Advanced Analytics for Large-Scale Failure-Time Data

Keivan Sadeghzadeh, PhD

Assistant Professor

University of Massachusetts Dartmouth

Abstract

The increasing accessibility of information has led to availability of massive failure-time data with large number of variables. Therefore, determining efficient and important explanatory variables provides an excellent opportunity to analyze causes and effects of failures. This research is motivated by the significance of the applied analytical methods for analyzing large-scale failure-time data in the presence of uncertainty and censored data to facilitate decision-making, increase productivity, and reduce costs.

Introduction

Failure-time data analysis has an inevitable role in predicting the probability of many failure occurrence such as hardware, software and human error. Thus, necessity of optimal and practical solutions for analysis of complex large-scale failure-time dataset is not only obvious but desired. Advanced analytics could be used to determine a subset of efficient and important explanatory variables that are significantly more valuable for analyzing, investigating, and making decision.

Objective

The objective is to apply large-scale data management methods and design procedures including a class of techniques for variable selection, classification, and reduction for complex large-scale failure-time data analysis as practical solutions to reduce redundant information, avoid data analysis difficulties, and facilitate decision-making process. This is obtained through variable efficiency and importance recognition which is the correlation of independent variables and failure-time.

Large-Scale Failure-Time Data

Failure-time data analysis considers the time until an event occurrence, focuses on predicting the probability of failure.

r	t	u_1	u_2	u_3	u_4	...	u_p
1	t_1	u_{11}	u_{12}	u_{13}	u_{14}	...	u_{1p}
2	t_2	u_{21}	u_{22}	u_{23}	u_{24}	...	u_{2p}
...	...	...	...	...	...	...	...
n	t_n	u_{n1}	u_{n2}	u_{n3}	u_{n4}	...	u_{np}

In many areas, there are great interests in time and causes of failures. Hospital readmission in medical sciences, delay in management, bankruptcy in business, divorce in phycology, arrest in criminology as well as collapse in engineering science, are samples of failure events which make changes in features and outcomes, and also affect the performance of the system.

Methods

Class I: Nonparametric Resampling Method for Kaplan–Meier (KM) Estimator Test

$$\hat{S}(t) = \prod_{i=1}^{t}\left(1 - \frac{d_i}{n_i}\right) \qquad m_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j)$$

Class II: Heuristic Randomized Method through Accelerated Failure Time Model

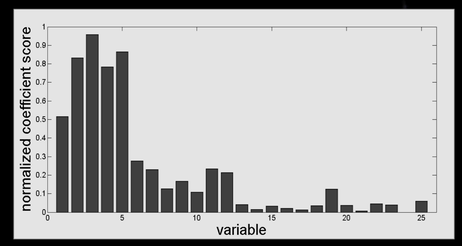
$$S(t|x) = S_0\left(\frac{t}{\psi(x)}\right), \psi(x) = e^{\beta X} \qquad \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\ln\boldsymbol{T}$$

Class III: Hybrid Clustering Algorithm using Weighted K-mean Method

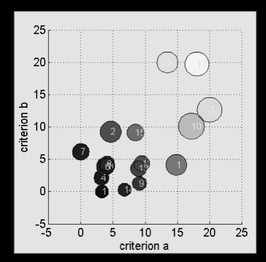
$$w_{ij} = f(v_{ij}^1, v_{ij}^2, \ldots, v_{ij}^k) \qquad Cost = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left|y_i - y_j\right|$$

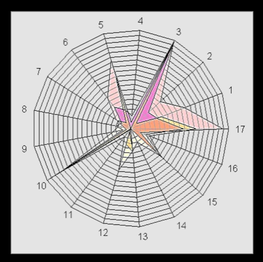
Experimental Results

Normalized coefficient score results for the simulation dataset by 3R algorithm for Class II [5]



Inefficient variable reduction for PBC dataset for Class I: Hybrid (Bubble) scatter [3] and Radar plot [1]





Performance of NTS & SSG methods based on the simulation numerical experiments for Class I [4]

	Simulation Model												
	#1		#2		#3		#4		#5		#6		Ave
Method	m	p	m	p	m	p	m	p	m	p	m	p	$\bar{p}$
NTS (F)	3	0.75	3	0.75	3	0.60	4	0.80	4	0.67	5	0.83	0.73
NTS (C)	3	0.75	3	0.75	4	0.80	4	0.80	5	0.83	5	0.83	0.79
SSG	2	0.50	3	0.75	3	0.60	4	0.80	4	0.67	4	0.67	0.66
Definition	4		4		5		5		6		6		

References

1. Hybrid Nonparametric Approach for Feature Selection in Nonlinear Failure-Time Data, (*Working Paper*)
2. Analytical clustering procedures in massive failure data, *IEEE Xplore*
3. Variable Selection Methods for Right-Censored Time-to-Event Data, *Journal of Quality and Reliability Engineering*
4. Nonparametric Data Reduction Approach for Large-Scale Survival Data Analysis, *IEEE Xplore*
5. Heuristic Ranking Classification Method for Complex Large-Scale Survival Data, *Advances in Intelligent Systems and Computing*