JCGIRM Journal of Corporate Governance, Insurance, and Risk Management (JCGIRM)
2014, Volume 1, Series 2

Pages 77-91

Multidisciplinary Decision-Making Approach to High-Dimensional Event History Analysis through Variable Reduction Methods

Keivan Sadeghzadeh[a,*], **Nasser Fard** [b]

[a]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, USA

[b]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, USA

ABSTRACT

As an analytical approach, decision-making is the process of finding the best option from all feasible alternatives. The application of decision-making process in economics, management, psychology, mathematics, statistics and engineering is obvious and this process is an important part of all science-based professions. Proper management and utilization of valuable data could significantly increase knowledge and reduce cost by preventive actions, whereas erroneous and misinterpreted data could lead to poor inference and decision-making. This paper presents a class of practical methods to analyze high-dimensional event history data to reduce redundant information and facilitate practical interpretation through variable inefficiency recognition. In addition, numerical experiments and simulations are developed to investigate the performance and validation of the proposed methods.

ARTICLE INFO

Keywords:

decision-making, logical model, event history analysis, time-to-event data, variable reduction

**Corresponding author:*

k.sadeghzadeh@neu.edu

(Keivan Sadeghzadeh)

Article Submitted 12-01-2014
Article Accepted 24-02-2014

**Article previously published in EJEM 2014, vol 1, No. 2

1. INTRODUCTION

Analytics data driven decision-making can substantially improve management decision-making process. In social science areas such as economics, business and management, decision-making is increasingly are based on the type and size of data, as well as analytic methods. It has been suggested that new methods to collect, use and interpret data should be developed to increase the performance of the decision makers (Lohr, S., 2012) (Brynjolfsson, E., 2012).

In the fields of economics, business and management, analyzing the collected data from different sources such as financial reports and consequently determining effective explanatory variables, specifically in complex and high-dimensional event history data provide an excellent opportunity to increase efficiency and reduce costs.

In economics, term event history analysis is used as an alternative to time-to-event analysis which has been used widely in the social sciences where interest is on analyzing time to events such as job changes, marriage, birth of children and so forth (Lee, E. T., and Wang, J. W., 2013). Some aspects make difficulty in analyzing this type of data using traditional statistical models. Dimensionality and non-linearity are among those (Allison, Paul D., 1984). Analysis of datasets with high number of explanatory variables requires different approaches and variable selection techniques could be used to determine a subset of variables that are significantly more valuable to (Yao, F., 2007) (Hellerstein, J., 2008) (Segaran, T., and Hammerbacher, J., 2009) (Feldman, D. et al., 2013) (Manyika, J. et al., 2011) (Moran, J., 2013) (Brown, B. et al., 2011).

The purpose of this study is to design a procedure including a class of methods for variable reduction via determining variable inefficiency in high-dimensional event history analysis where variable efficiency refers to the effect of a variable on event history data. As an outline, the concept of decision-making process, event history analysis, and relevant data analysis techniques are presented in Section 2. The logical model for the transformation of the explanatory variable dataset is proposed and three multidisciplinary variable selection methods and algorithms through variable efficiency are designed in Section 3. The results and comparison of results with well-known methods and simulation patterns are presented in Section 4. Finally, concluding remarks, including the advantages of the proposed methods are discussed in Section 5. The computer package that we use in this research is the MATLAB® R2011b programming environment.

2. BASES AND CONCEPTS

In this section, applied introductions to decision-making process and event history analysis as well as data analysis techniques are presented.

2.1. Decision-Making Process

Decision-making theories are classified based on two attributes: (a) Deterministic, which deals with a logical preference relation for any given action or Probabilistic, which postulate a probability function instead, and (b) Static, which assume the preference relation or probability function as time-independent or Dynamic which assume time-dependent events (Busemeyer, J. R., and Townsend, J. T., 1993). Historically, the Deterministic-Static decision-making is more popular decision-making process specifically under uncertainty. The assumption of decision-making in this study falls in this category as well.

As a process of making choices by setting objectives, gathering information, and assessing alternative choices in a decision-making process, broadly includes seven steps: (1) Defining the decision, (2)

Collecting information, (3) Identifying alternatives, (4) Evaluating the alternatives, (5) Selecting best alternative(s), (6) Taking action, (7) Review decision and consequences (Busemeyer, J. R., and Townsend, J. T., 1993).

