Statistics and Probability Letters 126 (2017) 114–118

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Generalized mixed effects regression trees

Ahlem Hajjem [a], Denis Larocque [b,*], François Bellavance [b]

[a] Department of Marketing, Université du Québec à Montréal, 320, rue Sainte-Catherine Est, Montréal (Québec), Canada, H2X 1L7
[b] Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal (Québec), Canada, H3T 2A7

ARTICLE INFO

Article history:
Received 7 July 2016
Received in revised form 20 February 2017
Accepted 21 February 2017
Available online 10 March 2017

Keywords:
Tree based methods
Clustered data
Mixed effects
Penalized quasi-likelihood algorithm
EM-algorithm

ABSTRACT

This paper presents generalized mixed effects regression trees, an extension of mixed effects regression trees to other types of outcomes. A simulation shows that the proposed method provides substantial improvements over standard trees when data are correlated.

1. Introduction

Tree based methods are classic data mining techniques. They have many advantages compared to parametric methods. For instance, they are able to detect automatically possible significant interactions between covariates, and they propose easily interpretable models that can be graphically displayed. They were extended to clustered and longitudinal data in Segal (1992), Abdollel et al. (2002), Eo and Cho (2013) and Loh and Zheng (2013). But these extensions do not allow observation-level (i.e., time-varying) covariates to be candidates in the splitting process and, consequently, (1) no random or subject-specific effect of these covariates is allowed, and (2) all repeated observations from a given subject cannot be split across different nodes. Hajjem et al. (2011) proposed a mixed effects regression tree (MERT) method. In contrast to the above extensions, MERT can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters since observation-level covariates are candidates in the splitting process. Sela and Simonoff (2012) and Fu and Simonoff (2014) independently proposed a similar approach, called random effects expectation–maximization (RE-EM) trees. MERT and RE-EM trees are designed for Gaussian response data. The main idea in both approaches is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence. An R (R Development Team, 2016) package is available for RE-EM trees. In a related problem, Bürgin and Ritschard (2015) proposed a tree-based method with varying coefficients for learning moderated relations with longitudinal data and ordinal responses. It is implemented in the R package `vcrpart`. Following the steps of the generalized linear mixed models (GLMMs), we propose a tree based method, named "generalized mixed effects regression tree" (GMERT), which is suitable for non-gaussian data (e.g., binary outcomes and count data). The proposed GMERT method can handle unbalanced clusters, and can incorporate observation-level covariates and their potential random effects. It allows observations within clusters to be split. This fact is crucial for the prediction performance, as will be shown in the simulation study.

[*] Corresponding author.
E-mail addresses: hajjem.ahlem@uqam.ca (A. Hajjem), denis.larocque@hec.ca (D. Larocque).

http://dx.doi.org/10.1016/j.spl.2017.02.033
0167-7152/© 2017 Elsevier B.V. All rights reserved.

