Investigation on imbalanced Cox regression based on simulations

Jing Wang

University of Connecticut

April 15, 2024

Overview

Background

- Imbalanceness is an issue for binary response
- Recent years many techniques have been developed to rebalance the data from both statistics and machine learning perspective.
- Survival analysis can also suffer imbalance issue
- Research on imbalance survival analysis is rarely touched by researchers.

- Logistic regression is a wildly used statistical model for **binary data**
- For a $y \in \{0, 1\}$ and $\boldsymbol{x}$, logistic regressions assume that

$$\mathbb{P}(y = 1|\boldsymbol{x}) = \frac{e^{\alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}}$$

- For this model
  - Response y: **Yes** or **No**, **Have a disease** or **Not have a disease**
  - Covariate $\boldsymbol{x}$: Age, Height, Weight, etc.
  - α**: Intercept**, $\boldsymbol{\beta}$: coefficients of $\boldsymbol{x}$.

Paper review

Highly imbalanced logistic regression

- What if $\mathbb{P}(y = 1)$ is very small (For example 0.5%)?
  - rare diseases, war, disaster
  - Resultant data is highly imbalanced: rare 1 and large amount of 0
- Wang (2020) proposes the following adjusted logistic regression to model highly imbalanced binary data:

$$\mathbb{P}(y = 1|\boldsymbol{x}) = \frac{e^{\alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}}$$

- Assume that $\boldsymbol{\beta}$ **is fixed and** $\alpha \to -\infty$ as $N \to \infty$, then

$$\mathbb{P}(y = 1) \to 0,$$

and

$$\frac{N_1}{N} = \frac{N_1}{N_0 + N_1} = \mathbb{P}(y = 1)\{1 + o_p(1)\} \to 0.$$

Paper review

Estimation of highly imbalanced logistic regression

- Denoting $\boldsymbol{z} = (1, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, the MLE estimator:

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \{ y_i \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log(1 + e^{\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\theta}}) \}$$

  is **asymptotic normal**:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx N\left(\mathbf{0}, \frac{\boldsymbol{V}_f}{N_1}\right).$$

- For regular logistic regression, we usually have

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx N\left(\mathbf{0}, \frac{\boldsymbol{V}_f}{N}\right)$$

- Information in data is essentially determined by N_1 instead of N.

Paper review

Methods to balance the data

- Imbalanceness may cause many issues in data analysis
- If imbalance rate is low, to get enough N_1, needs large amount of N
- In practice, imbalanceness may cause computational instability.
- In machine learning, **under-sample** and **over-sampling** is used.

Paper review

Methods to balance the data

Under-sampling

- If $y_i = 1$, include $(y_i, \boldsymbol{x}_i)$ into the sample
- If $y_i = 0$, include $(y_i, \boldsymbol{x}_i)$ with **probability** ρ

Over-sampling

- If $y_i = 1$, include $(y_i, \boldsymbol{x}_i)$ into the sample $1 + v$ times, where $v \sim Poisson(\lambda)$
- If $\delta_i = 0$, include $(y_i, \boldsymbol{x}_i)$ only once

Paper review

Methods to balance the data

- Wang (2020) also propose two estimators based on resultant balanced data set: **weighted** and **unweighted** estimator

Weighted method

- Taking samples with under- (or over-) sampling with ρ (or λ)
- If using under-sampling, weighting 0's with $\frac{1}{\rho}$,
- If using over-sampling, weighting 1's with $1/(1+\lambda)$
- Apply weighted logistic regression on sample

Unweighted method

- Taking samples with under- (or over-) sampling with ρ (or λ)
- **Directly apply logistic regression on the sample and obtain $\hat{\theta}_u$**
- **Adjust** $\hat{\alpha} = \hat{\alpha}_u + \log(\rho)$ **or** $\hat{\alpha} = \hat{\alpha}_u + \log(1+\lambda)$.

