Communications for Statistical Applications and Methods
2024, Vol. 31, No. 6, 703–708

https://doi.org/10.29220/CSAM.2024.31.6.703
Print ISSN 2287-7843 / Online ISSN 2383-4757

Comparison of the performance of various kernels for the survival prediction model

Seungyeoun Lee[1,a], Nayeon Kim[a], Beomseok Kim[b], Inyoung Kim[c]

[a]Department of Mathematics and Statistics, Sejong University, Korea;
[b]Department of Statistics, Korea University, Korea;
[c]Department of Statistics, Virginia Tech, Blacksburg, USA

Abstract

With the development of high-throughput technologies for producing genomic data, more advanced statistical methods such as regularization and machine learning techniques have been adapted to survival analysis. However, the clinical information such as age, gender and medical history plays a critical role in constructing a survival prediction model. Machine learning technique such as the support vector machine (SVM) can improve the predictability of the survival model by using the available clinical information. When implementing SVM for a predictive survival model, the clinical kernel was proposed by Daemen *et al.* by equalizing the influence of clinical variables and taking account of the range of these variables. However, this clinical kernel uses the same weight for all clinical variables without considering the different effect of those variables on the survival time. In this study, we proposed a simple kernel, called ensemble kernel, by combining a clinical kernel with model fitting. Since the proposed ensemble kernel is based on model fitting, two different kernels are considered by using either Cox model or accelerated failure time (AFT) model. We compare the performance of these two ensemble kernels with that of the linear kernel and the clinical kernel by the concordance index (C-index) using the four real data sets. While both linear and clinical kernels use all clinical variables in defining global kernels, the proposed two ensemble kernels can use only significant variables from either a Cox model or an AFT model. The comparative result shows that the proposed two ensemble kernels perform similarly as the existing clinical kernel does and the performance of four kernels vary according to data sets.

Keywords: kernel function, support vector machine, Cox model, accelerated failure model, C-index

1. Introduction

With the advent of high-throughput genotyping techniques, more advanced statistical methods such as regularization and machine learning techniques have been adapted to survival analysis. However, the clinical information such as age, gender, tumor size and the medical history is more critical in predicting the survival outcome for most diseases and examinations. Furthermore, it has shown that the clinical variables have similar power as profiles obtained from high-throughput technologies. To improve the predictability of the survival model, a machine learning technique such as the support vector machine (SVM) has been widely used by considering non-linear and more complex interactions between variables. When implementing SVM, the linear kernel function has several disadvantages as described in Daemen and Moore (2009), and Daemen *et al.* (2012) because it is based on the

[1]Corresponding author: Department of Mathematics and Statistics, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea. E-mail: leesy@sejong.ac.kr

Published 30 November 2024 / journal homepage: http://csam.or.kr

inner product for one or several variables. The inner product for continuous variables depends on the variable range (e.g. age from 20 to 80 years vs. Karno score from 0 to 10). For ordinal variables, the comparison of two patients with value 1 and 2 depends on the range of this variable. These patients will be less similar when the variable has three levels than when the variable has seven levels. Furthermore, when an ordinal variable has zero value for a possible category, the inner product for a patient with value zero will always be zero, independent of its dissimilarity with another patient. For a nominal variable, the inner product between two patients should only larger than zero when both patients have the same category. The clinical kernel proposed by Daemen and Moore (2009) and Daemen *et al.* (2012) standardizes the scale of continuous and ordinal variables to equally take account of all variables and then the total clinical kernel matrix is calculated as the sum of the individual kernel matrices from each clinical variable divided by the total number of clinical variables $\boldsymbol{p}$.

In this study, we propose to define the global clinical kernel matrix by taking the weighted average of the individual kernel matrices obtained from each clinical variable, in which the weight is the absolute value of the effect size of each clinical variable from either the fitted Cox model or an AFT model. Since the proposed method involves the fitting the model to obtain the effect size of each clinical variable, the proposed kernel is called ensemble kernel. According to the fitted model, we call it either the Cox ensemble kernel or the AFT ensemble kernel.

We compare the performance of two proposed ensemble kernels with that of linear and clinical kernels using the four different real data sets. The proposed ensemble kernels could use either all clinical variables or the subset of clinical variables by the selection of significant variables whereas the clinical kernel is supposed to use all variables. Depending on the selection of variables, we consider three different scenarios for the comparative study. In Scenario I, all clinical variables are considered, in Scenario II, only significant variables in a Cox model are considered, and in Scenario III, only significant variables in an AFT model are considered. For the comparison of the performance, data set is divided into training and test sets with 2 : 1 ratio and for the variable selection process, 5-fold-cross validation were implemented in Scenarios II and III. In addition, the performance of each kernel is compared by the coincidence index (C-Index) (Harrel *et al.*, 1982) using the four different clinical data sets.