2. Generalized mixed effects regression tree

The basic idea behind the proposed generalized mixed effects regression tree method is to replace the linear structure used to model the fixed effects component in the GLMM's linear predictor with a regression tree structure, while the random component is still represented using a linear structure as in GLMMs. For the estimation of the GMERT model, we use the penalized quasi-likelihood (PQL) method, and for the computation we use the expectation–maximization (EM) algorithm. Let $y_i = [y_{i1}, \ldots, y_{in_i}]^T$ denote the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $i = 1, \ldots, n$. Let $X_i = [x_{i1}, \ldots, x_{in_i}]^T$ denote the $n_i \times p$ matrix of fixed-effects covariates, and $Z_i = [z_{i1}, \ldots, z_{in_i}]^T$ denote the $n_i \times q$ matrix of random-effects covariates. Let b_i denote the $q \times 1$ unknown vector of random effects for cluster i. Then, conditional on the b_i, the GLMM assumes that the response vector y_i follows a distribution from the exponential family with density $f(y_i|b_i, \beta)$ where β is common for all the clusters and is the $p \times 1$ unknown vector of parameters for the fixed effects. The total number of observations is $N = \sum_{i=1}^{n} n_i$. Let $\mu_i = E(y_i|b_i)$ and $Cov(y_i|b_i) = \sigma^2 v_i(\mu_i)$, where σ^2 is a dispersion parameter that may or may not be known and $v_i(\mu_i) = diag[v(\mu_{i1}), \ldots, v(\mu_{in_i})]$ with $v(.)$ being a known variance function. Let $\eta_i = g(\mu_i)$ where $g(\mu_i) = [g(\mu_{i1}), \ldots, g(\mu_{in_i})]^T$ with $g(.)$ being a known link function. The GLMM is often written in the following form: $g(\mu_i) = \eta_i = X_i\beta + Z_ib_i$, $b_i \sim N(0, D)$, $i = 1, \ldots, n$, where D is the variance–covariance matrix of the random effects. A description of the PQL algorithm for GLMMs is detailed in Rodriguez (2008). The proposed generalized mixed effects regression tree (GMERT) model can be written as $\eta_i = f(X_i) + Z_ib_i$, $b_i \sim N(0, D)$, $i = 1, \ldots, n$, where all quantities are defined above except that the linear fixed part $X_i\beta$ is replaced by the function $f(X_i)$ that will be estimated with a standard regression tree model. Following the PQL approach, we can derive a MERT pseudo-model from the above GMERT model. More precisely, a first-order Taylor-series expansion yields the linearized response variable, $\tilde{y}_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$, and the MERT pseudo-model is defined as follows: $\tilde{y}_i = f(X_i) + Z_ib_i + e_i$. The GMERT algorithm is basically the PQL algorithm used to fit GLMMs where the weighted linear mixed effects (LME) pseudo-model is replaced by a weighted MERT pseudo-model. Consequently, the fixed-part $f(X_i)$ is estimated with a standard regression tree model. The GMERT algorithm is detailed below.

INITIALIZATION STEP. Set $M = 0$. Given initial estimates of the mean values, $\hat{\mu}_{ij}^{(0)}$, $j = 1, \ldots, n_i$, fit a weighted LME pseudo-model using the linearized pseudo responses, $\tilde{y}_i^{(0)} = g(\hat{\mu}_i^{(0)}) + (y_i - \hat{\mu}_i^{(0)})g'(\hat{\mu}_i^{(0)})$, and the weights, $W_i^{(0)} = diag(w_{ij}^{(0)})$ where $w_{ij}^{(0)} = (v_{ij}g'(\hat{\mu}_{ij}^{(0)})^2)^{-1}$. Set $m = 0$. Let $\hat{\sigma}_{(0)}^2$ and $\hat{D}_{(0)}$ be the estimates of this weighted LME pseudo-model.

OUTER LOOP. While non-convergence of $\hat{\eta}_i$, do:

INNER LOOP. While non-convergence of *GLL* (defined in point III below), set $m = m + 1$ and do:
I. Update $\hat{f}(X_i)$ and $\hat{b}_i$ using
(i) $\tilde{y}_{i(m)}^{*} = \tilde{y}_i^{(M)} - Z_i\hat{b}_{i(m-1)}$,
(ii) Let $\hat{f}_{(m)}(X_i)$ be an estimate of $f(X_i)$ obtained from a standard regression tree algorithm with $\tilde{y}_{i(m)}^{*}$ as responses, X_i as covariates, and W_i as weights, $i = 1, \ldots, n$,
(iii) $\hat{b}_{i(m)} = \hat{D}_{(m-1)}(W_i^{\frac{1}{2}(M)}Z_i)^T\hat{V}_{i(m-1)}^{-1}\left(W_i^{\frac{1}{2}(M)}\tilde{y}_i^{(M)} - W_i^{\frac{1}{2}(M)}\hat{f}_{(m)}(X_i)\right)$,
where $\hat{V}_{i(m-1)} = W_i^{\frac{1}{2}(M)}Z_i\hat{D}_{(m-1)}(W_i^{\frac{1}{2}(M)}Z_i)^T + \hat{\sigma}_{(m-1)}^2 I_{n_i}$, $i = 1, \ldots, n$.
II. Update $\hat{\sigma}^2$ and $\hat{D}$ using