Paper review

Methods to balance the data

Wang (2020) proved the following results

- For all the 4 methods, **the asymptotic variances are of order $\frac{1}{N_1}$**.
- Both under- and over-sampling will cause information loss
- For under-sampling, when $e^{\alpha}/\rho \to 0$, **no information loss**
- For over-sampling, when $\lambda = 0$ or ∞, **no information loss**
- For under-sampling. unweighted estimator with bias correction is more efficient
- For over-sampling. weighted estimator is more efficient

Paper review

Survival data and Cox regression

- In statistics, survival analysis studies time to a certain event
  - Time to death (Survival time), marriage,...
  - Time to develop a certain diseases
  - Time to malfunctions of a machine
- For survival analysis, due to practical issues, for example, limited resources, data usually suffer censorings
  - A study follows a large amount of patients for 5 years for a certain disease. However, some patients never develop that diseases in this 5 years. Then, for those patients, we only know their $t_i > 5$.

Paper review

Survival data and Cox regression

- A survival data set usually contains three parts $(t_i, \delta_i, \boldsymbol{x}_i)$
  - t_i: observed survival time
  - δ_i: indicator of event. If $\delta = 1$, event is observed, if $\delta = 0$, censoring is observed.
  - $\boldsymbol{x}_i$: covariates related to t_i, for example, age, height, weight,...
- A popular model for survival data is the Cox model:

$$h(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}.$$

  - $h(t|\boldsymbol{x})$ called hazard function, which is the risk of event at time t.
  - $h_0(t)$ is the baseline hazard function
  - Covariates $\boldsymbol{x}$ affect survival time through affecting hazard function.

- In Cox model, we do not assume a specific form of $h_0(t)$ and thus, there is no intercept term α in Cox model because

$$h(t|\boldsymbol{x}) = h_0(t)e^{\alpha+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}} = \tilde{h}_0(t)e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}},$$

  where $\tilde{h}_0(t) = h_0(t)e^{\alpha}$ and thus α is not estimable.
- From the above reason, we usually focus more on estimating coefficients $\boldsymbol{\beta}$.
- Note that the **lower $h_0(t)$**, the **lower risk of event** and thus the **lower rate of observed events**.

Background

Some question about survival analysis

- Data imbalance can be an issue for survival data.
- Sometimes, censored data points may be hundred times of events

For example, in practice

- Observed data collected from a successfully organ transplant
  - Death rate maybe too low
- Realtime surveillance for rare diseases with modern wearable devices
  - Time to catch the disease will have too much censors

For imbalance survival data, some questions need to be answered

- How imbalance influence Cox regression?
- Is method for balancing binary data can be apply to survival data?
- We use simulation to investigate these problems

Simulation setup

- There are three covariates in our simulations: age (x_1), treatment (x_2) and biomarker (x_3) in our simulation.
  - $X_1 \sim \text{Normal}(\mu = 30, \sigma = 2)$
  - $X_2 \sim \text{Bernoulli}(0.5)$
  - $X_3 \sim \text{Normal}(\mu = 10, \sigma = 1)$
- The hazard function of Cox model is

$$h(t|x_1, x_2, x_3) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3},$$

where $\beta_1 = -0.1$, $\beta_2 = -1$, and $\beta_3 = 0.5$

Simulation setup

- We let $h_0(t) = \lambda \gamma t^{\gamma-1}$ ($\lambda > 0$ and $\gamma > 0$) in our simulation.
  - $\gamma = 1.5$
  - $\log(\lambda) \in (\log(1 \times 10^{-7}), \log(1.4 \times 10^{-4}))$
    - ★ We equally select five λ in log scale
    - ★ Reduce the event rate ($\frac{N_1}{N}$) from 50% to 0.5%.
- The sample size (N) we tried in our simulations are: 600, 1000, 5000.

