Data Reconstruction Method for Learning to Censored Survival Data

Sang-Kyun Ko[1,2] and **Yung-Kyun Noh**[1,3]

[1] *Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea*
[2] *Ainex Co., Ltd., Seoul 06779, Korea*
[3] *School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Republic of Korea*
highsg19101@gmail.com, nohyung@hanyang.ac.kr

Lab. of Advanced Application for Intelligence Systems

Survival Analysis and Censored Data Problem

- **Survival analysis**, or time-to-event analysis, is a statistical methodology employed to examine the duration, called survival time T, until the incidence of a specific event within a population, considering covariates $x \in \mathbb{R}^D$ that can impact event occurrence.
- **Right-censored data** in survival analysis refers to observations where the event of interest has not occurred by the end of the study. This incomplete target value can only confirm that this data's survival time exceeds a specified threshold.
- **Concordance Index** (c-index), the commonly employed evaluation metric for survival analysis, can be utilized as test data in the censored data.

Data Reconstruction Method

- **Matrix survival data.**
  We aim to predict the data index with the highest survival time from a given data pair (matrix) via reconstruction.

$$\boldsymbol{M_n} = \begin{pmatrix} \boldsymbol{x_k} \\ \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{x_i} + \boldsymbol{x_j}) \\ \boldsymbol{x_k} \end{pmatrix}_{\boldsymbol{k}\in\{\boldsymbol{i},\boldsymbol{j}\},(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{\mathcal{A}}} , \boldsymbol{y_n} \in \{\mathbf{0},\mathbf{1},\mathbf{2}\}.$$

- **Classification Model.**
- The n-th prediction result $\hat{y}_n$ of the model for the reconstructed matrix $M_n \in \mathbb{R}^{3\times D}$ is define as follows:

$$\hat{y}_n = \underset{k\in\{0,1,2\}}{\operatorname{argmax}} f_k(M_n)\,,\ n = 1,\dots,\mathbb{N},\ \textstyle\sum_{k=0}^{2} f_k = 1.$$

- **Matrix Concordance Index.**

$$\boldsymbol{c^M} = \frac{\mathbf{1}}{\mathbb{N}} \sum_{\boldsymbol{n=1}}^{\mathbb{N}} \mathbb{I}_{(\hat{y}_n, \boldsymbol{y_n})}\,.$$

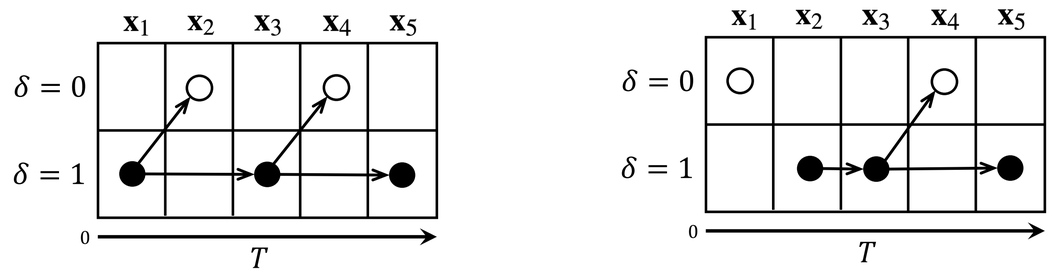
  Where $\mathbb{I}$ returns 1 if $T_i \neq T_j$ and $\hat{y}_n = y_n$, also it returns 0.5 if $T_i = T_j$ and $\hat{y}_n = y_n$. Otherwise returns 0.
- **Relaxation Matrix Concordance Index.**

$$\tilde{\boldsymbol{c}}^{\boldsymbol{M}} = \frac{\mathbf{1}}{\mathbb{N}} \sum_{\boldsymbol{n=1}}^{\mathbb{N}} \boldsymbol{f_{y_n}(M_n)}\,.$$

- **Surrogate Loss** directly optimize the c-index and utilize the censored data.

$$\boldsymbol{\mathcal{L}} = \frac{\mathbf{1}}{\mathbb{N}} \sum_{\boldsymbol{n=1}}^{\mathbb{N}} \left(\mathbf{1} - \boldsymbol{f_{y_n}(M_n)}\right).$$

Concordance Index (c-index)



The acceptable pair set $\mathcal{A}$ used in the c-index can be expressed as a graph, and the pairs connected in the Figure are as follows:
(left) : $(x_1, x_2), (x_1, x_3), (x_1, x_4), (x_1, x_5), (x_3, x_4), (x_3, x_5)$.
(right): $(x_2, x_3), (x_2, x_4), (x_2, x_5), (x_3, x_4), (x_3, x_5)$.