A major part of decision-making involves the analysis of a finite set of alternatives described in terms of evaluative criteria. The mathematical techniques of decision-making are among the most valuable factors of this process, which are generally referred to as realization in the quantitative methods of decision-making (Sadeghzadeh, K., and Salehi, M. B., 2010). With the increasing complexity and the variety of decision-making problems due to the huge size of data, the process of decision-making becomes more valuable (Brynjolfsson, E., 2012).

A brief review of event history analysis concept and definition of survival function is following.

2.2. Event History Analysis

Event history analysis consider the time until the occurrence of an event. The time can be measured in days, weeks, years, etc. Event history analysis is also known as time-to-event analysis which generally defined as a set of methods for analyzing such data where subjects are usually followed over a specified time period. Event history (time-to-event data) analysis has been used widely in the social sciences such as felons' time to parole in criminology, duration of first marriage in sociology, length of newspaper or magazine subscription in marketing and worker's compensation claims in insurance (Lee, E. T., and Wang, J. W., 2013) (Hosmer D. W. Jr., and Lemeshow, S., 1999) (Kalbfleisch, J. D., and Prentice, R. L., 2011).

Methods to analyze event history data can be categorized in parametric, semi-parametric and nonparametric methods. Parametric methods are based on survival function distributions such as exponential. Semi-parametric methods don't assume knowledge of absolute risk and estimates relative rather than absolute risk and this assumption is called the proportional hazards assumption. For moderate- to high-dimensional covariates, it is difficult to apply semi-parametric methods (Huang, J., Ma, S., and Xie, H, 2006). In nonparametric methods which are useful when the underlying distribution of the problem is unknown, there are no math assumptions. Nonparametric methods are used to describe survivorship in a population or comparison of two or more populations. The Kaplan-Meier Product Limit estimate is a nonparametric method which is the most commonly used nonparametric estimator of the survival function and has clear advantages since it does not require an approximation that results the division of follow-up time assumption (Lee, E. T., and Wang, J. W., 2013) (Holford, T. R., 2002).

The probability of the event occurring at time *t* is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (1)$$

In event history analysis, information on an event status and follow up time is used to estimate a survival function $S(t)$, which is defined as the probability that an object survives at least until time *t*:

$$S(t) = P\ (an\ object\ survives\ longer\ than\ t) = P\ (T > t) \quad (2)$$

From the definition of the cumulative distribution function:

$$S(t) = 1 - P(T \leq t) = 1 - F(t) \quad (3)$$

Accordingly survival function is calculated by probability density function as:

$$S(t) = \int_t^{\infty} f(u)du \quad (4)$$

In most applications, the survival function is shown as a step function rather than a smooth curve. Nonparametric estimate of $S(t)$ according to Kaplan–Meier (KM) estimator for distinct ordered event times t_1 to t_n is:

$$\hat{S}(t) = \prod_{i=1}^{t}(1 - \frac{d_i}{n_i}) \quad (5)$$

Where at each event time t_j there are n_j subjects at risk and d_j is the number of subjects which experienced the event.

A review of relevant used data analysis techniques in this study including discretization process as well as data reduction and variable selection methods is presented next.

2.3. Data Analysis Techniques

Discretization Process

Variables in a dataset potentially are a combination format of different data types such as dichotomous (binary), nominal, ordinal, categorical, discrete, and continuous (Interval). There are many advantages of using discrete values over continuous as discrete variables are easy to understand and utilize, more compact and more accurate. Quantizing continuous variables is called discretization process.

In the splitting discretization methods, continuous ranges are divided into sub-ranges by the user specified width considering range of values or frequency of the observation values in each interval, respectively called equal-width and equal-frequency. A typical algorithm for splitting discretization

process which quantifies one continuous feature at a time generally consists of four steps: (1) sort the feature values, (2) evaluate an appropriate cut-point, (3) split the range of continuous values according to the cut-point, and (4) stop when a stopping criterion satisfies.