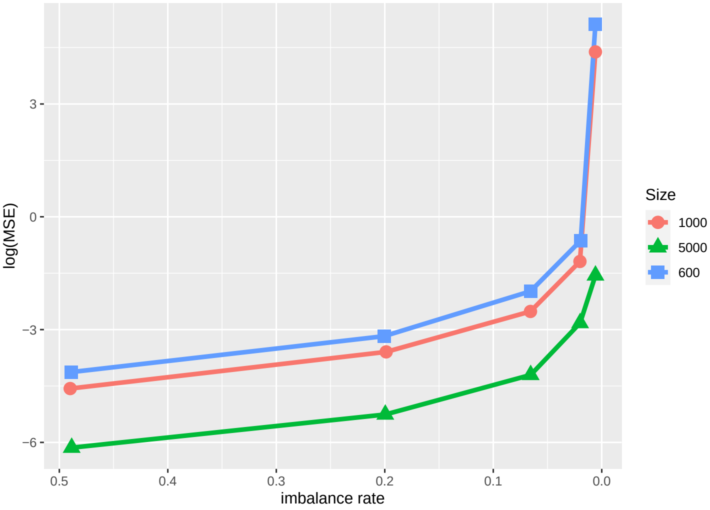
Effects of imbalance

- We first use simulation to check if imbalance affects Cox regression

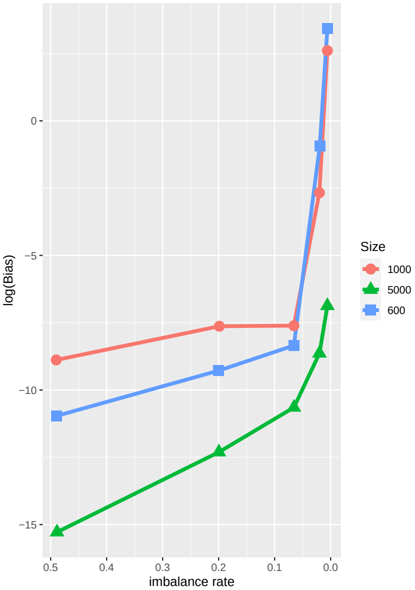
Methodology of simulation study

- Generate survival data with different total sample size and event rate
- Apply Cox regression on the generated data
- Iterate $S = 200$ times and compute empirical MSE, bias and variance

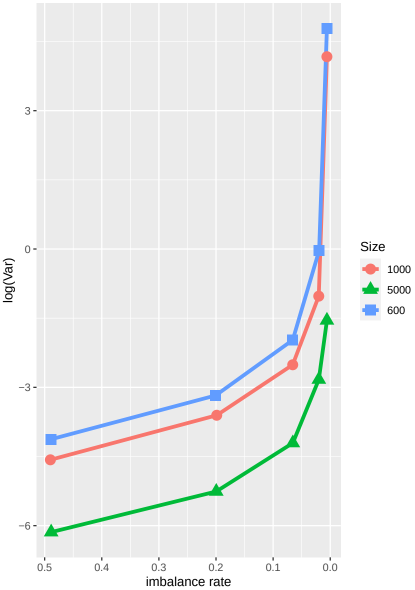
Effects of imbalance



Effects of imbalance



(a) Bias



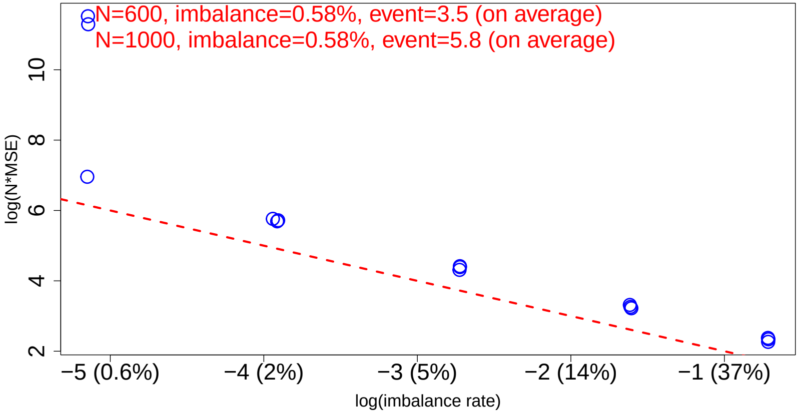
(b) Variance

Effects of imbalance

Relation between MSE and imbalance rate

Convergence rate

$$\log(N \times \mathrm{MSE}) = -\log(\tau) + \log(K) \Rightarrow \mathrm{MSE} = \frac{K}{\tau N} = \frac{K}{N_1}$$



Implication of the results

Implication from the simulation

- High imbalance rate will highly inflate MSE
- The amount information is essentially determined by N_1
- For finite sample, highly imbalance data is hard to analyze since observation of event is not sufficient

