Statistics and Its Interface Volume 9 (2016) 171–181

Energy bagging tree

Taoyun Cao, Xueqin Wang*, and Heping Zhang

This paper introduces Energy Bagging Tree (EBT) for multivariate nonparametric regression problems. The EBT makes use of a measure of dispersion constructed from a generalized Gini's mean difference as node impurity, and the tree split function therefore corresponds to the product of energy distance and descendants' proportions. As a nonparametric extension of the between-sample variation in the analysis of variance, this measure of dispersion serves well for EBT in understanding certain complex data. Extensive simulation studies indicate that EBT is highly competitive with existing regression tree methods. We also assess the performance of the EBT through a real data analysis on forest fires.

AMS 2000 subject classifications: Primary 62G08, 62H20; secondary 62P12.
Keywords and phrases: Multivariate nonparametric regression, Energy bagging tree, Energy distance, Generalized Gini's mean difference.

1. INTRODUCTION

Multivariate data, where the focus of multivariate refers to data with more than one response variable, have emerged in many scientific areas such as ecology, biometrics, econometrics, and medicine. It is often important to detect the relationships between multivariate response variables and explanatory variables. For instance, forest fires are a severe environmental issue that endangers human lives. It is important to timely assess the forest fire weather related factors and spatial location including rain, wind, temperature, relative humidity, x-axis spatial coordinate, y-axis spatial coordinate and examine the relationships with fuel moisture codes containing Fine Fuel Moisture Code, Duff Moisture Code, and Drought Code.

Multivariate regression trees (MRT), as a nonparametric data analysis tool, can explore the relationships between multivariate response variables and explanatory variables by building a tree-like model without assuming a specified relationship or a distribution for the response variables. This nonparametric regression method, as described by Death [6], is flexible for analyzing complex data, involving imbalance between covariates and nonlinear relationships between variables as well as high-order interactions, and its results are intuitive for interpretation. Consequently, it has become a practical and useful regression tool as a complement to the parametric regression models [6–12].

A number of impurity measures for MRT have been proposed in the literature comprising multivariate sums of squared deviations about the multivariate sample mean, the sums of squared pairwise dissimilarities, and the Manhattan distance [6]. The Mahalanobis distance is used as node impurity in [7]. Special cases and applications of MRT have been considered and presented by various authors [8–12].

However, as Zhang and Wang [13] pointed out, tree-based methods have two major limitations: tree structure can be unstable even with minor data perturbations, resulting in potentially unreliable prediction performance; and with a large number of variables and/or observations, one tree is either too complex or unlikely to summarize the essential information in the data. Bootstrapping and aggregating (Bagging) [1, 22] offered one option to alleviate these problems.

Bagging is proposed to improve the stability and accuracy of tree-based method used in nonparametric regression. It belongs to a broader class of ensemble methods through voting or model averaging [2–5]. So far, the bagging method has been developed primarily for a univariate response variable, however. The goal of this article is to extend bagging to the case with multivariate responses.

In this paper, we propose a novel bagging approach called Energy Bagging Tree (EBT) to nonparametric regression. EBT extends bagging by utilizing a measure of dispersion based on the generalized Gini's mean difference as the node impurity. This measure possesses useful properties for node splitting [14]. The split function turns out to be the product of energy distance with the descendants' proportions. And EBT reduces to bagging when there is only one response variable.

The energy distance was earlier introduced by Rizzo and Székely [15] as a nonparametric measure of the difference between two random variables. It can be used to compare two sets of multivariate data with arbitrary but same dimension, and hence can be used to test the heterogeneity (or homogeneity) of complex data.

The remainder of the paper is organized as follows: Section 2 introduces EBT, including impurity and split functions and generalization error. We reassure in Section 3 that bagging is a special case of EBT for a univariate response variable. Simulation studies are presented in Section 4, followed by a real data analysis. In our numerical examples, EBT is compared with other impurity functions for multivariate responses. We conclude with a few remarks in Section 5.

*Corresponding author.

2. ENERGY BAGGING TREE

2.1 Energy distance

Given two independent p-dimensional random variables X and Y with $\mathrm{E}\| X \|^\alpha < \infty$, $\mathrm{E}\| Y \|^\alpha < \infty$ for $\alpha \in (0, 2)$, where $\| \cdot \|$ denotes the Euclidean norm. The energy distance (with power α) between X and Y is defined as

$$\varepsilon_\alpha(X, Y) = 2E\|X - Y\|^\alpha - E\|X - X'\|^\alpha - E\|Y - Y'\|^\alpha,$$

where X' and Y' are the independent and identical random realizations of X and Y, respectively [15].

Specially, if X is uniformly distributed in the set $A = (a_1, ..., a_n)$ and Y is uniformly distributed in $B = (b_1, ..., b_m)$, let

$$g_\alpha(A, B) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \|a_i - b_j\|^\alpha, \tag{1}$$

be a generalized Gini's mean difference. Then, the energy distance of X and Y can be written as follows,

$$\varepsilon_\alpha(A, B) = 2g_\alpha(A, B) - g_\alpha(A, A) - g_\alpha(B, B). \tag{2}$$

Note that $\varepsilon_\alpha(A, B) = 0$ if and only if $A = B$, that is two sets A and B are exactly the same. $\varepsilon_\alpha(A, B)$ is an empirical energy distance between X and Y if A and B are the samples of X and Y, respectively.

2.2 Node impurity and split functions

Suppose there is a set of observations $A(\tau) = \{Y_j, j \in N(\tau)\}$ in node τ, and let $N(\tau)$ be the size of $A(\tau)$. We define the node impurity at τ as the generalized Gini's mean difference in equation (1),

$$i_\alpha(\tau) = g_\alpha(A(\tau), A(\tau)).$$

With a possible split s at τ, by the same token, $A(\tau_L)$ and $A(\tau_R)$ respectively denote the sets of observations in the left daughter node τ_L of node τ, and right daughter node τ_R of node τ, $N(\tau_L)$ and $N(\tau_R)$ denote their sizes and $p(\tau_L)$ and $p(\tau_R)$ denote their proportion in $A(\tau)$. Namely,

$$p(\tau_L) = \frac{N(\tau_L)}{N(\tau)},$$
$$p(\tau_R) = \frac{N(\tau_R)}{N(\tau)}.$$

$i_\alpha(\tau_L)$ and $i_\alpha(\tau_R)$ respectively denote the impurity function at two daughter nodes,

$$i_\alpha(\tau_L) = g_\alpha(A(\tau_L), A(\tau_L)),$$
$$i_\alpha(\tau_R) = g_\alpha(A(\tau_R), A(\tau_R)).$$

The split function is defined as

$$\phi_\alpha(s, \tau) = i_\alpha(\tau) - p(\tau_L) i_\alpha(\tau_L) - p(\tau_R) i_\alpha(\tau_R). \tag{3}$$