$$\hat{\sigma}_{(m)}^2 = N^{-1}\sum_{i=1}^{n}\left\{\hat{\epsilon}_{i(m)}^T\hat{\epsilon}_{i(m)} + \hat{\sigma}_{(m-1)}^2[n_i - \hat{\sigma}_{(m-1)}^2\text{trace}(\hat{V}_{i(m-1)})]\right\},$$

$$\hat{D}_{(m)} = n^{-1}\sum_{i=1}^{n}\left\{\hat{b}_{i(m)}\hat{b}_{i(m)}^T + [\hat{D}_{(m-1)} - \hat{D}_{(m-1)}(W_i^{\frac{1}{2}(M)}Z_i)^T\hat{V}_{i(m-1)}^{-1}W_i^{\frac{1}{2}(M)}Z_i\hat{D}_{(m-1)}]\right\},$$

where $\hat{\epsilon}_{i(m)} = W_i^{\frac{1}{2}(M)}\tilde{y}_i^{(M)} - W_i^{\frac{1}{2}(M)}\hat{f}_{(m)}(X_i) - W_i^{\frac{1}{2}(M)}Z_i\hat{b}_{i(m)}$.
III. Update the generalized log-likelihood value using

$$GLL(f(X_i), b_i|y) = \sum_{i=1}^{n}\{\hat{\epsilon}_{i(m)}^T(\hat{\sigma}_{(m)}^2 I_{n_i})^{-1}\hat{\epsilon}_{i(m)} + \hat{b}_{i(m)}^T\hat{D}_{(m)}^{-1}\hat{b}_{i(m)} + \log|\hat{D}_{(m)}| + \log|\hat{\sigma}_{(m)}^2 I_{n_i}|\}.$$

UPDATING STEP. Set $M = M + 1$. Update $\hat{\eta}_i$, $\hat{\mu}_i$, $\tilde{y}_i$, w_{ij} and W_i, using $\hat{\eta}_i^{(M)} = \hat{f}_{(m)}(X_i) + Z_i\hat{b}_{i(m)}$, $\hat{\mu}_i^{(M)} = g^{-1}(\hat{\eta}_i^{(M)})$, $\tilde{y}_i^{(M)} = g(\hat{\mu}_i^{(M)}) + (y_i - \hat{\mu}_i^{(M)})g'(\hat{\mu}_i^{(M)})$, $w_{ij}^{(M)} = (v_{ij}g'(\hat{\mu}_{ij}^{(M)})^2)^{-1}$, $W_i^{(M)} = diag(w_{ij}^{(M)})$.

The GMERT model can be used to get the predicted response for two categories of new observations: (1) one that belongs to a cluster among those used to fit model (1), and (2) one that belongs to a cluster not included in the sample used to fit this model. To predict the response for a new observation from category 1, we use both its corresponding fixed component prediction and the predicted random part corresponding to its cluster. This is a cluster-specific estimate. For a new observation from category 2, we can only use its corresponding fixed component prediction (i.e., the random part is set to 0).

Table 1
Data generating processes (DGP) of the tree structure used for the simulation study with binary responses.

DGP	Data structure										
	Fixed component							Random component			
	Effect	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	Structure	Effect	d_{11}	d_{22}
1	Large	0.10	0.20	0.80	0.20	0.80	0.90	No random effect	–	0.00	0.00
2	Small	0.20	0.40	0.70	0.30	0.60	0.80				
3	Large	0.10	0.20	0.80	0.20	0.80	0.90	Random intercept	Small	4.00	0.00
4									Large	10.00	0.00
5	Small	0.20	0.40	0.70	0.30	0.60	0.80		Small	0.50	0.00
6									Large	4.00	0.00
7	Large	0.10	0.20	0.80	0.20	0.80	0.90	Random intercept and covariate	Small	2.00	0.05
8									Large	5.00	0.25
9	Small	0.20	0.40	0.70	0.30	0.60	0.80		Small	0.25	0.01
10									Large	2.00	0.05