In this study, discretization of explanatory variables of event history dataset assumed unsupervised, static, global and direct in order to reach a top-down splitting approach and transformation of all types of variables in dataset into a logical (binary) format. Briefly, static discretization is dependent of classification task, global discretization uses the entire observation space to discretize, and direct methods divide the range of *k* intervals simultaneously. For a comprehensive study of discretization process, see (Liu, Huan, et al., 2002).

Data Reduction and Variable Selection Methods

Data reduction techniques are categorized in three main strategies, including dimensionality reduction, numerosity reduction, and data compression (Han, J. et al, 2006) (Tan, P. et al., 2006). Dimensionality reduction as the most efficient strategy in the field of large-scale data deals with reducing the number of random variables or attributes in the special circumstances of the problem. All dimensionality reduction techniques are also classified as feature extraction and feature selection approaches. Feature Extraction is defined as transforming the original data into a new lower dimensional space through some functional mapping such as PCA and SVD (Motoda, H., and Huan, L., 2002) (Addison, D. et al., 2003). Feature selection is denoted as selecting a subset of the original data (features) without a transformation in order to filter out irrelevant or redundant features, such as filter methods, wrapper methods and embedded methods (Saeys, Y. et al., 2007) (Guyon, I., and Elisseeff, A., 2003).

Variable selection is a necessary step in a decision-making process dealing with a large-scale data. There is always uncertainty when researchers aim to collect most important variables specifically in the presence of big data. Variable selection for decision-making in many fields is mostly guided by expert opinion (Casotti, M., n.d.). The computational complexity of all the possible combinations of the p variables from size 1 to p, could be overwhelming, where the total number of combinations are $2^p - 1$. For example, for a dataset of 20 explanatory variables, the number all possible combinations is $2^{20} - 1 = 1048575$.

Next section presents proposed methodology for multidisciplinary decision-making approach based on proposed analytical model, designed methods and heuristic algorithms for explanatory variable subset selection.

3. METHODOLOGY

In this section, first proposed analytical model for transformation of the explanatory variable dataset to reach the logical representation as a sort of binary variables is presented. Next, in order to select most significant variables in terms of inefficiency, designed variable selection methods and heuristic clustering algorithms are introduced.

3.1. Logical model

A multipurpose and flexible model for a type of event history data with a large number of variables when the correlation between variables is complicated or unknown is presented. The logical model is to simplify the original covariate dataset into a logical dataset by transformation lemma. Next, we show the validation of this designed logical model by correlation transformation (Sadeghzadeh, K., and Fard, N, in press) (Sadeghzadeh, K., and Fard, N, 2014).

The original event history dataset may include any type of explanatory. Many time-independent variables are even binary or interchangeable with a binary variable such as dichotomous variable. Also, interpretation of binary variable is simple, understandable and comprehensible. In addition, the model is appropriate for fast and low-cost calculation. The General schema of high-dimensional event history dataset includes *n* observations with *p* variables as shown in Table 1.

Table 1: Schema for high-dimensional event history dataset

Obs. #	Time to Event	Variables			
		Var. 1	Var. 2	…	Var. *p*
1	t_1	u_{11}	u_{12}	…	u_{1p}
2	t_2	u_{21}	u_{22}	…	u_{2p}
…	…	…	…	…	…
n	t_n	u_{n1}	u_{n2}	…	u_{np}

Each array of *p* variables vectors will take only two possible values, canonically 0 and 1. As discussed in Section 2, discretization method is applied to values by dividing the range of values for each variable into 2 equally sized parts. We define w_{ij} as an initial splitting criterion equal to arithmetic mean of maximum and minimum value of u_{ij} for $i = 1 \dots n, j = 1 \dots p$. The criteria w_{ij} could be defined by expert using experimental or historical data as well. For any array u_{ij} in the *n*-by-*p* dataset matrix $\boldsymbol{U} = [u_{ij}]$, then allocate a substituting array v_{ij} as 0 if $u_{ij} < w_{ij}$ and 1 if $u_{ij} \geq w_{ij}$. The proposed model assumes any array with a value of 1 as desired for expert and 0 otherwise. In other words, $v_{ij} = 0$ represent the lack of the *j*th variable in the *i*th observation. The result of the transformation is an *n*-by-*p* dataset matrix $\boldsymbol{V} = [v_{np}]$ which will be used in the following methods and algorithms. Also, we define time-to-event vector $\boldsymbol{T} = [t_n]$ including all observed event times. The logical model initially could be

satisfied by proper design of data collection process by based on Boolean logic to generate binary attributes.