It should be noted that

$$i_\alpha(\tau) = g_\alpha(A(\tau), A(\tau)) = \frac{1}{N(\tau)^2} \sum_{i=1}^{N(\tau)} \sum_{j=1}^{N(\tau)} \|Y_i - Y_j\|^\alpha.$$

Thus

$$\begin{aligned}\phi_\alpha(s, \tau) = & \frac{1}{N(\tau)^2} \sum_{i=1}^{N(\tau)} \sum_{j=1}^{N(\tau)} \|Y_i - Y_j\|^\alpha \\ & - \frac{N(\tau_L)}{N(\tau)} \frac{1}{N(\tau_L)^2} \sum_{u=1}^{N(\tau_L)} \sum_{v=1}^{N(\tau_L)} \|Y_u - Y_v\|^\alpha \\ & - \frac{N(\tau_R)}{N(\tau)} \frac{1}{N(\tau_R)^2} \sum_{g=1}^{N(\tau_R)} \sum_{h=1}^{N(\tau_R)} \|Y_g - Y_h\|^\alpha,\end{aligned}$$

among

$$\begin{aligned}\sum_{i=1}^{N(\tau)} \sum_{j=1}^{N(\tau)} \|Y_i - Y_j\|^\alpha = & \sum_{u=1}^{N(\tau_L)} \sum_{v=1}^{N(\tau_L)} \|Y_u - Y_v\|^\alpha \\ + \sum_{g=1}^{N(\tau_R)} \sum_{h=1}^{N(\tau_R)} \|Y_g - Y_h\|^\alpha & + 2 \sum_{l=1}^{N(\tau_L)} \sum_{r=1}^{N(\tau_R)} \|Y_l - Y_r\|^\alpha.\end{aligned}$$

So we have obtained the following equation,

$$\phi_\alpha(s, \tau) = p(\tau_L) p(\tau_R) \varepsilon_\alpha(A(\tau_L), A(\tau_R)). \tag{4}$$

The equation (4) connects the split function in equation (3) with the energy distance defined in equation (2). By the properties of the energy distance deduced in [15], $\phi_\alpha(s, \tau)$ is a measure of between-node dispersion. If α takes the value of 2 and the response variable is univariate, $\phi_2(s, \tau)$ is proportional to the between-sample variation (sum of squared deviations from the mean) in the analysis of variance. However, if $\alpha = 1$, $\phi_1(s, \tau)$ differs from the between-sample dispersion induced by the sum of absolute deviations in [6] even for a univariate variable. Therefore, our proposed split function is not an extension for the measure of dispersion induced by L_α-norm when $0 < \alpha < 2$.

We should note that the split function is closely related to the energy distance, maximizing the split function is not equivalent to maximizing the energy distance, because the sizes of daughter nodes are not fixed. In fact, the splitting optimization tends to balance between the sizes of the daughter nodes and the energy distance.

Once the node impurity function and splitting criterion are defined, we can follow the same algorithm in [1] to grow trees.

2.3 Generalization error

After having the node impurity function and splitting criterion, it is important to assess how closely a result of predicted (function of the covariables or inputs) fits the data

(the outputs). This is now common practice in supervised learning problems, generalization error has been regard as an index for measuring. In the following we review it.

Given a sample $S = \{(X_i, Y_i), i = 1, \cdots, n\}$, where $X_i = (X_{i1}, \ldots, X_{iu})$ is the ith observed value of $X = (X_1, \ldots, X_u)$ and $Y_i = (Y_{i1}, \ldots, Y_{iv})^T$ is the ith observed value of $Y = (Y_1, \ldots, Y_v)^T$, $i = 1, \cdots, n$.

First, we draw B bootstrap samples from S. Each bootstrap sample is obtained by sampling with replacement. Second, we grow the bth multivariate regression tree T^{*b} from the bth bootstrap sample, $1 \le b \le B$. Let $\hat{\mu}^{*b}(\cdot)$ denote the bth prediction function for a new observation. A bagging prediction of Y with explanatory variables X is obtained by averaging the predictions:

$$\hat{\mu}_{bag}(X) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^{*b}(X).$$

We use the out-of-bag (OOB) samples to assess the accuracy of the above bagging prediction. For $(X_i, Y_i) \in S$, let O_i be the set of indices of b's such that the bth bootstrap sample does not contain (X_i, Y_i). The OOB prediction of Y_i is

$$\hat{\mu}_{OOB}(X_i) = \frac{1}{|O_i|}\sum_{b\in O_i}\hat{\mu}^{*b}(X_i),$$

here $|O_i|$ denotes the size of O_i.

Finally, we use the following generalization error as a measure of prediction accuracy:

$$PE_{bag} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{OOB}(X_i))^T(Y_i - \hat{\mu}_{OOB}(X_i)). \tag{5}$$

3. UNIVARIATE REGRESSION BAGGING

In this section, we relate EBT for $\alpha = 2$ to bagging method [1] where the sum of squares about the mean (SSM) is used as the impurity function.

Specifically, in bagging, the node impurity $i(\tau)$ at node τ was defined as follows:

$$i(\tau)_{ssm} = \sum_{i=1}^{N(\tau)}(Y_i - \bar{Y}(\tau))^2,$$

where $\bar{Y}$ is the average of Y_i's within node τ, $N(\tau)$ is the number of samples within node τ. And its split function was chosen as follows:

$$\phi(s,\tau) = i(\tau)_{ssm} - i(\tau_L)_{ssm} - i(\tau_R)_{ssm}, \tag{6}$$

where s is an allowable split, τ_L and τ_R are the left and right daughter nodes of node τ resulting from split s, respectively.

When $\alpha = 2$, the node impurity $i_2(\tau)$ for EBT is written as follows:

$$i_2(\tau) = \frac{1}{N(\tau)^2}\sum_{i=1}^{N(\tau)}\sum_{j=1}^{N(\tau)}(Y_i - Y_j)^2, \tag{7}$$

where Y_i and Y_j are within node τ, $N(\tau)$ is the number of samples within node τ as before. Equation (3) with $\alpha = 2$ is its corresponding split function. Namely,

$$\phi_2(s,\tau) = i_2(\tau) - p(\tau_L)i_2(\tau_L) - p(\tau_R)i_2(\tau_R). \tag{8}$$

It follows from equation (2.4) in [15] that

$$i_2(\tau) = \frac{2}{N(\tau)}i(\tau)_{ssm}. \tag{9}$$

Thus

$$\phi_2(s,\tau) = \frac{2}{N(\tau)}\phi(s,\tau). \tag{10}$$

Considering that $N(\tau)$ is a constant when splitting node τ, the impurity and split functions between EBT($\alpha = 2$) and bagging are in effect the same for univariate response variable.

