Statistics and Its Interface Volume 0 (0000) 1–8

Supplement to "A Modified Classification Tree Method for Personalized Medicine Decisions"

Wan-Min Tsai, Heping Zhang, Eugenia Buta, Stephanie O'Malley and Ralitza Gueorguieva

1. DATA GENERATION IN SIMULATION STUDY

1.1 Scenario with sizeable treatment-covariate interactions

Design 1 (no noise). We generated 1000 subjects in total and, according to the contingency table in Table 1, assigned 500 subjects to treatment A (250 with outcome $Y = 0$ and 250 with outcome $Y = 1$) and 500 to treatment B (250 with outcome $Y = 0$ and 250 with outcome $Y = 1$). We only describe in detail the covariate generation process for subjects on treatment A. A similar process was used to generate covariates for subjects on treatment B.

Among the 250 subjects assigned to treatment A and outcome 0, 126 subjects had a covariate X_1 value generated that was ≤ 0.5 (drawn from a discrete uniform distribution on the grid from 0.01 to 0.5 by increments of 0.0001) and 124 had $X_1 > 0.5$ (drawn from a discrete uniform distribution on the grid from 0.51 to 1.5 by increments of 0.0001). Similarly, among those assigned to treatment A and outcome 1, 224 had an X_1 value generated that was ≤ 0.5 and 26 had $X_1 > 0.5$.

Among the 126 subjects assigned to treatment A, outcome 0, and $X_1 \leq 0.5$, 111 had X_3 (ordinal categorical covariate) drawn from a discrete uniform distribution on 1 to 3 and 15 had X_3 drawn from a discrete uniform distribution on 4 to 5. Among the 111 subjects assigned to treatment A, outcome 0, $X_1 \leq 0.5$, and $X_3 \leq 3$, we randomly selected 20 to get $X_2 = 0$ and the rest (91 subjects) got $X_2 = 1$.

Design 2 (some noise variables). We added the following noise variables to the design 1 dataset: 10 continuous variables drawn from a standard normal distribution rounded to 4 decimal places, 10 binary variables with probability of success 0.2, and one three-level nominal variable drawn from a discrete uniform distribution on 1 to 3. The noise variables are independent of each other, the outcome and covariates.

Design 3 (many noise variables). Following the approach used to generate data for Design 2, we created the Design 3 dataset by adding the following noise variables to the design 1 dataset: 75 continuous variables, 10 binary variables, 4 three-level nominal variables, 3 four-level nominal variables and 3 five-level nominal variables.

1.2 Scenario with small treatment-covariate interactions

We simulated data with small treatment-covariate interactions using Table 2 and the steps outlined above for generation of sizeable treatment-covariate interactions data.

1.3 Scenario with no treatment-covariate interactions

Design 1 (no noise). Data for 1000 subjects were generated according to the following logistic regression model with main effects only: logit $P(Y = 1) = -1+0.3I(T = B)+0.5X_1+0.5I(X_3 \geq 3)+0.5X_2$, where I represents the indicator function, X_1 was drawn from a discrete uniform distribution on the grid from 0.51 to 1.5 by increments of 0.0001, X_2 was drawn from a discrete uniform distribution on 0 to 1, X_3 was drawn from a discrete uniform distribution on 1 to 5, and 500 subjects were assigned to each treatment (A and B).

Design 2 and 3. We added noise using the same method as the one described above for sizeable interactions scenario.

Table 1. Contingency table for sizeable treatment-covariate interactions data generation.

Y	Frequency	Treatment	Frequency	X_1	Frequency	X_3	Frequency	X_2	Frequency
0	250	A(=0)	500	$\leq$0.5	126	$\leq$3	111	0	20
								1	91
						>3	15	0	10
								1	5
				>0.5	124	$\leq$3	80	0	40
								1	40
						>3	44	0	27
								1	17
1	250			$\leq$0.5	224	$\leq$3	89	0	60
								1	29
						>3	135	0	10
								1	125
				>0.5	26	$\leq$3	20	0	10
								1	10
						>3	6	0	3
								1	3
0	250	B(=1)	500	$\leq$0.5	224	$\leq$3	84	0	74
								1	10
						>3	140	0	125
								1	15
				>0.5	26	$\leq$3	20	0	10
								1	10
						>3	6	0	2
								1	4
1	250			$\leq$0.5	126	$\leq$3	116	0	46
								1	70
						>3	10	0	5
								1	5
				>0.5	124	$\leq$3	80	0	40
								1	40
						>3	44	0	18
								1	26

Table 2. Contingency table for small treatment-covariate interactions data generation.

Y	Frequency	Treatment	Frequency	X_1	Frequency	X_3	Frequency	X_2	Frequency
0	250	A(=0)	500	≤ 0.5	160	≤ 3	100	0	50
								1	50
						>3	60	0	35
								1	25
				>0.5	90	≤ 3	42	0	18
								1	24
						>3	48	0	24
								1	24
1	250			≤ 0.5	190	≤ 3	100	0	30
								1	70
						>3	90	0	55
								1	35
				>0.5	60	≤ 3	28	0	12
								1	16
						>3	32	0	16
								1	16
0	250	B(=1)	500	≤ 0.5	175	≤ 3	100	0	60
								1	40
						>3	75	0	29
								1	46
				>0.5	75	≤ 3	40	0	20
								1	20
						>3	35	0	20
								1	15
1	250			≤ 0.5	175	≤ 3	100	0	60
								1	40
						>3	75	0	31
								1	44
				>0.5	75	≤ 3	40	0	20
								1	20
						>3	35	0	20
								1	15