Decision trees-based EM algorithm for semi-parametric mixture cure rate model

Wisdom Aselisewine[1] & Suvra Pal[1]

1. University of Texas at Arlington, Department of Mathematics.

INTRODUCTION

Recent studies on cure rate focus on achieving flexibility and precision on the latency part, whiles fewer studies focus on the incidence part of the model. Due to the inability of the existing logistics model to capture non-linear effects of covariates on incidence part, researchers have redirected efforts to identifying more robust and flexible modeling technique that can accommodate both linear and non-linear effects. Supervised machine learning techniques namely, support vector machines and neural network, have been lately proposed and justified to outperformed the traditional logistics model. However, the problem of interpret-ability still hangs on the balance due to the trade-off between interpret-ability and predictive accuracy in these proposed models.

OBJECTIVES

- To integrate machine learning techniques with cure rate model and come out with a novel cure rate model that can capture non-linearity in lifetime data.
- To develop an efficient estimation algorithm for cure rate models.

METHODOLOGY

Model: *Mixture cure rate model[1].*

$S_p(t_i; x_i, z_i) = 1 - \pi(z_i) + \pi(z_i)\, S_0(t)^{\exp(x_i\beta)}$.

- **Latency Part:** *Cox PH model.*

$H(t_i; x_i) = \lambda_0(t) \exp(x_i\beta)$**.**

- **Incidence Part:** *Decision trees[2].*

$\min_T$ Error(T, D) + $\lambda|T|$ such that Nk ≥ Nm

Estimation of parameters.

- **Latency part.**

EM algorithm[3].

- **Incidence part.**

EM algorithm.
Platt scaling [4].
Multiple imputation.

Simulated Data

Three different settings[5]:

- Scenario 1: $\pi(z) = \frac{\exp(0.3-5z_1-3z_2)}{1+\exp(0.3-5z_1-3z_2)}$
- Scenario 2: $\pi(z) = \frac{\exp(0.3-5z_1^2-3z_2^2)}{1+\exp(0.3-5z_1^2-3z_2^2)}$
- Scenario 3:

$\pi(z) = \exp(-\exp(0.3 - 8\cos z_1 - 5\sin z_2))$.

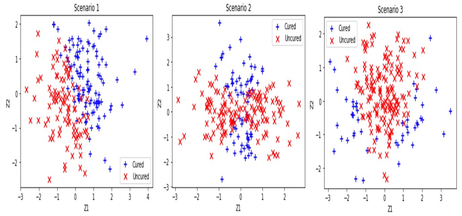

Figure 1: Simulated cured and uncured observations for the three scenarios.

RESULTS

Table 1. Comparison of Bias and MSE of the uncured probability.

n	Setting	Bias		MSE	
		D-Trees	Logistics	D-Trees	Logistics
200	1	.215	.130	.080	.032
	2	.216	.390	.095	.195
	3	.162	.230	.061	.101
400	1	.190	.105	.068	.021
	2	.194	.390	.083	.195
	3	.127	.214	.042	.102

Real Data Application.

Data description.

- Data from a study on leukemia patients.
- Event of interest is relapse or death due to leukemia bone marrow transplantation among 137 patients.

Results for the latency part.

Table 2: Comparison of standard deviation of the latency parameters.

Param.	Standard deviation		P-value	
	D-Trees	Logistics	D-Trees	Logistics
β_1	.235	.302	.113	.098
β_2	.264	.304	.033	.029

RESULTS

Results for the incidence part.

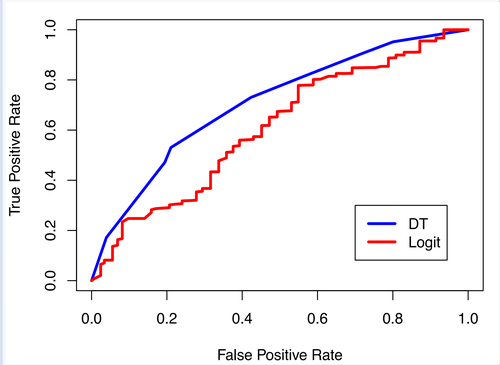

Figure 2: ROC curves corresponding to the leukemia data set.

AUC values.

- Decision trees: 0.7157.
- Logistics model: 0.6092.

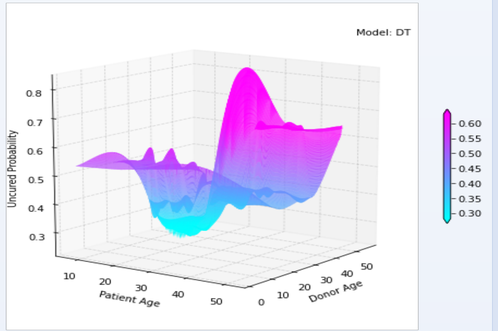

Figure 3: A 3-dimensional plot of uncured probabilities as a function of Patient age and Donor age.

CONCLUSIONS

- Decision trees outperformed the standard logistics model in capturing non-linear covariates effects.
- We proposed the decision trees-based EM algorithm to model the incidence part of the cure model since they are also easy to interpret and closely mimic human decision making process.

Future works

- Extend proposed model to high dimensional covariates and covariates subject to measurement error.
- Extend proposed model to capture competing risk scenario.

References

[1] Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc 11:15–53.

[2] Bertsimas D, Dunn J (2017) Optimal classification trees. Machine Learning 106(7):1039–1082.

[3] Platt J et al (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advanced Large Margin Classification 10:61–74

[4] Balakrishnan, N. and Pal, S. (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes. Statistical Methods in Medical Research, 25, 1535–1563.

[5] Li, P., Peng, Y., Jiang, P., Dong, Q. (2020). A support vector machine based semiparametric mixturecure model. Computational Statistics, 35 (3), 931945.