4. MULTIVARIATE REGRESSION BAGGING

In this section, we report simulation experiments and a real data analysis to demonstrate the potential of EBT in tree construction for multivariate response variables. We compare its performance with that of existing MRT methods in the context of bagging: bagging(SAM), bagging(SSM) and bagging(SMD) in order to perform an unbiased comparison. The bagging(SAM) uses the sums of the Manhattan distance (the sums of absolute pairwise deviations) as the impurity functions [6], and bagging(SSM) uses the sums of squares about the mean as the impurity functions [6], and bagging(SMD) exploits the sums of Mahalanobis distance as the impurity function [7]. As presented in [6] and [7], equation

$$\phi(s,\tau) = i(\tau) - i(\tau_L) - i(\tau_R),$$

is the corresponding split function in bagging(SAM), bagging(SSM) and bagging(SMD).

Note that equations (9) and (10) still hold for multivariate responses, and that EBT($\alpha = 2$) and bagging(SSM) are equivalent. Therefore, we primarily compare the following four methods, EBT($\alpha = 1$), bagging(SAM), bagging(SSM) and bagging(SMD). Also notice that $\alpha = 1$ is the simplest choice within interval $\alpha \in (0, 2]$ for energy distance, which is why we consider EBT($\alpha = 1$).

4.1 General settings

To assess the performance, we consider the following three criteria similar to those in [11] and [18].

1. 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors. The generalization errors have emerged as one of the most commonly used indices to evaluate the predictive power of ensemble methods.
2. Mean of tree complexity is measured by the number of terminal nodes.
3. Frequency of a variable being selected in a tree.

Table 1. Distributions of predictor variables $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ used in example 1 and 2, where Ga denotes Gamma distribution with shape parameter $k = 2$ and scale parameter $\theta = 2$, three dependent structures among the X variables: independent, weakly dependent, strongly dependent

	Independent	Weakly dependent	Strongly dependent
X_1	Ga	$X_4 + X_5$	$X_4 + 0.1$ Ga
X_2	Ga	Ga	Ga
X_3	Ga	Ga	Ga
X_4	Ga	Ga	Ga
X_5	Ga	Ga	Ga
X_6	Ga	Ga	Ga

We simulated bivariate response variable $Y = (Y_1, Y_2)'$, and normalized the values of the response variables to have zero mean and unit variance. The sample size n is fixed to be 100, and $\epsilon = (\epsilon_1, \epsilon_2)'$ is a random error vector generated from the bivariate Cauchy distribution, the scale matrix $\sum = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, where correlation ρ is set to be 0, 0.5 and 0.8 to evaluate the impact of correlation of Y_1 and Y_2 on the method. All simulations are replicated 100 times.

4.2 Simulation study

Example 1. In this example, data are generated as follows: $X = (X_1, ..., X_6)$ are six predictor variables, Y_1, Y_2 are response variables, and the distributions of the X variables are given in Table 1. We consider three dependent structures among the X variables: independent, weakly dependent, strongly dependent. And,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} f(X) + \epsilon_1 \\ f(X) + \epsilon_2 \end{pmatrix},$$

where

(1.a): $f(X) = 2X_1 + 2X_2 + 2X_3$, or
(1.b): $f(X) = 5X_1X_2I(A_1) + \exp(-\sqrt{X_1^2 + X_2^2})I(A_2) + \sin(10\pi X_1X_2)I(A_3) + (X_1 + X_2)I(A_4)$.

Here $I(\cdot)$ is the indicator function. $A_1 = \{X_1 \leq 4, X_2 \leq 4\}$, $A_2 = \{X_1 > 4, X_2 \leq 4\}$, $A_3 = \{X_1 \leq 4, X_2 > 4\}$, and $A_4 = \{X_1 > 4, X_2 > 4\}$.

It is worth noting that we included some noise variables in the X that are irrelevant to the responses. Furthermore, the regression function is linear in model (1.a), and contains a piecewise function with interaction terms, exponential function, periodic and non-monotonous function terms in model (1.b).

Table 2 provides a summary from 100 simulation runs. We can see that EBT($\alpha = 1$) has advantage in balancing between Q_1, median, Q_3 quantiles of generalization errors and tree complexity. EBT($\alpha = 1$) owns the lowest degree of tree complexity with the loss of the prediction performance. Figure 1 displays the bar graphs of the frequencies when a variable is selected in a tree by the four methods in model (1.a). It suggests that when predictors X are independent, EBT($\alpha = 1$) has higher probability of selecting true predictors X_1, X_2 and X_3, and the lowest probability of selecting noise predictors X_4, X_5 and X_6. This is still the case for the weakly dependent X variables. With strongly dependent X variables, EBT($\alpha = 1$) maintains the lowest probability of selecting noise predictors X_5 and X_6, and relatively high probability of selecting the true predictors, especially for X_2 and X_3, while bagging(SAM) has the lowest probability of

Table 2. The 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors and the mean complexity in bagging in 100 replications for model (1.a)

Independent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.120	1.588	2.074	10.7
	bagging(SSM)	1.146	1.580	2.092	10.9
	bagging(SAM)	1.176	1.557	1.981	18.1
	bagging(SMD)	1.919	1.990	2.045	10.2
0.5	EBT($\alpha = 1$)	1.105	1.651	2.003	10.8
	bagging(SSM)	1.101	1.653	2.042	11.0
	bagging(SAM)	1.127	1.609	1.980	18.5
	bagging(SMD)	1.930	2.018	2.074	9.9
0.8	EBT($\alpha = 1$)	1.051	1.694	2.126	10.1
	bagging(SSM)	1.084	1.739	2.156	10.3
	bagging(SAM)	1.073	1.652	2.058	17.7
	bagging(SMD)	1.899	2.010	2.068	8.9
Weakly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.169	1.599	2.080	10.1
	bagging(SSM)	1.146	1.610	2.147	10.3
	bagging(SAM)	1.130	1.537	2.106	17.3
	bagging(SMD)	1.826	1.971	2.052	10.1
0.5	EBT($\alpha = 1$)	0.912	1.282	1.741	10.7
	bagging(SSM)	0.907	1.300	1.782	10.9
	bagging(SAM)	0.917	1.298	1.729	17.8
	bagging(SMD)	1.770	1.905	2.015	10.8
0.8	EBT($\alpha = 1$)	0.919	1.384	1.983	10.1
	bagging(SSM)	0.924	1.418	2.068	10.3
	bagging(SAM)	0.932	1.345	1.939	17.5
	bagging(SMD)	1.821	1.947	2.014	10.0
Strongly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.223	1.581	1.997	10.5
	bagging(SSM)	1.256	1.530	1.999	10.7
	bagging(SAM)	1.168	1.505	1.859	17.8
	bagging(SMD)	1.901	1.999	2.044	9.8
0.5	EBT($\alpha = 1$)	1.139	1.728	2.201	10.2
	bagging(SSM)	1.173	1.717	2.138	10.3
	bagging(SAM)	1.147	1.704	2.045	17.2
	bagging(SMD)	1.939	2.003	2.033	9.4
0.8	EBT($\alpha = 1$)	1.122	1.652	2.174	10.0
	bagging(SSM)	1.151	1.679	2.310	10.2
	bagging(SAM)	1.183	1.645	2.089	17.7
	bagging(SMD)	1.894	2.009	2.052	9.1

