H., and Baek, J. G. (2018). A Kernel Fisher Discriminant Analysis-Based Tree Ensemble Classifier: Kfda Forest. *International Journal of Industrial Engineering*, 25(5).

Kim, H., and Loh, W. Y. (2001). Classification Trees with Unbiased Multiway Splits. *Journal of The American Statistical Association*, *96*(454): 589–604.

Kohavi, R. (1996). Scaling Up The Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. *KDD*, 96: 202–207.

Kumar, M. A., and Gopal, M. (2010). A Hybrid SVM Based Decision Tree. *Pattern Recognition*, *43*(12): 3977–3987 .

Kuruzovich, J., and Lu, Y. (2017, January). Entrepreneurs' Activities on Social Media and Venture Financing. in Proceedings of The 50th Hawaii International Conference on System Sciences. Retrieved from: http://hdl.handle.net/10125/41390.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1–2), 161–205.

Lee, S., and Jun, C. H. (2018). Fast Incremental Learning of Logistic Model Tree Using Least Angle Regression. *Expert Systems with Applications*, *97*, 137-145.

Liang, B., Wu, P., Tong, X., and Qiu, Y. (2020). Regression and Subgroup Detection for Heterogeneous Samples. *Computational Statistics*, 1-26.

Loh, W. Y., and Shih, Y. S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*, *7*(4): 815–840.

López-Chau, A., Cervantes, J., López-García, L., and Lamont, F. G. (2013). Fisher's Decision Tree. *Expert Systems with Applications*, *40*(16): 6283–6291.

Madzarov, G., Gjorgjevikj, D., and Chorbev, I. (2009). A Multi-Class SVM Classifier Utilizing Binary Decision Tree. *Informatica*, *33*(2).

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., and Consonni, V. (2013). Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals. *Journal of Chemical Information and Modeling*, *53*(4): 867-878.

Menkovski, V., Christou, I. T., and Efremidis, S. (2008). Oblique Decision Trees Using Embedded Support Vector Machines in Classifier Ensembles. in *Proceedings of The Seventh IEEE International Conference on Cybernetic Intelligent Systems* (Pp. 1–6). IEEE.

Nozza, D., Fersini, E., and Messina, E. (2016, November). Deep Learning and Ensemble Methods for Domain Adaptation. in Tools with Artificial Intelligence (ICTAI) 28th International Conference On, *2016*. IEEE, 184–189.

Osmanović, A., Abdel-Ilah, L., Hodžić, A., Kevric, J., and Fojnica, A. (2017). Ovary Cancer Detection Using Decision Tree Classifiers Based on Historical Data of Ovary Cancer Patients. In. *IFMBE Proceedings* Proceedings of The International Conference on Medical and Biological Engineering (CMBEBIH). Springer: Singapore, 503–510.

Quinlan, J. R. (2014). *C4.5: Programs for Machine Learning*. Elsevier.

Ravi, K., and Ravi, V. (2017). A Novel Automatic Satire and Irony Detection Using Ensembled Feature Selection and Data Mining. *Knowledge-Based Systems*, *120*: 15–33.

Sankaranarayanan, H. B., Vishwanath, B. V., and Rathod, V. (2016, September). An Exploratory Analysis for Predicting Passenger Satisfaction at Global Hub Airports Using Logistic Model Trees. in *Research in Computational Intelligence and Communication Networks (ICRCICN), 2016 Second International Conference On*, 285–290.

Sumner, M., Frank, E., and Hall, M. (2005, October). Speeding Up Logistic Model Tree Induction. In. *Lecture Notes in Computer Science* European Conference on Principles of Data Mining and Knowledge Discovery. Springer: Berlin, Heidelberg, (675–683).

Trincado, J. L., Sebestyén, E., Pagés, A., and Eyras, E. (2016). The Prognostic Potential of Alternative Transcript Isoforms Across Human Tumors. *Genome Medicine*, *8*(1): 85.

Wickramarachchi, D. C., Robertson, B. L., Reale, M., Price, C. J., and Brown, J. (2016). HHCART: An Oblique Decision Tree. *Computational Statistics & Data Analysis*, *96*: 12–23.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2): 492-514.