Gene selection in Cox regression model based on a new adaptive elastic net penalty

Oday Isam Alskaal *  
odayesam@gmail.com

Zakariya Yahya Algamal**  
zakariya.algamal@uomosul.edu.iq

Abstract:
The common issues of high dimensional gene expression data for survival analysis are that many of genes may not be relevant to their diseases. Gene selection has been proved to be an effective way to improve the result of many methods. The Cox regression model is the most popular model in regression analysis for censored survival data. In this paper, a new adaptive elastic net penalty with Cox regression model is proposed, with the aim of identification relevant genes and provides high classification accuracy, by combining the Cox regression model with the weighted L1-norm. Experimental results show that the proposed method significantly outperforms two competitor methods in terms of the area under the curve and the number of the selected genes.

Keywords: Cox regression model; penalized method; elastic net, gene selection.

This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction
The problem of analyzing time to event data arises in a number of applied fields, such as medicine, biology, public health, and epidemiology (Cockeran et al., 2019; Emura et al., 2012). Nowadays, high dimensional gene expression data are increasingly used for modeling various clinical outcomes to facilitate disease diagnosis, disease prognosis, and prediction of treatment outcome (Jian Huang et al., 2014).

Regression modeling is a standard practice to study jointly the effects of multiple predictors on a response. The Cox regression model is ubiquitous in the analysis of time-to-event data. When the number of predictors is large, building a Cox regression model including all of them is undesirable because it has low prediction accuracy and is hard to interpret (Karabey & Tutkun, 2017; Leng & Helen Zhang, 2006). For these reasons, variable selection has become an important focus in Cox regression modeling.

Penalized methods are very effective variable selection methods. These methods combine the Cox regression model with a penalty to perform variable selection and estimation simultaneously.
Gene selection in Cox regression model based on a new adaptive elastic net penalty

With different penalties, several Cox regression models can be applied, among which are, LASSO, which is called the least absolute shrinkage and selection operator (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), elastic net (Zou & Hastie, 2005), adaptive LASSO (Zou, 2006), and adaptive elastic net (Zou & Zhang, 2009).

Unquestionably, elastic net is considered to be one of the most popular procedures in the class of penalized methods. However, elastic net has a limitation: It applies the same amount of the penalty to all variables. Thus it is an inconsistent variable selection method (Algamal & Lee, 2015; Zou & Zhang, 2009).

To increase the power of informative gene selection, in the present study, a new adaptive elastic net with Cox regression model is proposed. More specifically, a new weight inside L1-norm is proposed, which can correctly reduce the estimation error. This weight will reflect the importance amount of each gene. Experimentally, comparisons between our proposed gene selection method and other competitor methods are performed. The experimental results prove that the proposed method is very effective for selecting the relevant genes with high prediction accuracy.

2. Panelized Cox regression model

Survival analysis is the statistical branch studying time-to-event data, or more precisely the time elapsing from a well-defined initiating event to some particular endpoint. The Cox regression model is one of the most popular and useful models in survival analysis (Cox, 1972).

Consider an analysis with time-to-event outcome, we denote the observed triplet as \( \{(t_i, \delta_i, x_i) : i = 1, \ldots, n\} \) where \( t_i \) is the survival time if \( \delta_i = 1 \) and censored time if \( \delta_i = 0 \) and \( x_i = (x_{i1}, \ldots, x_{ip}) \) is a \( p \)-dimensional explanatory variables. Under the proportional hazards framework, the Cox regression model (CRM) can be defined as

\[
h(t_i | x_i) = h_0(t_i) \exp\left( \beta^T x_i \right),
\]

where \( h_0(t_i) \) is the baseline hazard function and \( \beta = (\beta_1, \ldots, \beta_p)^T \) is a \( p \times 1 \) vector of unknown regression coefficients. Assuming that the subjects are statistically independent of each other, the joint probability of all realized events is the following partial likelihood

\[
L(\beta) = \prod_{i=1}^{n} \frac{\exp(\beta^T x_i)}{\sum_{j \in R_i} \exp(\beta^T x_j)}.
\]

where \( R_i \) is the set of subjects that are at risk just before time \( t_i \).