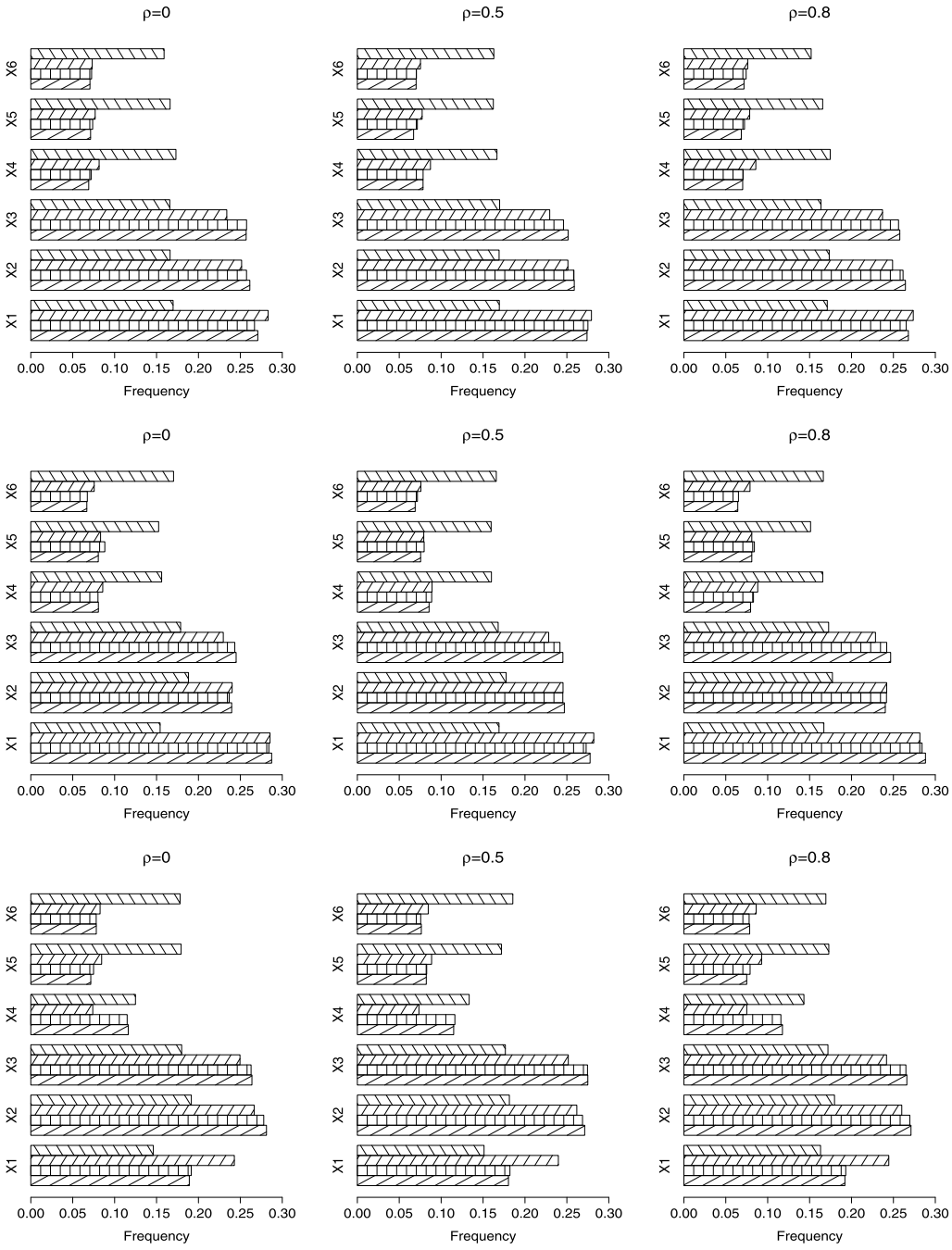

Figure 1. Frequencies of a variable selected by the four methods for model (1.a). Here in the rectangle oblique line (//) with $30°$*: EBT(*$\alpha = 1$*), vertical line (|): bagging(SSM), oblique line (/) with* $60°$*: bagging(SAM), backslash (\): bagging(SMD). The distributions of* X*'s and three dependent structures among the X variables are given in Table 1. Horizontal axis shows frequencies of a variable selected, vertical axis shows six predictor variables. The first line denotes independent structure among the X variables, weakly dependent structure and strongly dependent structure in the second line and third line, respectively. And each line shows three values of* ρ*: 0, 0.5 and 0.8.*

selecting X_4 in the condition of strongly dependent between X_1 and X_4.

Table 3 reveals that EBT($\alpha = 1$) is superior to the other methods when variables X are weakly dependent. It is clear that EBT($\alpha = 1$) owns the best prediction performance with the lowest degree of tree complexity when predictors X are weakly dependent. In addition to this, EBT($\alpha = 1$) still has advantage in balancing between Q_1, median, Q_3 quantiles

Table 3. The 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors and the mean complexity in bagging in 100 replications for model (1.b)

Independent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha=1$)	1.372	1.748	2.109	11.2
	bagging(SSM)	1.465	1.818	2.200	11.3
	bagging(SAM)	1.192	1.643	1.956	15.7
	bagging(SMD)	2.022	2.050	2.096	8.5
0.5	EBT($\alpha=1$)	1.115	1.466	2.016	11.4
	bagging(SSM)	1.171	1.523	2.042	11.7
	bagging(SAM)	1.071	1.332	1.816	15.9
	bagging(SMD)	2.014	2.048	2.093	8.3
0.8	EBT($\alpha=1$)	1.159	1.648	2.119	10.9
	bagging(SSM)	1.255	1.679	2.180	11.2
	bagging(SAM)	1.097	1.569	2.001	15.9
	bagging(SMD)	2.014	2.036	2.098	7.5
Weakly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha=1$)	1.278	1.646	2.077	8.4
	bagging(SSM)	1.360	1.813	2.168	8.9
	bagging(SAM)	1.359	1.769	2.070	12.7
	bagging(SMD)	1.972	2.032	2.097	9.2
0.5	EBT($\alpha=1$)	1.131	1.609	2.166	8.3
	bagging(SSM)	1.356	1.739	2.228	8.8
	bagging(SAM)	1.328	1.703	2.086	12.8
	bagging(SMD)	1.957	2.027	2.086	8.8
0.8	EBT($\alpha=1$)	1.100	1.506	2.055	8.3
	bagging(SSM)	1.234	1.682	2.105	8.8
	bagging(SAM)	1.195	1.687	2.057	13.1
	bagging(SMD)	1.898	2.027	2.072	9.0
Strongly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha=1$)	1.166	1.516	1.871	10.8
	bagging(SSM)	1.212	1.594	1.964	11.1
	bagging(SAM)	1.130	1.477	1.785	15.1
	bagging(SMD)	2.011	2.053	2.106	8.4
0.5	EBT($\alpha=1$)	0.993	1.425	1.799	10.8
	bagging(SSM)	1.055	1.441	1.922	11.1
	bagging(SAM)	0.930	1.319	1.787	15.1
	bagging(SMD)	1.998	2.041	2.073	8.5
0.8	EBT($\alpha=1$)	1.153	1.453	1.838	10.9
	bagging(SSM)	1.184	1.500	1.851	11.2
	bagging(SAM)	1.044	1.414	1.761	15.3
	bagging(SMD)	2.003	2.041	2.082	8.0