To validate the robustness of this model we show that the change of correlation between variables before and after transformation is not significant and the logical dataset has followed the same pattern and behavior as the original; in terms of correlation of covariates. We define correlation matrix for each of original and transformed dataset based on Pearson product-moment correlation coefficient; $M = [m_{ij}]$ and $N = [n_{ij}]$ where $i = 1 \dots n, j = 1 \dots p$, where n_{ij} and m_{ij} denote covariance of variables i and j for original and transformed dataset respectively as follows:

$$n_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(u_{ik} - \bar{u}_i)(u_{jk} - \bar{u}_j) \qquad i = 1 \dots p, j = 1 \dots p \tag{6}$$

$$m_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j) \qquad i = 1 \dots p, j = 1 \dots p \tag{7}$$

where (u_{ik}, v_{ik}) and $(\bar{u}_i, \bar{v}_i)$ represent value of variable i in observation k and mean of variable i in each dataset respectively, and similarly the second parenthesis in equations(6) and (7) are defined for variable j.

The experimental fitted line for the scatter plot of m_{ij} and n_{ij} for any dataset is $y = a + bx$ where b is positive small and a is not significant. For instance, Figure 1 shows the primary biliary cirrhosis (PBC) dataset (Section 4) for an experimental result of an uncensored data with the fitted line of $y = 0.6356x + 0.0116b$.

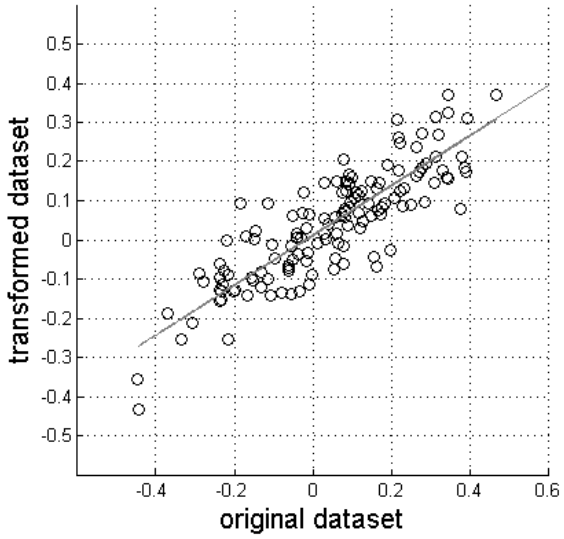

Figure 1: Comparison of covariate correlations in the original and the transformed dataset. Fitted polynomial for the uncensored PBC dataset (Section 4) is $y = 0.0116 + 0.6356x$

The proposed logical model validation and verification of the robustness were presented comprehensively in (Sadeghzadeh, K., and Fard, N, in press) and (Sadeghzadeh, K., and Fard, N, 2014).

In order to select the most significant variables in terms of inefficiency, methods and algorithms are presented next.

3.2. Designed Methods and Heuristic Algorithms

We design a class of methods applying on proposed logical model to select inefficient variables in a high-dimensional event history datasets. The major assumption to design appropriate methods for this purpose is that the variable which is completely inefficient solely can provide a significant performance improvement when engaged with others, and two variables that are inefficient by themselves can be efficient together (Guyon, I., and Elisseeff, A., 2003). Based on this assumption, we design three methods and heuristic algorithms to select inefficient variables in event history datasets with high-dimensional covariates. We use Kaplan-Meier estimator in this study to estimate survival probabilities as a function of time. The *n*-by-*p* matrix $\boldsymbol{V}$ is the prepared transformed logical dataset according to Section 3.1, where *n* is the number of observations, *p* is the number of variables, and *k* is the estimated subset size to select for calculation parts in the algorithms.