The estimation of the regression parameters of Eq. (1) is commonly carried out by minimizing the partial log likelihood function (Eq. (2)) as
\[ \hat{\beta}_{CRM} = \arg\min_{\beta} \left( \log L(\beta) \right) = -\sum_{i=1}^{n} \left[ \beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j) \right]. \] (3)

Panelized Cox regression model (PCRM) adds a nonnegative penalty term to Eq. (1), such that the size of variable coefficients can be controlled. Several penalty terms have been discussed in the literature considering the Cox regression model (Du et al., 2010; Fu et al., 2017; Gui & Li, 2005; Hossain & Ahmed, 2014; Hou et al., 2013; H. H. Huang & Liang, 2018; J. Huang et al., 2013; Jiang & Liang, 2018; Kauermann, 2005; Li et al., 2014; Lin & Halabi, 2017; Liu et al., 2014; Park & Ha, 2018; Shi et al., 2019; Suchting et al., 2019; Wang et al., 2019; Wu et al., 2012; Zhang & Lu, 2007). The LASSO method, proposed by Tibshirani (1996), is one of the popular penalty terms. The LASSO performs variable selection and estimation simultaneously by constraining the log-likelihood function of variable coefficients. Generally, the PCRM is defined as

\[ PCRM = \sum_{i=1}^{n} \left[ \beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j) \right] - \lambda P(\beta), \] (4)

where \( \lambda P(\beta) \) is the penalty term that regularized the estimates. The penalty term depends on the positive tuning parameter, \( \lambda > 0 \), which controls the tradeoff between fitting the data to the model and the effect of the regularization. In other words, it controls the amount of shrinkage. For the \( \lambda = 0 \), we obtain the CRM solution in Eq. (3). In contrast, for large values of \( \lambda \), the influence of the penalty term on the coefficient estimates increases.

Without loss of generality, it is assumed that the explanatory variables are standardized, \( \sum_{i=1}^{n} x_{ij} = 0 \) and \( (n^{-1})\sum_{i=1}^{n} x_{ij}^2 = 1 \), \( \forall j \in \{1, 2, ..., p\} \). The estimation of the vector \( \beta \) using LASSO is obtained by minimizing Eq. (4) as (Bradic et al., 2011; Goeman, 2010; Tibshirani, 1997)

\[ \hat{\beta}_{LASSO}^{Cox} = \arg\min_{\beta} \left[ -\sum_{i=1}^{n} \left[ \beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j) \right] + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \] (5)

Equation (5) can be efficiently solved by the coordinate descent algorithm (Simon et al., 2011).

The LASSO has an advantage in that it is computationally feasible in high-dimensional data. On the other hand, the LASSO has a drawback. The LASSO lacks the oracle properties, as stated in Fan and Li (Fan & Li, 2001) because it is equally penalize all the coefficients. In addition, the LASSO cannot handle the effect of grouping. When the pairwise correlations among a group of genes are very high, then the LASSO tends to select only one gene from the whole group and does not take into account which one is selected. To alleviate the first drawback, Zou (2006) and proposed the adaptive LASSO in which adaptive weights are used for penalized different coefficients in the penalty. The basic idea behind the adaptive LASSO is that by assigning a higher weight to the small
coefficients and lower weight to the large coefficients, it is possible to reduce the selection bias; therefore, it can consistently select the model. Furthermore, the adaptive LASSO solution is continuous from its definition, which enables it to enjoy oracle properties.

To deal with the second drawback, Zou and Hastie (2005) proposed the elastic net penalty. However, the elastic net does not enjoy the oracle property even though it performs much better in classification accuracy. As a result, the adaptive elastic net was proposed by Zou and Zhang (2009).