of generalization errors and tree complexity when predictors X are independent or strongly dependent. Figure 2 displays the bar graphs of the frequencies when a variable is selected in a tree by the four methods in model (1.b). EBT($\alpha=1$) has the highest probability of selecting true predictors X_1 and X_2 and the lowest probability of selecting noise predictors X_3, X_4, X_5, and X_6 in the case of X are weakly dependent. However, bagging(SAM) has the lowest probability of selecting X_4 in the condition of strongly dependent between X_1 and X_4 as in model (1.a). EBT($\alpha=1$) and bagging(SSM) work very well in selecting the true variables X_1 and X_2, meantime excluding the noise variables for this relatively complex regression function. We should remind that bagging(SSM) is equivalent to EBT($\alpha=2$).

Example 2. In the first example, the bivariate Y depends on a single regression function $f(X)$. In this example, distributions of predictor variables are used as before, we consider Y depending on two regression functions $g(X)$ and $h(X)$. That is,

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\begin{pmatrix} g(X)+\epsilon_1 \\ h(X)+\epsilon_2 \end{pmatrix}.$$

Here,

(2.a): $g(X)=\sin(\pi X_1X_2)-X_3, h(X)=5X_1X_2+2X_3$.

We see that $g(X)$ and $h(X)$ are not linear and include interaction and periodic terms.

As before, the advantage of EBT($\alpha=1$) in terms of balancing between Q_1, median, Q_3 quantiles of generalization errors and tree complexity are clear from Table 4 no matter what the dependent structures the variables X have. It should be noted that bagging(SAM) also has comparable performance with EBT($\alpha=1$) and bagging(SSM) based on Figure 3. Moreover Figure 3 confirms the favorable performance of the three methods relative to bagging(SMD) for model (2.a).

Example 3. In this example, we have constructed data similar as the regression function involving five variables in [25]. We first simulate $X_1, X_2, \ldots, X_{10}$ *i.i.d.* from Uniform(0,1). Then, the values of Y is defined as follows:

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\begin{pmatrix} f(X)+\epsilon_1 \\ 10f(X)+\epsilon_2 \end{pmatrix},$$

where

(3.a): $f(X)=10\sin(\pi X_1X_2)+20(X_3-0.5)^2+10X_4+5X_5$.

Note that $X_6, X_7, \ldots, X_{10}$ are noise variables. And linear terms, interaction terms, periodic terms and high-order terms appear in model (3.a).

Table 5 compares the performance of the four methods in this example, and EBT($\alpha=1$) is favorable to the other methods in terms of balancing between Q_1, median, Q_3 quantiles of generalization errors and tree complexity. It is clear from Figure 4 that EBT($\alpha=1$) can distinguish the true variables X_1, X_2, X_3, X_4, X_5 from the noise variables $X_6, X_7, X_8, X_9, X_{10}$.

In summary, through three simulation examples and four models, Tables 2–5 and Figures 1–4 demonstrate that EBT($\alpha=1$) performs better than the competing methods from the view of three criteria presented in Section 4.1. And EBT has an obvious advantage in distinguishing the true variables from the noise variables for all models in simulation study. It is of great importance for variable selection.