Recalling $\boldsymbol{V}$ which is constructed by *k* observation vectors corresponding to each of the variables, $\boldsymbol{D} = [d_{kp}]$ as a *k*-by-*p* matrix is a selected subset of $\boldsymbol{V}$ and *k* is defined as the number of observations in any subset of $\boldsymbol{V}$, where $k \leq n$. For any variable *i*, we define vector $\boldsymbol{O}^i$ as a time-to-event vector which includes failure times of any observation *j* the value of v_{ij} is one. Similarly, we define vector $\boldsymbol{Z}^i$ including failure times of any observation *j* where the value of v_{ij} is zero. The vectors $\boldsymbol{R}$ and $\boldsymbol{S}$ are defined as follow:

$$r_i = \begin{cases} t_i, & \sum di. \geq 0 \\ 0, & Otherwise \end{cases} \qquad i = 1 \dots n \tag{8}$$

$$s_i = \begin{cases} t_i, & \prod di. = 1 \\ 0, & Otherwise \end{cases} \qquad i = 1 \dots n \tag{9}$$

Vector $\boldsymbol{R}$ is constructed by all non-zero arrays *r* and similarly vector $\boldsymbol{S}$ is constructed by all non-zero arrays *s*.

We propose three methods and algorithms to select inefficient variables as follows:

Singular Variable Effect Algorithm

The objective of Singular Variable Effect (SVE) method is to determine the efficiency of a variable by analyzing the effect of the presence of any variable singularly in comparison with its absence in a transformed logical dataset. For p variable, we aim to set vector $\boldsymbol{\Delta} = [\delta i]$ where $i = 1 \dots p$ to rank the efficiency of the variables. The preliminary step for the highest efficiency in this method is to initially clustering the variables based on the correlation coefficient matrix of original dataset, $\boldsymbol{M}$, and choose a representative variable from each highly correlated cluster and eliminate the other variables from the dataset. For instance, for any given dataset, if three variables are highly correlated, only one of them is selected randomly and the other two are eliminated from the dataset. The result of this process assures that the remaining variables for applying methods and heuristic algorithms are not highly correlated.

As an outcome of the SVE procedure, if one hopes to reduce the number of variables in the dataset for further analysis, could eliminate less efficient identified variables or if aims to concentrate on a reduced number of variables, could choose a category of more efficient identified variables as well. Heuristic algorithm for SVE method is:

---

for $i = 1$ to p do
Calculate $\boldsymbol{O}^i$ and $\boldsymbol{Z}^i$ for variable i observation vector in dataset $\boldsymbol{V}$
Compare $\boldsymbol{T}$ and $\boldsymbol{O}^i$ with Wilcoxon rank sum test
Save the test score for variable i as α_i
Compare $\boldsymbol{T}$ and $\boldsymbol{O}^i$ with Wilcoxon rank sum test
Save the test score for variable i as β_i
Calculate $\delta_i = \alpha_i - \beta_i$
end for
Return $\boldsymbol{\Delta} = [\delta_p]$ as the variable efficiency vector

---

Splitting Semi-Greedy Clustering Algorithm

Splitting Semi-Greedy (SSG) method to select an inefficient variable subset is proposed. A clustering procedure through randomly splitting approach to select the best local subset according to a defined criterion incorporated. In this method we use block randomization which is designed to randomize subjects into equal sample sizes groups. A nonparametric test is used to test a null hypothesis that whether two samples are drawn from the same distribution, as compared to a given alternative hypothesis. Wilcoxon rank sum test is used in this method.

The concept of this method is inspired by the semi-greedy heuristic (Feo, T. A., and Resende, M. G., 1995) (Hart, J. P., and Shogan, A. W., 1987) and tabu search (Gendreau, M., and Potvin, J. Y., 2005). Criterion of this search is similar to *The Nonparametric Test Score (NTS)* method (Sadeghzadeh, K., and Fard, N, in press) which is to collect the most inefficient variable subset via Wilcoxon rank sum test score. At each of l trials, all p variables from the transformed logical dataset $\boldsymbol{V}$ are randomly clustered into subsets of size k variables, where one cluster possibly contains less than k variables and the number

of clusters is equal to $\lceil p/k \rceil$. To calculated score summation for each variable over all trials, a randomization dataset matrix $\boldsymbol{\Xi} = [\xi_{lk}]$ where each row is formed by k variable identification numbers in any selected subsets for all l trials. Comprehensive experimental results for validation of the proposed methods by comparison with similar methods are presented next. Heuristic algorithm for SSG method is:

---

for i = 1 to l do
Split the data into equally sized subsets
Compose the dataset $\boldsymbol{D}$ for each subset
Calculate $\boldsymbol{R}$ over the $\boldsymbol{D}$ for each subset
Compare $\boldsymbol{T}$ and $\boldsymbol{R}$ with Wilcoxon rank sum test and save the test score for each subset one by one
Select a subset with the highest test score
Save the test score for variables in the selected subset as $\xi_{i(k+1)}$
end for
Assume $\boldsymbol{\Theta} = [\theta_p]$ as the reverse variable efficiency vector where initially each array as the cumulative contribution score corresponding to a variable is zero
for i = 1 to l do
for j = 1 to k do
Add the value of $\xi_{i(k+1)}$ to the cumulative contribution score θ_p of the variable i based on its identification number = ξ_{ij}
end for
end for
Return $\boldsymbol{\Theta} = [\theta_p]$ as the variable inefficiency vector

---

Weighted Time Score Algorithm

The Weighted Time Score (WTS) method is a variable clustering technique which selects set of size k variables from the transformed logical dataset $\boldsymbol{V}$ and calculates the score of each variable. The first step is to determine the observations in a selected subset which all k variables are 1 for that observation and eliminate other observation from subset. Cumulative time score over the vector $\boldsymbol{T}$ credit each of variables in the subset. Final score of all variables is reached by aggregation of those credits in l trials. Randomization algorithm randomly chooses a defined l subset of k from the $\boldsymbol{V}$, transformed logical dataset of p variable. We define a randomization dataset matrix $\boldsymbol{\Psi} = [\psi_{lk}]$ where each row is formed by k variable identification numbers in any selected subsets for overall l subsets. Heuristic algorithm for WTS method is:

---

for i = 1 to l do
Compose the dataset $\boldsymbol{D}_i$ for variable set i in $\boldsymbol{\Psi}$ including variables ψ_{ij} where j = 1 to k
Calculate $\boldsymbol{S}_i$ over the dataset $\boldsymbol{D}_i$
Calculate $\sum t_i$ for $\boldsymbol{S}_i$ as a time score
Save the time score for variables in subset i as $\psi_{i(k+1)}$
end for

Assume $\boldsymbol{\Omega} = [\omega_p]$ as the reverse variable efficiency vector where initially each array as the cumulative contribution score corresponding to a variable is zero
for $i = 1$ to l do
for $j = 1$ to k do
Add the value of $\psi_{i(k+1)}$ to the cumulative contribution score ω_p of the variable i based on its identification number = ψ_{ij}
end for
end for
Return $\boldsymbol{\Omega} = [\omega_p]$ as the variable inefficiency vector

The experiment results for these algorithms are followed in Section 4.

4. RESULTS AND ANALYSIS

To evaluate the performance of the designed methods, first well-known and publicly available primary biliary cirrhosis (PBC) dataset (Fleming and Harrington 1991) is considered as the sample collected dataset. These dataset includes 111 uncensored complete observations and 17 explanatory variables in addition to event times for each observation. In order to obtain an approximate value of desired number of variables in any selected subset, we use principal component analysis (PCA) scree plot criterion (Sadeghzadeh, K., and Fard, N, in press) (Sadeghzadeh, K., and Fard, N, 2014). For the original uncensored PBC dataset, approximation of this number is 3.

To verify the performance of the proposed methods, the result of these methods and algorithms for the transformed logical uncensored PBC dataset is compared with the results of Nonparametric Test Score (NTS) method (Sadeghzadeh, K., and Fard, N, in press), Random Survival Forest (RSF) method (Ishwaran, H. et al., 2008) (Ishwaran, I., and Kogalur, U. B., 2007), Additive Risk Model (ADD) (Ma, S., Kosorok, M. R., and Fine, J. P., 2006), and Weighted Least Square (LS) method (Huang, J., Ma, S., and Xie, H, 2006) for similar dataset variable selection, given in Table 1. A comprehensive comparison of NTS, RSF, ADD and LS performance with other relevant methods in high-dimensional time-to-event data analysis such as Cox's Proportional Hazard Model, LASSO and PCR has been presented in (Huang, J., Ma, S., and Xie, H, 2006) (Ishwaran, H. et al., 2008) (Ma, S., Kosorok, M. R., and Fine, J. P., 2006).