In panelized Cox regression model, the elastic net (ELASTIC) (Suchting et al., 2019) and its adaptive version (AELASTIC) are defined, respectively, as

\[
\hat{\beta}_{Elastic}^{Cox} = \arg\min_{\beta} \left[ -\sum_{i=1}^{n} \beta^T x_i - \log \sum_{j=1}^{m} \exp(\beta^T x_j) \right] + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2,
\]

where \( \lambda_1 \) and \( \lambda_2 \) are two non-negative tuning parameters and \( w_j = (w_{j1},...,w_{jp})^T \) is a data-driven weight vector. It depends on the root \( n \)-consistent initial values of \( \hat{\beta} \) and \( w_j = (|\hat{\beta}_j|)^{\gamma} \), where \( \gamma \) is a positive constant. For the low dimensional data, initial values of \( \hat{\beta} \) can be the unpenalized maximum partial likelihood estimator. While in the case of the high dimensional data, initial values of \( \hat{\beta} \) can be the elastic net estimates.

3. The proposed weight

In the context of gene expression data problems, the goal of gene selection is to improve prediction performance, to provide faster and more cost-effective genes, and to achieve a better knowledge of the underlying problem. High dimensionality can negatively influence the performance of the Cox regression regression model by increasing the risk of overfitting and lengthening the computational time. Therefore, removing irrelevant and noisy genes from the original microarray gene expression data is essential for applying Cox regression regression model to analyze the microarray gene expression data.

It is worthwhile to highlight that our contribution of this paper comes from the following issues. First, although PCRM with ELASTIC can be applied directly to the high dimensional gene expression data, this method may select irrelevant genes because ELASTIC has the inconsistent property in gene selection. In other words, the estimates of the PCRM with ELASTIC can be biased for large coefficients because larger coefficients will take larger penalties. Second, in PCRM, the genes are usually standardized. However, the standardization process may be unreasonable when the variances of genes showing important effect.
Motivated by these issues, a consistent identification of the true underlying genes is essential to improve the classification accuracy. As a result, the standard deviation for each gene is proposed as a weight inside $L_1$-norm, where

$$w_j = \frac{1}{s\hat{d}_j}, \quad j = 1, 2, \ldots, p,$$

where $s\hat{d}_j$ is the standard deviation for each gene. According to Eq. (8), the gene with low value of standard deviation will receive relatively large amount of weight, while the gene with high value of standard deviation will receive small amount of weight. By this weighting procedure, the ELASTIC can reduce the inconsistent property in gene selection.

The detailed of the our proposed weight computation is described in as

**Step 1:** Find $w_j, \quad j = 1, 2, \ldots, p$.

**Step 2:** Define $\tilde{x}_i = w_j x_i$.

**Step 3:** Solve the PCRM,

$$\hat{\beta}_{\text{Proposed}} = \arg \min_{\beta} \left\{ -\sum_{i=1}^{n} \left[ \beta^T \tilde{x}_i - \log \sum_{j=1}^{p} \exp(\beta^T \tilde{x}_j) \right] + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

4. **Real Application**

To evaluate the performance of the proposed method, three real gene datasets were used. A brief introduction and summary of the used datasets are given in Table 1. The first dataset is the Diffuse large B-cell lymphoma dataset (DLBCL) (Rosenwald et al., 2002). There are 240 lymphoma patients’ samples. Each patient’s data consists 7399 gene expression measurements, and its survival time, including censored or not.

The second dataset is the Dutch breast cancer dataset (DBC) (van Houwelingen et al., 2006). In this dataset, there was 295 breast cancer patients’ information collected in this dataset. Each patient’s data consist 4919 gene expression measurements.