Survival Data

- D-dimensional random variable vector x that concerns the occurrence of an event.
- $\delta \in \{0,1\}$ that indicates whether an event occurs ($\delta = 0$ indicates no event has occurred).
- Event waiting time (survival time) $T \in \mathbb{R}$.
- $\mathcal{D} = \{x_i, T_i, \delta_i\}_{i=1}^{N}$.
- The **c-index** of the survival time predictive model f for the set $\mathcal{A}$ is defined as follows:

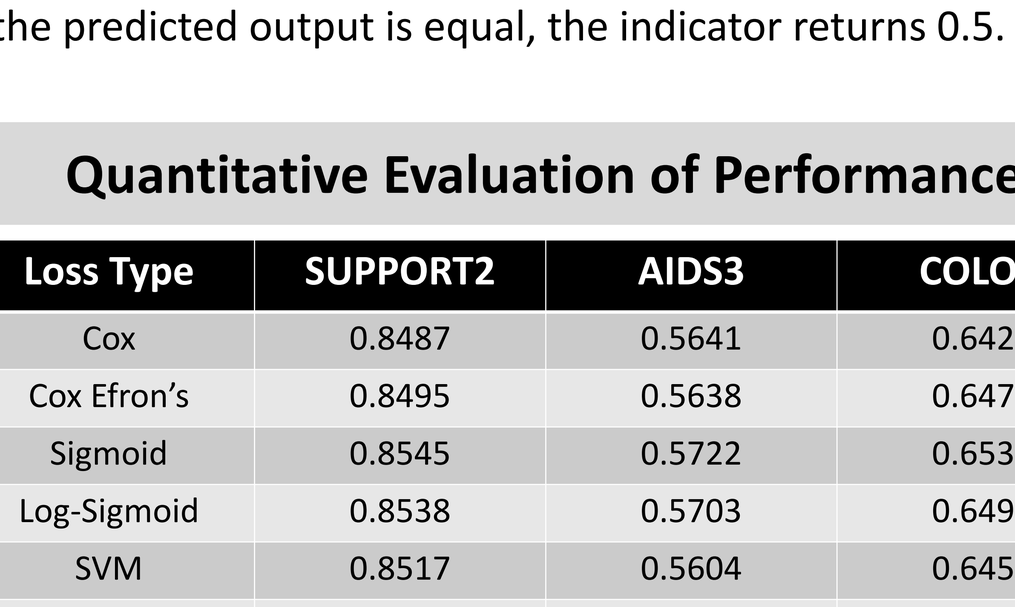
$$c(x_i, T_i, \delta_i) = \frac{1}{|\mathcal{A}|} \sum_{(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{\mathcal{A}}} \mathbb{I}_{(f(x_i), f(x_j))}\,.$$

- Where $|\mathcal{A}|$ is the number of elements in the set $\mathcal{A}$, and the indicator function $\mathbb{I}_{(f(x_i),f(x_j))}$ returns 1 if the predicted order is concordance. And if each survival time are same and the predicted output is equal, the indicator returns 0.5.

Quantitative Evaluation of Performance

Loss Type	SUPPORT2	AIDS3	COLON
Cox	0.8487	0.5641	0.6428
Cox Efron's	0.8495	0.5638	0.6471
Sigmoid	0.8545	0.5722	0.6536
Log-Sigmoid	0.8538	0.5703	0.6496
SVM	0.8517	0.5604	0.6453
Boost	0.8537	0.5714	0.6391
Wasserstein	0.8542	0.5596	0.6475
Ours	**0.8548**	**0.5555**	**0.6524**

- Proposed method shows **competitive performance with other SOTA methods** in survival analysis



Impact of employing the censored data.
0.612 (0%) -> 0.647 (100%)