To conclude this section, we detail a special case. For clustered data with a binary response variable, i.e., $y_{ij} = \mu_{ij} + \varepsilon_{ij}$ with $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma^2 v_{ij} = \sigma^2 \mu_{ij}(1 - \mu_{ij})$, the commonly used mixed effects logistic regression model with the logit link function is $\eta_{ij} = g(\mu_{ij}) = logit(\mu_{ij}) = \ln[\frac{\mu_{ij}}{1-\mu_{ij}}] = x_{ij}^T\beta + z_{ij}^T b_i$. The GMERT model in the binary response case and its corresponding MERT pseudo-model are respectively defined as follows: $\eta_{ij} = \ln[\frac{\mu_{ij}}{1-\mu_{ij}}] = f(x_{ij}) + z_{ij}^T b_i$, and $\tilde{y}_{ij} = \eta_{ij} + e_{ij}$, where $e_{ij} = (y_{ij} - \mu_{ij})g'(\mu_{ij}), g'(\mu_{ij}) = [\mu_{ij}(1-\mu_{ij})]^{-1}$, and $Var(e_{ij}) = \sigma^2[\mu_{ij}(1-\mu_{ij})]^{-1}$. The weights to be used in the GMERT algorithm are $w_{ij} = \mu_{ij}(1 - \mu_{ij})$.

3. Simulation

In this section, we investigate the performance of the GMERT method for a binary outcome in comparison to a standard classification tree. The GMERT method was implemented in R by means of a repeated call to the MERT algorithm (Hajjem et al., 2011), which uses the CART implementation in the `rpart` package (Therneau and Atkinson, 1997). In the initialization step of GMERT, we used $\hat{\mu}_{ij}^{(0)} = .25$ if $y_{ij} = 0$ and $\hat{\mu}_{ij}^{(0)} = .75$ if $y_{ij} = 1$. We set to five the maximum depth of any node of the final tree, to 50 the minimum number of observations that must exist in a node in order for a split to be attempted, and to 10 the minimum number of observations in any terminal node. The largest tree is grown then pruned automatically based on minimum ten-folds cross-validated error. The simulation design used has a hierarchical structure of 100 clusters with 60 observations each. The first ten observations in each cluster form the training sample, and the other 50 observations are left for the test sample. Consequently, models are built from 1000 observations (100 clusters of 10 observations). The fixed part of the outcome, $f(x_{ij})$, is generated based on a tree structure. Eight random variables, X_1 to X_8, independent and uniformly distributed in the interval [0, 10] are generated. Only the first five are used as predictors. The conditional or cluster-specific expectation (conditional probability of success), μ_{ij}, is generated based on a tree model with six terminal nodes given by $\mu_{ij} = g^{-1}(g(\varphi) + z_{ij}^T b_i)$, where g is the logit link function and $\varphi = \varphi_1 I(X_1 \leq 5, X_2 \leq 5) + \varphi_2 I(X_1 \leq 5, X_2 > 5, X_4 \leq 5) + \varphi_3 I(X_1 \leq 5, X_2 > 5, X_4 > 5) + \varphi_4 I(X_1 > 5, X_3 \leq 5, X_5 \leq 5) + \varphi_5 I(X_1 > 5, X_3 \leq 5, X_5 > 5) + \varphi_6 I(X_1 > 5, X_3 > 5)$. The values φ_1 to φ_6 are the typical probabilities (i.e., probability of success when the random effects b_i equal zero), and $b_i \sim N(0, D)$, for $i = 1, \ldots, 100, j = 1, \ldots, 60$. The binary response values y_{ij} are generated according to a Bernoulli distribution with probability μ_{ij}. Two different scenarios are selected for the fixed components (see Table 1). In the large fixed effects scenario, the probabilities are chosen so that when there is no random effect, the standard classification tree is able to recover the true number of leaves most of the time (i.e., about 95% of times). In the small fixed effects scenario, the probabilities are chosen so that when there is no random effect, the standard classification tree is much less able to recover the true number of leaves (i.e., about 55% of times). The random components are generated based on the following three different scenarios (see Table 1): (1) No random effects (NRE), i.e. $D = 0$; (2) Random intercept (RI), i.e. $z_{ij} = 1$ for $i = 1, \ldots, 100$, and $j = 1, \ldots, 60$, and $D = d_{11} > 0$; 3) Random intercept and covariate (RIC) which is a RI and a linear random effect for X_1. More precisely, $z_{ij} = [1, x_{1ij}]$ for $i = 1, \ldots, 100, j = 1, \ldots, 60$, and $D = (d_{ij}), d_{11} > 0, d_{22} > 0$, and $d_{12} = d_{21} = 0$. Within each fixed effects scenario with random effects, we consider two levels (low and high) for the between-clusters covariance matrix D. For each scenario, we adjust three models: (1) a standard (STD) classification tree model, (2) a random intercept (RI) classification tree model, and (3) a random intercept and covariate (RIC) classification tree model. In addition, using the *glmmPQL* function of R, we fitted two parametric mixed effects logistic regression models (GLMM and MElog). GLMM uses all covariates as predictors and the true random effects structure. MElog uses as predictors the true fixed part structure and the true random effects structure. Clearly, this model is not a real competitor since it is not possible in practice to specify this parametric structure without knowing the true underlying data generating process. The MElog model only serves as a benchmark for comparing the performance of the GMERT model. The simulation results are obtained by means of 100 runs.