Implication of the results

Problem in practice and possible solution

- Reasonable estimates may involve massive data because we need large amount of data to observe sufficient events
- For online system or observed data, computation complexity maybe heavy for massive data and too waste to record too much censors
- Sampling may relief imbalance issue and reduce computational burden

Main idea

Similarity between logistic regression and Cox model

Assuming a proportional hazard model $h(t; \boldsymbol{x}_i) = h_0(t)e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}$, for survivial data $\{(t_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^N$

$$
\begin{aligned}
L(\beta; \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\delta}) &= \prod_{i=1}^N e^{-H(t_i)e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \prod_{i=1}^N \{h(t_i)\}^{\delta_i} \\
&= \prod_{i=1}^N \exp\left\{-e^{\log\int_0^{t_i} h_0(s)\mathrm{d}s + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right\} \prod_{i=1}^N \left\{e^{\log h_0(t_i) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right\}^{\delta_i}.
\end{aligned}
$$

Assuming that $h_0(t)$ is very **small**, for example, $\log\int_0^{t_i} h_0(s)\mathrm{d}s \to -\infty$, we have $e^{\log\int_0^{t_i} h_0(s)\mathrm{d}s + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \approx 0$ and thus

$$
\exp\left\{-e^{\log\int_0^{t_i} h_0(s)\mathrm{d}s + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right\} \approx \frac{1}{1 + e^{\log\int_0^{t_i} h_0(s)\mathrm{d}s + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}},
$$

since $e^{-x} \approx \frac{1}{1+x}$, when $x \to 0$.

Main idea

Similarity between logistic regression and Cox model

Now, we have the likelihood function of $\{(t_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^N$ is

$$L(\beta; \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\delta}) \approx \prod_{i=1}^N \frac{1}{1 + e^{\log \int_0^{t_i} h_0(s)\mathrm{d}s + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}}} \prod_{i=1}^N \left\{ e^{\log h_0(t_i) + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right\}^{\delta_i}.$$

The likelihood function of logistic regression of $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$

$$L(\beta; \boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^N \frac{1}{1 + e^{\alpha_i + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}}} \prod_{i=1}^N \left(e^{\alpha_i + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)^{y_i},$$

if α_i are different (commonly used as *offset* option in R).

Implication

- The likelihoods are similar for imbalanced survival data
- Maybe methodology for balancing binary data can be used for imbalanced survival data.

Methodology

Sampling techniques

- Remember that **events** are essential information
- We propose to use under-sampling method because it reduces computational complexity
- **Under sampling** is widely used in binary data and proposed in Keret and Gorfine (2023) for survival data

Under-sampling for survivial data

- If $\delta_i = 1$, include $(t_i, \delta_i, \boldsymbol{x}_i)$ into the sample
- If $\delta_i = 0$, include $(t_i, \delta_i, \boldsymbol{x}_i)$ with probability ρ

Methodology

Sampling techniques

Features of under-sampling

- The resultant sample is of size $N_1 + \rho N_0$
- We can set ρ by ourselves to balance the data as we need
- The resultant estimator is usually biased due to different sampling probability between 0 and 1
- Easy to apply to massive data and streaming data

Methodology

Estimation techniques

To debias **Cox regression**, we can use **weighted method**

Weighted method

- For sampled data, we weight censors with $\frac{1}{\rho}$
- This is essentially approximating the original likelihood and thus should be similar with full data MLE.
- Proposed in Keret and Gorfine (2023).

- If ρ is small (usually for imbalanced data), the variance will be inflated
- We give censors more weights, which is not reasonable

Methodology

Estimation techniques

Remember that we mentioned similarity between logistic regression and Cox regression. Let ν_i denote the indicator that if i-th data is sampled. We have

$$\mathbb{P}(y_i = 1|\boldsymbol{x}_i, \nu_i = 1) = \frac{\mathbb{P}(\nu_i = 1|y_i = 1, \boldsymbol{x}_i)\mathbb{P}(y_i = 1|\boldsymbol{x}_i)}{\sum_{t=0}^{1} \mathbb{P}(\nu_i = t|y_i = t, \boldsymbol{x}_i)\mathbb{P}(y_i = t|\boldsymbol{x}_i)}$$
$$= \frac{e^{\alpha - \log\rho + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\alpha - \log\rho + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}$$