Each number in Tables 2 and 3 represents a specific variable in experiment dataset. For example, in Table 2, variable #1 is a selected as an inefficient variable by all methods.

Table 2: Selected inefficient variables in all proposed methods and comparison to NTS, RSF, ADD, and LS method results

Method	Selected Inefficient Variables
SVE	1, 3, 5, 10, 13, 17
SSG	1, 3, 5, 10, 15, 17

WTS	1, 3, 5, 10, 15, 17
NTS	1, 3, 5, 6, 10, 15, 17
RSF	1, 3, 5, 12, 13, 14, 15, 17
ADD	1, 2, 5, 12, 14
LS	1, 2, 3, 14, 15, 17

From the results shown on Tables 2, the SSG and WTS methods have a same performance. More than 80% of inefficient variables which has been detected by other methods (NTS, RSF, LS and ADD) are collected by proposed algorithms at significantly shorter calculation period, where the robustness of this class of methods has examined for several sample datasets.

To show variable inefficiency through three designed methods SVE, SSG, and WTS, graphical representation for the experiment results for uncensored PBC dataset is depicted in Figure 2. Each variable with larger radius and more distance from the center is less efficient and an ideal candidate to remove from dataset if it is desired.

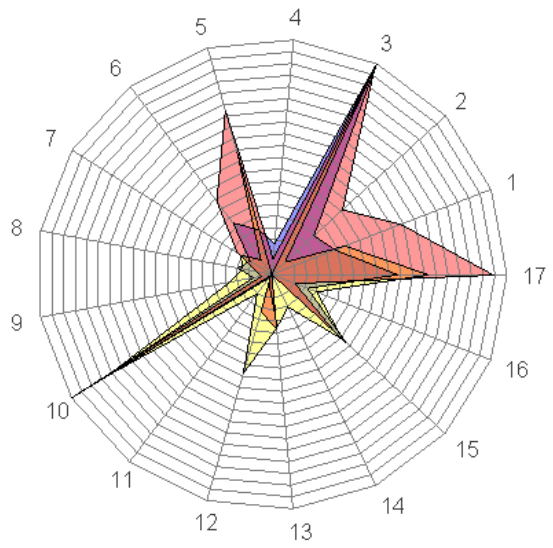

Figure 2: Radar plot of inefficient variables: Normalized inefficiency results from the transformed logical uncensored PBC dataset by SVE algorithm (red), SSG algorithm (green), and WTS algorithms (yellow).

As another validation of the proposed methods, a simulation is designed. We set n = 400 observations and p = 15 variables and simulated event times from a pseudorandom algorithm. We also set first five variables inefficient, where first two are absolutely inefficient. Some variable vectors are set as a linear function of event time data in addition to constant and periodic binary numbers as well as normal and exponential distributed pseudorandom numbers as independent values of explanatory variables. The results of methods and algorithms applying the simulated data are presented in Table 3. These results are compared with the results from NTS method. From the simulation defined pattern the comparison verifies the performance of all proposed methods.

Table 3: Selected inefficient variables in all proposed methods and comparison to NTS results and simulation defined pattern

Method	Selected Inefficient Variables (No.)
SVE	1, 2, 3, 10
SSG	1, 2, 3
WTS	1, 2, 3, 5, 12
NTS	1, 2, 5
Definition	1, 2, 3, 4, 5

Inefficiency analysis results for the simulation experiment shows that variables with identification number 1, 2 and 3 are detected as inefficient variables by all proposed methods. To reduce the number of variables in the dataset for further analysis, these explanatory variables are the best candidates to be eliminated from the dataset.

5. CONCLUSIONS

The proposed logical model, designed variable selection methods, and heuristic clustering algorithms in this paper are beneficial to explanatory variable reduction through an inefficient variable selection approach to obtain an appropriate variable subset in high-dimensional and large-scale event history data in order to avoid difficulties in decision-making.

By using such novel methods in the fields of economics, business and management, data analysis and decision-making processes will be faster, simpler and more accurate. For example, in business applications, many explanatory variables in a customer survey are defined based on cause and effect analysis process data or similar analytic process outcome. In most cases, correlations of these explanatory variables are complicated and unknown, and it is important to simply understand the efficiency of each variable in the survey. These procedures potentially applicable solutions for many problems in a vast area of science and technologies are presented.