The performance of the methods is judged based on their predictive accuracy on the test set as measured by: (1) the predictive mean absolute deviation PMAD $= \frac{\sum_{i=1}^{100}\sum_{j=1}^{50}|\mu_{ij}-\hat{\mu}_{ij}|}{5000}$ and (2) the predictive misclassification rate PMCR $=$

Table 2
Results of the simulation.

DGP	Fixed effect	Random effect	Fitted model	PMAD (%)					PMCR (%)				
				Avg.	Med.	Min	Max	SD	Avg.	Med.	Min	Max	SD
1	Large	No random effect	STD	3.09	3.05	1.48	6.38	0.97	15.71	15.67	13.92	18.20	0.79
			RI	3.86	3.66	1.28	8.85	1.46	16.86	16.58	14.54	21.44	1.52
			RIC	4.17	3.98	1.31	8.85	1.49	16.85	16.60	14.52	21.78	1.55
			GLMM	21.44	21.38	20.34	22.8	0.57	29.76	29.83	27.78	31.20	0.72
			MElog	2.48	2.36	0.78	4.80	0.87	15.49	15.39	13.86	17.62	0.71
2	Small		STD	4.97	4.64	1.73	11.98	1.89	29.33	28.94	26.94	34.68	1.63
			RI	6.35	5.95	2.23	13.36	2.81	31.23	30.41	26.74	38.82	2.83
			RIC	6.32	5.82	2.43	12.52	2.68	31.00	30.18	26.66	38.68	2.70
			GLMM	15.12	15.04	14.33	16.29	0.43	36.87	36.84	35.08	38.56	0.75
			MElog	2.73	2.72	0.86	5.34	0.82	27.72	27.76	26.22	29.12	0.68
3	Large	Random intercept	STD	21.70	21.48	17.44	26.50	1.68	26.49	26.23	21.90	30.90	1.81
			RI	9.20	9.12	7.10	12.13	0.99	19.82	19.78	17.36	22.18	1.11
			RIC	9.69	9.58	7.10	14.87	1.20	20.08	20.02	17.82	22.86	1.16
			GLMM	18.73	18.67	16.93	20.52	0.75	26.53	26.55	23.32	30.22	1.12
			MElog	8.40	8.48	6.26	9.94	0.62	19.13	19.13	16.56	21.24	0.87
4			STD	30.24	29.97	25.29	35.50	1.98	33.65	33.23	28.92	41.14	2.58
			RI	8.59	8.52	6.80	11.42	0.84	16.45	16.46	12.20	20.16	1.16
			RIC	9.37	9.27	7.28	13.61	1.05	16.93	16.85	14.50	20.06	1.15
			GLMM	15.14	15.23	13.28	16.84	0.82	20.73	20.72	18.14	24.14	1.28
			MElog	7.59	7.57	6.06	9.14	0.65	15.69	15.73	11.82	18.34	1.07
5	Small		STD	12.56	12.36	10.40	15.97	1.30	31.70	31.36	29.06	36.44	1.67
			RI	10.71	10.54	7.81	15.44	1.58	31.37	31.17	28.14	36.58	1.62