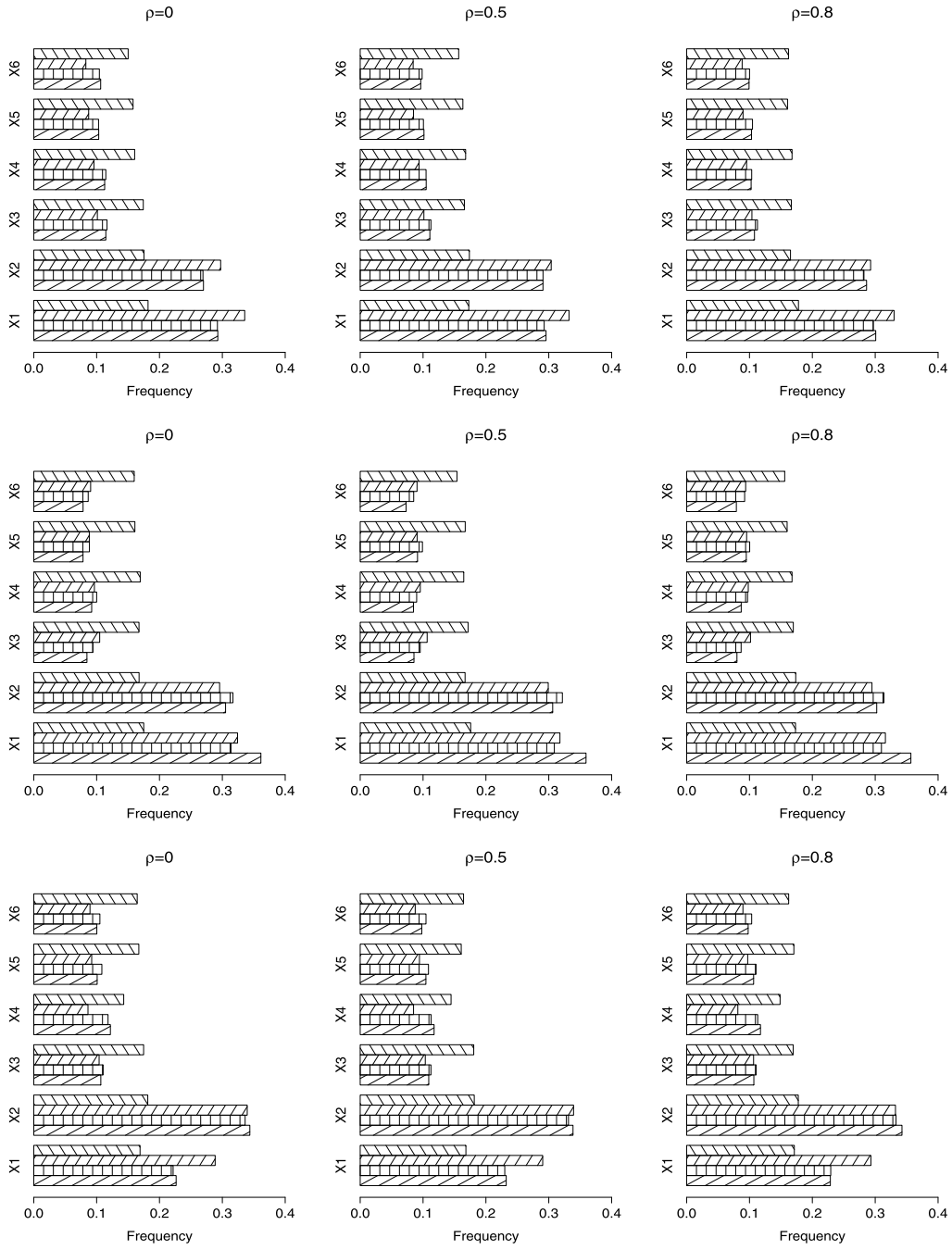

Figure 2. Frequencies of a variable selected by the four methods for model (1.b). Here in the rectangle oblique line (//) with $30°$*: EBT(*$\alpha = 1$*), vertical line (|): bagging(SSM), oblique line (//) with* $60°$*: bagging(SAM), backslash (\): bagging(SMD). The distributions of* X*'s and three dependent structures among the X variables are given in Table 1. Horizontal axis shows frequencies of a variable selected, vertical axis shows six predictor variables. The first line denotes independent structure among the X variables, weakly dependent structure and strongly dependent structure in the second line and third line, respectively. And each line shows three values of* ρ*: 0, 0.5 and 0.8.*

4.3 Case study

Forest fires are a severe environmental issue while endangering human lives. A fast detection is critical to control and prevent such a disaster.

We use the Forest Fires dataset with a total of 517 entries, which is available from UCI Machine Learning Repository. Paulo Cortez and Anibal Morais [16] used the data to predict the burned area of forest fires via Data Mining approach. In

Table 4. The 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors and the mean complexity in bagging in 100 replications for model (2.a)

Independent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.318	1.415	1.584	10.5
	bagging(SSM)	1.327	1.434	1.602	10.7
	bagging(SAM)	1.260	1.374	1.494	17.9
	bagging(SMD)	1.962	2.000	2.058	9.8
0.5	EBT($\alpha = 1$)	1.295	1.395	1.525	10.7
	bagging(SSM)	1.319	1.427	1.552	10.9
	bagging(SAM)	1.273	1.371	1.506	18.3
	bagging(SMD)	1.959	2.016	2.069	10.1
0.8	EBT($\alpha = 1$)	1.300	1.404	1.579	10.3
	bagging(SSM)	1.344	1.456	1.605	10.5
	bagging(SAM)	1.253	1.407	1.571	17.6
	bagging(SMD)	1.978	2.021	2.088	9.8
Weakly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.249	1.374	1.481	10.7
	bagging(SSM)	1.279	1.390	1.543	10.9
	bagging(SAM)	1.223	1.321	1.421	18.2
	bagging(SMD)	1.892	1.936	1.993	10.4
0.5	EBT($\alpha = 1$)	1.293	1.402	1.518	10.3
	bagging(SSM)	1.320	1.416	1.542	10.5
	bagging(SAM)	1.262	1.369	1.474	17.8
	bagging(SMD)	1.886	1.947	2.017	10.1
0.8	EBT($\alpha = 1$)	1.260	1.369	1.474	10.8
	bagging(SSM)	1.270	1.380	1.484	10.9
	bagging(SAM)	1.204	1.329	1.427	18.3
	bagging(SMD)	1.902	1.955	2.026	10.5
Strongly dependent X					
ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.276	1.382	1.555	10.4
	bagging(SSM)	1.313	1.436	1.598	10.6
	bagging(SAM)	1.242	1.349	1.464	17.9
	bagging(SMD)	1.936	2.000	2.059	9.8
0.5	EBT($\alpha = 1$)	1.240	1.369	1.524	10.4
	bagging(SSM)	1.278	1.409	1.584	10.6
	bagging(SAM)	1.208	1.327	1.469	17.8
	bagging(SMD)	1.929	1.976	2.031	10.1
0.8	EBT($\alpha = 1$)	1.256	1.354	1.520	10.6
	bagging(SSM)	1.286	1.412	1.564	10.8
	bagging(SAM)	1.225	1.351	1.488	18.1
	bagging(SMD)	1.911	1.969	2.030	10.2

this work, we mainly explore the relationships between fuel moisture codes and six predictors, where fuel moisture codes contain Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), and Drought Code (DC). The six predictors are rain, wind, temperature, relative humidity, x-axis spatial coordinate, y-axis spatial coordinate.

In our analysis, we made use of the resampling method. In every run, we chose 100 bootstrap samples, and replicated 100 times. Table 6 provides the three quantile results of the generalization errors. We can see that EBT($\alpha = 1$)'s generalization errors are relatively lower than the other methods.

Table 5. The 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors and the mean complexity in bagging in 100 replications for model (3.a)

ρ	method	Q_1	median	Q_3	Nodes
0	EBT($\alpha = 1$)	1.265	1.422	1.684	12.3
	bagging(SSM)	1.259	1.459	1.701	12.5
	bagging(SAM)	1.270	1.421	1.610	19.7
	bagging(SMD)	1.793	1.885	1.954	11.9
0.5	EBT($\alpha = 1$)	1.229	1.438	1.677	12.3
	bagging(SSM)	1.221	1.449	1.745	12.5
	bagging(SAM)	1.227	1.414	1.612	19.8
	bagging(SMD)	1.803	1.889	1.972	11.7
0.8	EBT($\alpha = 1$)	1.219	1.416	1.696	12.2
	bagging(SSM)	1.228	1.442	1.781	12.5
	bagging(SAM)	1.241	1.377	1.628	19.7
	bagging(SMD)	1.791	1.908	1.970	11.5

Table 6. The 25% (Q_1), 50% (median), 75% (Q_3) quantiles of generalization errors in Forest Fires data set in 100 replications

method	Q_1	median	Q_3
EBT($\alpha = 1$)	1.722	1.957	2.670
bagging(SSM)	1.693	1.982	2.791
bagging(SAM)	1.722	1.984	2.721
bagging(SMD)	2.258	2.539	3.206

Figure 5 suggests that temperature and relative humidity are vital for FFMC, DMC and DC. Not surprisingly, the selected variables are important weather conditions for forest fires.