The third dataset is the Lung cancer dataset (LC) (Beer et al., 2002). This dataset contains 86 lung cancer patients’ information including 7129 gene expression measurements, survival time and whether the survival time is censored.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample</th>
<th>Gene</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>240</td>
<td>7399</td>
<td>102</td>
</tr>
<tr>
<td>DBC</td>
<td>295</td>
<td>4919</td>
<td>207</td>
</tr>
<tr>
<td>LC</td>
<td>86</td>
<td>7129</td>
<td>62</td>
</tr>
</tbody>
</table>
To demonstrate the usefulness of the proposed method, comparative experiments with the ELASTIC and AELASTIC are conducted. To do so, each gene expression dataset is randomly partitioned into the training dataset and the test dataset, where 70% of the sample are selected for training dataset and the rest 30% are selected for testing dataset. For a fair comparison and for alleviating the effect of the data partition, all the used methods are evaluated, for their classification performance metrics using 10 folds cross validation, averaged over 100 partitioned times. Depending on the training dataset, the tuning parameter value, $\lambda$, for each used method was fixed as $0 \leq \lambda_1, \lambda_2 \leq 50$. To assess how well the model predicts the outcome, the idea of time-dependent receiver-operator characteristics (ROC) curves for censored data and area under the curve (AUC) as our criteria. The real application results are summarized in Tables 2 – 4.

Table 2 shows the average results of different used methods applied to the three real datasets. It is obviously that the numbers of genes selected by the proposed method are much more than those of the AELASTIC and the ELASTIC method. Among the other two methods, the proposed method selected the largest subset of genes. For example, in DBC dataset, the proposed method selected 97 gens out of 4919 genes comparing to 89 and 91 selected genes by ELASTIC and AELASTIC, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ELASTIC</th>
<th>AELASTIC</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>126</td>
<td>133</td>
<td>147</td>
</tr>
<tr>
<td>DBC</td>
<td>89</td>
<td>91</td>
<td>97</td>
</tr>
<tr>
<td>LC</td>
<td>79</td>
<td>88</td>
<td>101</td>
</tr>
</tbody>
</table>

In order to test the prediction accuracy of the different used methods, their average values of AUC for both the training and testing dataset were given in Tables 3 and 4, respectively. In the observation of Table 3, in terms of AUC, the proposed method achieved a maximum accuracy of 96.1%, 95.8% and 97.8% for DLBCL, DBC, and LC datasets, respectively. Furthermore, it is clear from the results that the proposed method outperformed the AELASTIC for all datasets. This improvement in AUC is mainly due to the proposed method ability in taking into account the new weight. Moreover, the proposed method improved the classification accuracy compared to ELASTIC. The improvements were 7.9%, 6.7%, and 6.7% for the DLBCL, DBC, and LC datasets, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ELASTIC</th>
<th>AELASTIC</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>0.882</td>
<td>0.922</td>
<td>0.961</td>
</tr>
<tr>
<td>DBC</td>
<td>0.891</td>
<td>0.929</td>
<td>0.958</td>
</tr>
<tr>
<td>LC</td>
<td>0.911</td>
<td>0.939</td>
<td>0.978</td>
</tr>
</tbody>
</table>
It can also be seen from Table 4 that the proposed method has the best results in terms of the AUC for the testing dataset. The proposed method has the largest AUC of 94.1%, 94.0%, and 95.2% for the DLBCL, DBC, and LC datasets, respectively. This indicated that the proposed method significantly succeeded in identifying the patients who are in fact having the cancer with a probability of greater than 0.94.

Table 4: The AUC results for the testing dataset

<table>
<thead>
<tr>
<th></th>
<th>ELASTIC</th>
<th>AELASTIC</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>0.861</td>
<td>0.917</td>
<td>0.941</td>
</tr>
<tr>
<td>DBC</td>
<td>0.825</td>
<td>0.918</td>
<td>0.940</td>
</tr>
<tr>
<td>LC</td>
<td>0.891</td>
<td>0.927</td>
<td>0.952</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper presents a new adaptive penalized Cox regression model by combining the Cox regression model with the weighted elastic net penalty to identify the relevant genes in gene expression data. Our proposed method was experimentally tested and compared with other existing methods. The superior prediction performance of the proposed method was shown through the AUC. Meeting this criterion nominates the proposed method as a promising gene selection method.

References