			RIC	10.79	10.69	7.86	15.43	1.53	31.38	31.12	28.46	36.72	1.58
			GLMM	16.17	16.09	15.08	16.99	0.40	36.14	36.09	33.88	37.96	0.77
			MElog	8.21	8.17	6.61	9.89	0.60	28.87	28.85	27.46	30.64	0.64
6			STD	26.77	26.80	21.53	30.53	1.47	39.32	39.21	34.96	46.10	2.35
			RI	11.20	11.08	8.91	14.73	1.10	24.00	24.03	20.66	30.40	1.43
			RIC	11.40	11.20	9.32	14.66	1.02	24.09	24.06	20.94	30.30	1.42
			GLMM	13.83	13.88	12.42	15.73	0.65	25.54	25.6	22.12	28.98	1.34
			MElog	9.01	8.94	7.65	10.95	0.67	22.56	22.49	19.64	26.62	1.23
7	Large	Random intercept and covariate	STD	20.37	20.48	16.33	23.62	1.24	25.31	25.34	21.76	28.30	1.21
			RI	10.86	10.74	9.25	13.83	0.88	20.87	20.85	18.50	23.32	0.93
			RIC	10.58	10.47	8.43	14.14	0.98	20.83	20.79	18.02	23.54	1.02
			GLMM	20.42	20.43	18.78	21.77	0.70	29.00	29.00	26.38	31.26	0.99
			MElog	9.61	9.53	8.10	12.49	0.70	20.04	19.95	17.68	22.32	0.85
8			STD	30.90	30.92	27.43	35.56	1.60	34.34	33.97	30.06	42.52	2.37
			RI	12.37	12.35	9.91	15.76	0.98	18.15	18.20	15.10	20.82	1.14
			RIC	10.67	10.52	8.63	14.73	1.12	17.28	17.29	14.68	21.12	1.10
			GLMM	15.61	15.67	12.95	17.81	1.03	20.68	20.68	16.38	24.02	1.41
			MElog	9.45	9.39	7.91	11.35	0.74	16.42	16.37	14.16	18.60	0.92
9	Small		STD	12.86	12.64	10.21	17.48	1.45	31.81	31.15	29.00	37.92	1.85
			RI	11.12	10.73	8.87	16.57	1.62	31.36	30.93	28.12	36.82	1.86
			RIC	11.04	10.62	8.50	16.19	1.65	31.35	30.83	28.24	36.12	1.85
			GLMM	16.51	16.49	15.78	17.66	0.36	36.14	36.09	34.18	38.84	0.83
			MElog	8.79	8.73	7.77	10.44	0.50	29.01	28.99	26.98	30.78	0.71
10			STD	25.42	25.18	21.48	28.76	1.58	39.02	38.90	34.26	46.26	2.59
			RI	13.11	13.05	10.67	15.86	1.16	25.98	25.89	22.42	29.12	1.4
			RIC	12.54	12.48	10.23	15.16	1.09	25.84	25.72	22.74	29.82	1.38
			GLMM	15.27	15.27	13.48	17.01	0.77	27.53	27.49	24.02	31.12	1.57
			MElog	10.41	10.34	8.89	12.84	0.71	24.24	24.39	21.24	26.94	1.19