5. DISCUSSION

Bagging is known to be effective in exploring complex data structures. However, the conventional bagging method is generally used for univariate response only. In this paper, we have attempted to generalize bagging method for handling multivariate responses by using generalized Gini's mean difference as node impurity in constructing a tree during bagging. As such, the node split function corresponds to an adjusted energy distance. It should be noted that the split function of EBT is not an extension to the measure of dispersion induced by L_α-norm, $0 < \alpha < 2$. The results from both simulation and real data analysis show that the proposed Energy Bagging Tree, EBT($\alpha = 1$) has its advantage than the existing MRT methods—bagging(SSM), bagging(SAM) and bagging(SMD)—when they are modified for bagging. We also noted that bagging(SSM) is equivalent to EBT($\alpha = 2$).

The advantage of bagging(SSM) lies in normal distribution for random error vector ϵ. Bagging(SAM) has advantage for analysis of complex data structure, such as eco-

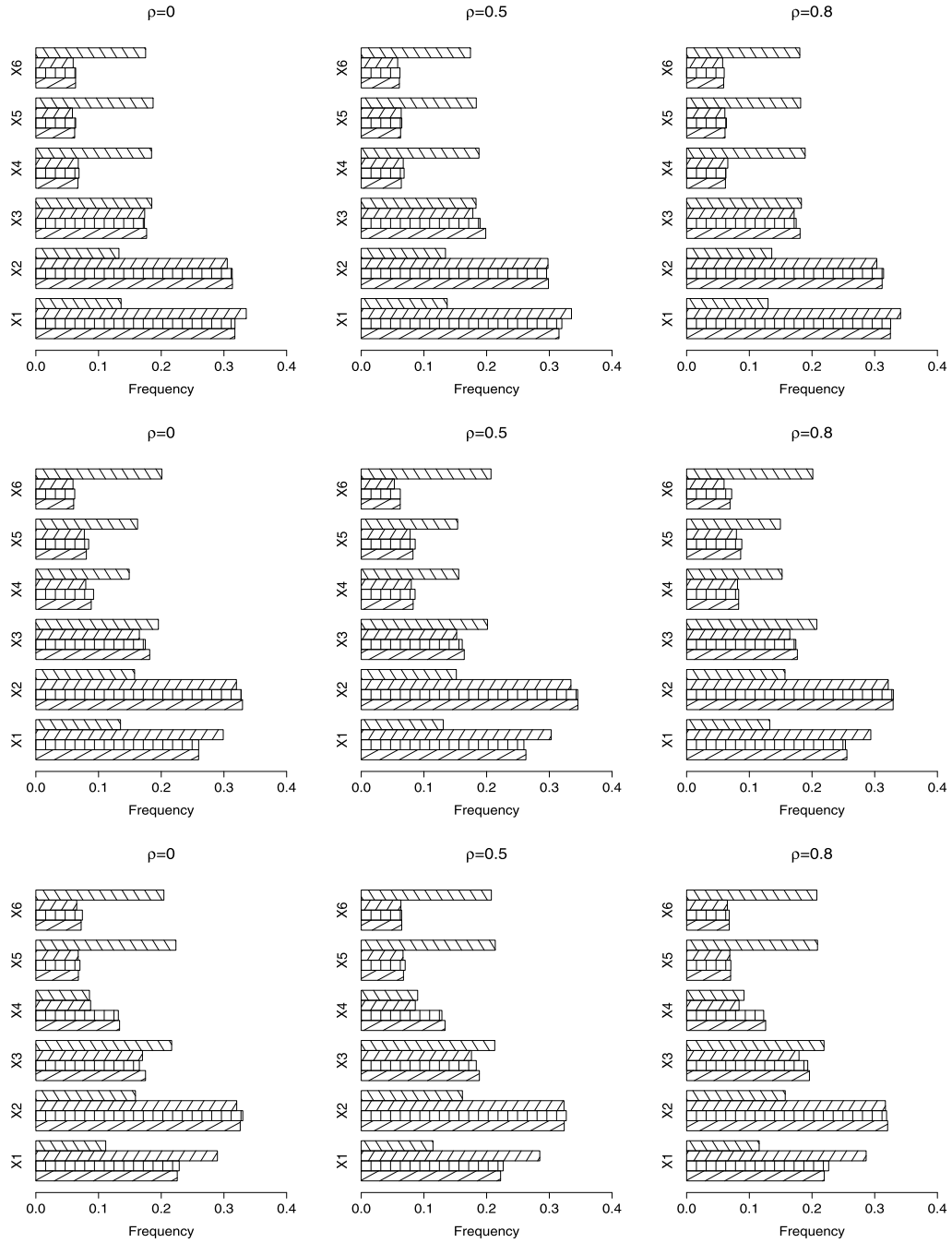

Figure 3. Frequencies of a variable selected by the four methods for model (2.a). Here in the rectangle oblique line (/) with $30°$*: EBT(*$\alpha = 1$*), vertical line (|): bagging(SSM), oblique line (/) with* $60°$*: bagging(SAM), backslash (\): bagging(SMD). The distributions of* X*'s and three dependent structures among the X variables are given in Table 1. Horizontal axis shows frequencies of a variable selected, vertical axis shows six predictor variables. The first line denotes independent structure among the X variables, weakly dependent structure and strongly dependent structure in the second line and third line, respectively. And each line shows three values of* ρ*: 0, 0.5 and 0.8.*

logical data with high-order and logarithm relationships between variables. And Bagging(SMD) has advantage for analysis of data with the simultaneous cooccurrence of several dependent variables, due to its impurity function is a variation of the approach of dealing with longitudinal data.

The main advantage of our proposed EBT is in two aspects: one is that it works for multivariate response variables

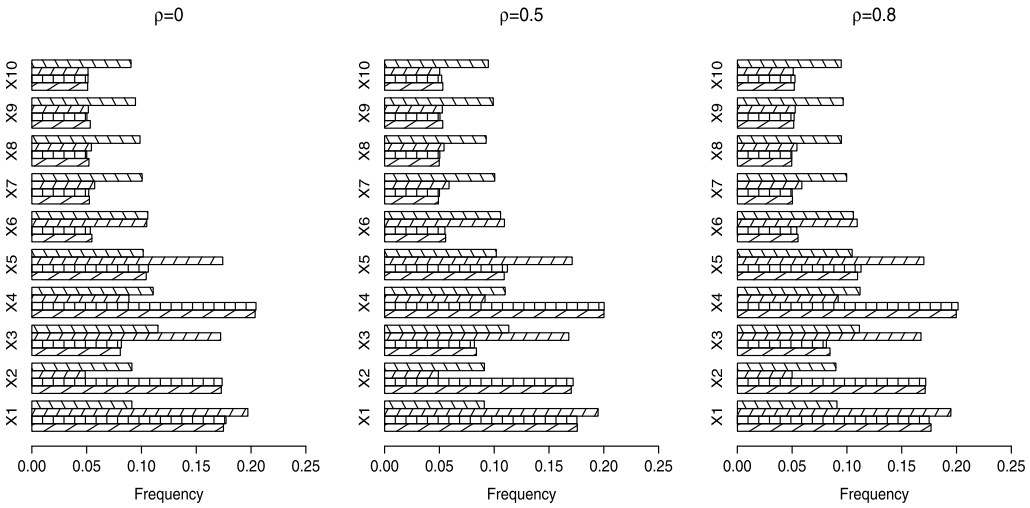

Figure 4. Frequencies of a variable selected by the four methods for model (3.a). Here in the rectangle oblique line (//) with $30°$: EBT($\alpha = 1$), vertical line (|): bagging(SSM), oblique line (//) with $60°$: bagging(SAM), backslash (\): bagging(SMD). The distributions of X's is from Uniform (0,1). Horizontal axis shows frequencies of a variable selected, vertical axis shows ten predictor variables. And three values of ρ: 0, 0.5 and 0.8.