2. Methods

Let $k_x(i, j)$ be the clinical kernel of clinical variable x between the individual i and j. The clinical kernel is defined as

$$k_x(i, j) = \frac{(\max - \min) - |x_i - x_j|}{(\max - \min)},$$

when a clinical variable x is continuous or an ordinary and

$$k_x(i, j) = \begin{cases} 1, & \text{if } x_i = x_j, \\ 0, & \text{if } x_i \neq x_j, \end{cases}$$

when a clinical variable x is nominal. These clinical kernels are combined as a global and heterogeneous kernel matrix, $K(i, j) = (1/p) \sum_{l=1}^{p} k_{xl}(i, j)$. Here the global kernel matrix is the sum of the individual clinical kernel divided by the number of variables.

We propose to take the weighted average of the individual clinical kernel, in which the weight is the effect size of each clinical variable from the fitted model. For example, the effect of the l^{th} clinical

Table 1: The covariate information of randomly selected three individuals from a dataset of pancreatic cancer

ID	Alcohol history	Postoperative	Sex	Maximum tumor	N stage
15	No	Yes	Male	4.6	3
42	Yes	No	Female	5.0	1
49	No	Yes	Female	3.2	1

variable is estimated as β_l from a Cox model. Then the ensemble global kernel matrix is defined as $K(i,j) = (\sum_{l=1}^{p} |\beta_l| k_{xl}(i,j))/\sum_{l=1}^{p} |\beta_l|$ between the individual i and j.

We propose two different ensemble kernels, called a Cox ensemble kernel and an AFT ensemble kernel according to the fitted model and compare the performance of these two ensemble kernels with that of linear kernel and clinical kernel using the four real data sets.

Here we compare the proposed ensemble kernels with the clinical kernel with an example. Suppose that we randomly select three individuals from a dataset of pancreatic cancer and their information is given in the Table 1.

Among five covariates, only one variable, maximum tumor, is continuous and the other four variables are categorical variables. The corresponding clinical kernel function to each covariate is defined as follows:

$$K_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},\ K_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},\ K_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},\ K_4 = \begin{bmatrix} 1 & 0.96 & 0.86 \\ 0.96 & 1 & 0.82 \\ 0.86 & 0.82 & 1 \end{bmatrix},\ K_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Here K_1, K_2, K_3, K_4 and K_5 are the clinical kernel functions corresponding to alcohol history, postoperative, sex, maximum tumor and N stage, respectively. Then the clinical global kernel matrix is the average of these five kernel functions given as

$$K_{cli} = \frac{K_1 + K_2 + K_3 + K_4 + K_5}{5} = \begin{bmatrix} 1 & 0.1920 & 0.5720 \\ 0.1920 & 1 & 0.5640 \\ 0.5720 & 0.5640 & 1 \end{bmatrix}.$$

On the other hand, a Cox ensemble kernel and the an AFT ensemble kernel are given as

$$K_{Cox} = \frac{1.2814K_1 + 1.3718K_2 + 0.8209K_3 + 0.2191K_4 + 0.2545K_5}{3.9477} = \begin{bmatrix} 1 & 0.0533 & 0.7198 \\ 0.0533 & 1 & 0.3179 \\ 0.7198 & 0.3179 & 1 \end{bmatrix}.$$

$$K_{AFT} = \frac{0.6074K_1 + 0.5742K_2 + 0.3212K_3 + 0.0993K_4 + 0.2380K_5}{1.8401} = \begin{bmatrix} 1 & 0.0518 & 0.6886 \\ 0.0518 & 1 & 0.3481 \\ 0.6886 & 0.3481 & 1 \end{bmatrix}.$$

When comparing the elements of three global kernel matrices, those of the Cox ensemble kernel and the AFT ensemble kernel have the similar values while those of the clinical global kernel matrix are different from the previous two ensemble kernels. For example, the kernel value between the patient having ID of 15 and one having ID of 42 is 0.1920 in the clinical kernel whereas the corresponding values are 0.0533 and 0.0518 in the Cox ensemble and the AFT ensemble kernels, respectively. These two patients have different values for alcohol history, postoperative, sex and N stage except for maximum tumor size as shown in Table 1. Therefore, the closeness of these two patients should