$\frac{\sum_{i=1}^{100}\sum_{j=1}^{50}|y_{ij}-\hat{y}_{ij}|}{5000}$, where $\hat{\mu}_{ij}$ and $\hat{y}_{ij}$ are, respectively, the predicted probability and the predicted class of observation j in cluster i in the test data set. The misclassification rate depends on the cutpoint value used to classify the observations, in particular, when the data has a nested structure with clusters having different sizes in the training and the test data sets. In this study, the cutpoint was selected as the value such that, in the training set, the proportion of observations assigned to class 1 is the closest to the actual proportion of class 1. Table 2 presents a summary of the PMAD (columns 5–9) and the PMCR (columns 10–14) calculated over the 100 runs. In terms of predictive accuracy (PMAD and PMCR), we note that when random effects are present (DGPs 3–10 in Table 1), the mixed effects classification trees (RI or RIC) do much better than the standard classification tree (STD) even with a wrong specification of the random component part. When there is no random effect (DGPs 1 and 2), then the standard classification tree algorithm does slightly better than the mixed effects classification trees. This means that it is much more worse to neglect the clustering nature of the data than it is to add unrequired random effects. As expected, the differences between the models are greater for the PMAD than the PMCR because the latter is based on a dichotomization of the estimated probability. Indeed, a model could fairly well classify a binary outcome even when the estimated probabilities are not very precise. The mixed effects classification trees are also often close to the best possible model, the MElog, which is unknown in practice. Also as expected, a GLMM with main effects only performs poorly for these tree structured DGPs.

4. Discussion and conclusion

This paper extends the MERT approach to the case of non Gaussian response variables. The simulation results in the binary case show substantial improvements of the predictive accuracy over the standard classification tree, whenever random effects are present. However, ensemble methods such as random forests and boosting can greatly improve the predictive performance of trees. Hence, further improvement of the predictive accuracy of the GMERT method could be achieved if we use it as the base learner in an ensemble algorithms. Hajjem et al. (2014) proposed a mixed effects random forest (MERF) method. However, MERF was designed for a continuous response. Extending MERF to the discrete case remains for future work. Supplementary material provides an R program implementing GMERT, additional simulations, a data example, and discussions about: other types of outcome, other tree algorithms, PQL, convergence of GMERT, (Bürgin and Ritschard, 2015). It is available online from the Statistics & Probability Letters web site.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), by Le Fonds de recherche du Québec Nature et technologies (FRQNT), and by HEC Montréal. The authors wish to thank two reviewers for their relevant comments that lead to an improved version of this article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2017.02.033.

References

Abdollel, M., LeBlanc, M., Stephens, D., Harrison, R.V., 2002. Binary partitioning for continuous longitudinal data: Categorizing a prognostic variable. Stat. Med. 21, 3395–3409.
Bürgin, R., Ritschard, G., 2015. Tree-based varying coefficient regression for longitudinal ordinal responses. Comput. Statist. Data Anal 86, 65–80.
Eo, S.-H., Cho, H., 2013. Tree-structured mixed-effects regression modeling for longitudinal data. J. Comput. Graph. Statist. 23, 740–760.
Fu, W., Simonoff, J.S., 2014. Unbiased Regression trees for longitudinal and clustered data. Comput. Statist. Data Anal. 88, 53–74.
Hajjem, A., Bellavance, F., Larocque, D., 2011. Mixed effects regression trees for clustered data. Statist. Probab. Lett. 81, 451–459.
Hajjem, A., Bellavance, F., Larocque, D., 2014. Mixed-effects random forest for clustered data. J. Stat. Comput. Simul. 84, 1313–1328.
Loh, W., Zheng, W., 2013. Regression trees for longitudinal and multiresponse data. Ann. Appl. Stat. 7, 495–522.
R Development Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, www.R-project.org.
Rodriguez, G., 2008. Multilevel Generalized Linear Models. In: De Leeuw, J., Meijer, E. (Eds.), In Handbook of Multilevel Analysis. Springer, New York, pp. 335–376.
Segal, M.R., 1992. Tree-structured methods for longitudinal data. J. Amer. Statist. Assoc. 87, 407–418.
Sela, R.J., Simonoff, J.S., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. 86, 169–207.
Therneau, T.M., Atkinson, E.J., 1997. An Introduction to Recursive Partitioning Using the Rpart Routines. In: Technical Report 61, Department of Health Science Research, Mayo Clinic, Rochester.