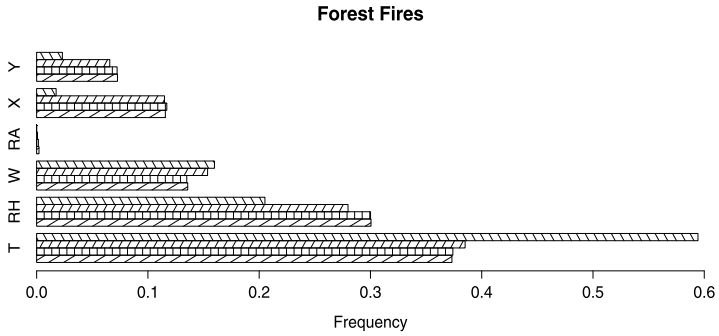

Figure 5. Frequencies of a variable selected by the four methods in the Forest Fires. Here in the rectangle oblique line (//) with $30°$: EBT($\alpha = 1$), vertical line (|): bagging(SSM), oblique line (//) with $60°$: bagging(SAM), backslash (\): bagging(SMD). Horizontal axis shows frequencies of a variable selected, vertical axis shows six predictor variables. Here T: temperature, RH: relative humidity, W: wind, RA: rain, X: x-axis spatial coordinate, Y: y-axis spatial coordinate.

and extends the scope of bagging tree. In the meantime, it covers bagging as a special case for univariate response variable and it covers multivariate sums of squared deviations about the multivariate sample mean [6] as a special case for multivariate response variables. The other is that it has the potential to be applied to variable selection because EBT has obvious advantage in distinguishing the true variables from the noise variables for four models in simulation studies.

Some issues deserve further study. For example, it may be interesting to explore whether we can choose α values to further improve the performance of EBT. We should note the limitation of our proposed EBT in that it deals with the quantitative response variables only. For multivariate discrete responses we need to further develop our method.

ACKNOWLEDGEMENT

Xueqin Wang's research is partially supported by NSFC for Excellent Young Scholar (11322108), NCET (12-0559), NSFC (11001280), RFDP (20110171110037). Heping Zhang's research is partially supported by grant R01DA016750-08 from the U.S. National Institute on Drug Abuse, a 1,000-plan scholarship from the Chinese Government, and the overseas and Hong Kong, Macau Young Scholars Collaborative Research Fund from the National Natural Science Foundation of China (11328103).

Received 21 January 2015

REFERENCES

[1] Breiman (1996). Bagging predictors. *Machine Learning*. **24** 123–140.

[2] Hothorn, T. and Lausen, B. (2003). Bagging tree classifiers for laser scanning images: a data- and simulation-based strategy. *Artificial Intelligence in Medicine*. **27** 65–79.
[3] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*. **40** 139–157.
[4] Hothorn, T. and Lausen, B. (2005). Bundling classifiers by bagging trees. *Computational Statistics and Data Analysis*. **49** 1068–1078. MR2143058
[5] Prasad, A. M., Iverson, L. R., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*. **9** 181–199.
[6] Death, G. (2002). Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*. **83** 1105–1117.
[7] Larsen, D. R. and Speckman, P. L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics*. **60** 543–549. MR2067013
[8] Davidson, T. A., Sayer, C. D., Perrow, M., Bramm, M., and Jeppesen, E. (2010). The simultaneous inference of zooplanktivorous fish and macrophyte density from sub-fossil cladoceran assemblages: a multivariate regression tree approach. *Freshwater Biology*. **55** 546–564.
[9] Hamann, A., Gylander, T., and Chen, P.-y. (2011). Developing seed zones and transfer guidelines with multivariate regression trees. *Tree Genetics Genomes*. **7** 399–408.
[10] Questier, F., Put, R., Coomans, D., Walczak, B., and Vander Heyden, Y. (2005). The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*. **76** 45–54.
[11] Hsiao, W.-C. and Shih, Y.-S. (2007). Splitting variable selection for multivariate regression trees. *Statist. Probab. Lett.* **77** 265–271. MR2339030
[12] Molinaro, A. M., Dudoit, S., and van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*. **90** 154–177. MR2064940
[13] Zhang, H. and Wang, M. (2009). Search for the smallest random forest. *Statistics and Its Interface*. **2** 381–396. MR2540095
[14] Yitzhaki, S. (2003). Gini's Mean difference: a superior measure of variability for non-normal distributions. *International Journal of Statistics*. **lxi** 285–316. MR2025523
[15] Rizzo, M. L. and Székely, G. J. (2010). Disco analysis: a nonparametric extension of analysis of variance. *The Annals of Applied Statistics*. **4** 1034–1055. MR2758432
[16] Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. Home page: http://www.dsi.uminho.pt/pcortez.
[17] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer, New York. MR2445017
[18] Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*. **7** 495–522. MR3086428
[19] Segal, M. R. (1992). Tree structured methods for longitudinal data. *J. Amer. Statist. Assoc.* **87** 407–418.
[20] Zhang, H. (1998). Classification trees for multiple binary responses. *J. Amer. Statist. Assoc.* **93** 180–193.
[21] Zhang, H. and Ye, Y. (2008). A tree-based method for modeling a multivariate ordinal response. *Stat. Interface*. **1** 169–178. MR2425353
[22] Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*. Springer, New York. MR2674991
[23] Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Data Mining and Knowledge Discovery*. **1** 80–87.
[24] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
[25] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*. **19** 1–67. MR1091842
[26] Ahmad, A. and Brown, G. (2014). Random projection random discretization ensembles: ensembles of linear multivariate decision trees. *IEEE Transactions on Knowledge and Data Engineering*. **26** 1225–1239.

Taoyun Cao
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: caotaoyun@126.com

Xueqin Wang
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China

Zhongshan School of Medicine
Sun Yat-Sen University
Guangzhou, 510080
China
E-mail address: wangxq88@mail.sysu.edu.cn

Heping Zhang
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China

Department of Biostatistics
Yale University School of Public Health
New Haven, CT 06520-8034
USA
E-mail address: heping.zhang@yale.edu