Next step in this study is to considering event data and time-to-event models including new types of dependent variables through well-known models such as accelerated failure time and applying heuristic algorithms especially in the field of artificial intelligence.

REFERENCES

Addison, D. et al., 2003. *A Comparison of Feature Extraction and Selection Techniques.* s.l., s.n.

Allison, Paul D., 1984. *Event History Analysis: Regression for Longitudinal Event Data.* s.l.:Sage.

Brown, B. et al., 2011. *Are you ready for the era of 'big data',* s.l.: McKinsey Quarterly.

Brynjolfsson, E., 2012. *A Revolution in Decision-Making Improves Productivity,* s.l.: MIT.

Busemeyer, J. R., and Townsend, J. T., 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological Review.*

Casotti, M., n.d. *Variable Selection Methods: An Introduction,* s.l.: Molecular Descriptors.

Feldman, D. et al., 2013. Turning Big Data into Tiny Data: Constant-Size Coresets for k-means, PCA and Projective Clustering. *SODA.*

Feo, T. A., and Resende, M. G., 1995. Greedy Randomized Adaptive Search Procedures. *Journal of Global Optimization.*

Gendreau, M., and Potvin, J. Y., 2005. Tabu Search. In: *Search Methodologies.* s.l.:Springer.

Guyon, I., and Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research.*

Han, J. et al, 2006. *Data Mining: Concepts and Techniques.* s.l.:Morgan Kaufmann.

Hart, J. P., and Shogan, A. W., 1987. Semi-Greedy Heuristics: An Empirical Study. *Operations Research Letters.*

Hellerstein, J., 2008. *Parallel Programming in the Age of Big Data,* s.l.: Gigaom Blog.

Holford, T. R., 2002. *Multivariate Methods in Epidemiology.* s.l.:Oxford University Press.

Hosmer D. W. Jr., and Lemeshow, S., 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* s.l.:Wiley.

Huang, J., Ma, S., and Xie, H, 2006. Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. *Biometrics.*

Ishwaran, H. et al., 2008. Random Survival Forests. *The Annals of Applied Statistics.*

Ishwaran, I., and Kogalur, U. B., 2007. *Random Survival Forests for R,* s.l.: Rnews.

Kalbfleisch, J. D., and Prentice, R. L., 2011. *The Statistical Analysis of Failure Time Data.* s.l.:John Wiley & Sons.

Lee, E. T., and Wang, J. W., 2013. *Statistical Methods for Survival Data Analysis.* s.l.:John Wiley & Sons.

Liu, Huan, et al., 2002. Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery.*

Lohr, S., 2012. *The Age of Big Data,* s.l.: New York Times.

Ma, S., Kosorok, M. R., and Fine, J. P., 2006. Additive Risk Models for Survival Data with High-Dimensional Covariates. *Biometrics.*

Manyika, J. et al., 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity,* s.l.: McKinsey Global Institute.

Moran, J., 2013. *Is Big Data a Big Problem for Manufacturers?,* s.l.: Sikich.

Motoda, H., and Huan, L., 2002. Feature Selection, Extraction and Construction. *Communication of IICM.*

Sadeghzadeh, K., and Fard, N, 2014. *Applying Data Clustering and Data Reduction Methods in High-Dimensional Survival Data Analysis.* s.l., s.n.

Sadeghzadeh, K., and Fard, N, in press. Nonparametric Data Reduction Approach for Large-Scale Survival Data Analysis. *IEEE Xplore.*

Sadeghzadeh, K., and Salehi, M. B., 2010. Mathematical Analysis of Fuel Cell Strategic Technologies Development Solutions in the Automotive Industry by the TOPSIS Multi-Criteria Decision Making Method. *International Journal of Hydrogen Energy.*

Saeys, Y. et al., 2007. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics.*

Segaran, T., and Hammerbacher, J., 2009. *Beautiful Data: The Stories Behind Elegant Data Solutions.* s.l.:s.n.

Tan, P. et al., 2006. *Introduction to Data Mining.* s.l.:WP Co.

Yao, F., 2007. Functional Principal Component Analysis for Longitudinal and Survival Data. *Statistica Sinica.*