The likelihood of sampled data is

$$\prod_{i=1}^{N^*} \mathbb{P}(y_i = 1|\boldsymbol{x}_i, \nu_i = 1) = \prod_{i=1}^{N^*} \frac{1}{1 + e^{\alpha - \log\rho + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \prod_{i=1}^{N^*} \left(e^{\alpha - \log\rho + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^{y_i}$$

- Only the intercept term is different

Methodology

Estimation techniques

- For logistic regression, we only need to adjust intercept term
- For Cox regression, we only estimate β and do not care $h_0(t)$
- Maybe it is possible to directly apply Cox regression on sampled data

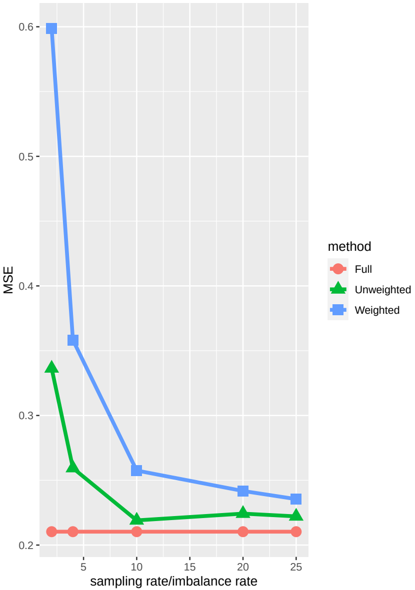
Weighted method

- Taking samples with negative sampling with user-defined ρ
- Weighting censors with $\frac{1}{\rho}$, apply weighted Cox regression on sample

Unweighted method

- Taking samples with negative sampling with user-defined ρ
- Directly apply Cox regression on the sample

Estimation efficiency



(a) MSE

(b) Bias

Consistent inferential results

	coef	exp(coef)	se(coef)	z	p
age	-0.19684	0.82132	0.08813	-2.233	0.0255
trt	-0.71910	0.48719	0.38426	-1.871	0.0613
biomk	0.14332	1.15410	0.18204	0.787	0.4311

Likelihood ratio test=9.38 on 3 df, p=0.02466
n= 5000, number of events= 31

(a) Full

	coef	exp(coef)	se(coef)	z	p
age	-0.1817	0.8338	0.0866	-2.098	0.0359
trt	-0.7493	0.4727	0.3846	-1.948	0.0514
biomk	0.1387	1.1488	0.1726	0.804	0.4215

Likelihood ratio test=9.53 on 3 df, p=0.02298
n= 668, number of events= 31

(b) Unweighted

	coef	exp(coef)	se(coef)	z	p
age	-0.18639	0.82995	0.08673	-2.149	0.0316
trt	-0.76358	0.46600	0.38457	-1.986	0.0471
biomk	0.14006	1.15034	0.17194	0.815	0.4153

Likelihood ratio test=9.93 on 3 df, p=0.01918
n= 668, number of events= 31

(c) Weighted

- When $\frac{N_0}{N_1} = 25$ in resultant sample, inference are also very similar
- We only use 668 observations instead of 5000

Computational complexity

Table: Computaional time (in seconds)

Sampling rate	Unweighted	Weighted	Full data
0.02	3.92	4.37	26.65
0.05	4.47	4.73	26.65
0.125	6.38	7.00	26.65

- Computational time reduced a lot if use a small sample

- This will be more attractive for massive streaming data if we also use an online version of Cox regression because we can reject a lot of censors without losing too much estimation efficiency

Conclusions

Conclusion for this project:

- Imbalance may cause estimation issue for Cox regression
- Balancing methods for classification can be applied to survival data
- It is even possible to obtain similar results with very small data size when dealing with massive imbalance survival data

Future extension for this project:

- We limited our scope to large scale imbalance data. For finite sample size, it may need other techniques for example, penalization methods
- Nonuniform sampling can be considered
- Combining negative sampling with online updating of Cox regression maybe interesting and of practical interests.

References I

Keret, N. and Gorfine, M. (2023). Analyzing big ehr data—optimal cox regression subsampling procedure with rare events. *Journal of the American Statistical Association* **0**, 0, 1–14.

Wang, H. (2020). Logistic regression for massive data with rare events. In *ICML*.

Thank you!