Table 2: Comparison of test C-index under three scenarios I, II and III

Data set	**Kernel**	**Scenario I**		**Scenario II**		**Scenario III**	
		train C-index	**test C-index**	**train C-index**	**test C-index**	**train C-index**	**test C-index**
		mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
Pancreatic	Linear	0.6539 (0.0782)	0.6206 (0.0655)	0.6628 (0.0869)	0.6305 (0.0705)	0.6560 (0.0841)	0.6220 (0.0650)
	Clinical	**0.7648** (0.0369)	**0.6909** (0.0566)	**0.7686** (0.0425)	**0.6775** (0.0604)	**0.7685** (0.0360)	**0.6811** (0.0541)
	Cox sensemble	0.7525 (0.0392)	0.6814 (0.0559)	0.7635 (0.0472)	0.6653 (0.0616)	0.7603 (0.0366)	0.6709 (0.0557)
	AFT ensemble	0.7542 (0.0376)	0.6794 (0.0344)	0.7604 (0.0467)	0.6678 (0.0625)	0.7604 (0.0344)	0.6711 (0.0573)
Veteran	Linear	0.7154 (0.0170)	**0.6998** (0.0397)	0.7149 (0.0169)	**0.7008** (0.0350)	0.7146 (0.0173)	**0.7002** (0.0342)
	Clinical	**0.7909** (0.0337)	0.6967 (0.0383)	**0.7605** (0.0284)	0.6942 (0.0371)	**0.7607** (0.0245)	0.6965 (0.0370)
	Cox ensemble	0.7741 (0.0318)	0.6967 (0.0383)	0.7588 (0.0255)	0.6949 (0.0371)	0.7599 (0.0228)	0.6961 (0.0392)
	AFT ensemble	0.7760 (0.0168)	0.6960 (0.0388)	0.7582 (0.0254)	0.6945 (0.0368)	0.7592 (0.0250)	0.6954 (0.0388)
Lung	Linear	0.5325 (0.0168)	0.5255 (0.0389)	0.5653 (0.0415)	0.5559 (0.0455)	0.5616 (0.0390)	0.5598 (0.0486)
	Clinical	**0.6678** (0.0201)	**0.6407** (0.0325)	**0.6670** (0.0208)	**0.6316** (0.0341)	**0.6690** (0.0229)	**0.6304** (0.0369)
	Cox ensemble	0.6611 (0.0207)	0.6348 (0.0339)	0.6612 (0.0197)	0.6292 (0.0348)	0.6623 (0.0233)	0.6268 (0.0375)
	AFT ensemble	0.6619 (0.0154)	0.6347 (0.0338)	0.6622 (0.0210)	0.6282 (0.0347)	0.6638 (0.0230)	0.6264 (0.0362)
Melanoma	Linear	0.8775 (0.0251)	**0.8802** (0.0314)	0.8776 (0.0154)	0.8804 (0.3130)	0.8775 (0.0154)	**0.8802** (0.0314)
	Clinical	**0.8984** (0.0161)	0.8532 (0.0258)	**0.8965** (0.0284)	0.8576 (0.0264)	**0.8987** (0.0315)	0.8526 (0.0271)
	Cox ensemble	0.8795 (0.0161)	0.8750 (0.0311)	0.8843 (0.0171)	0.8821 (0.0332)	0.8803 (0.0167)	0.8780 (0.0313)
	AFT ensemble	0.8792 (0.0161)	0.8731 (0.0315)	0.8842 (0.0159)	**0.8824** (0.0323)	0.8803 (0.0162)	0.8775 (0.0320)

[*] Bold number represents the maximum C-index among four kernels

be very low based on the covariates and the two ensemble kernels reflect this difference more than the clinical kernel. On the other hand, the kernel value between the patient having ID of 15 and one having of ID 49 is 0.5720 in the clinical kernel whereas the corresponding values are 0.7198 and 0.6886 in the Cox ensemble and the AFT ensemble kernels, respectively. As shown on Table 1, these two patients belong to the same categories of alcoholic history and postoperative though they are male and female, have different N stages. The covariates of alcohol history and postoperative have the significant effect on the survival time with p-values of less than 0.05 while sex, N stage and maximum tumor have no significant effect. The large values in the Cox ensemble and the AFT ensemble kernels show the closeness between the patients having ID 15 and one having of ID 49 more than that in the clinical kernel. Since the variable selection process is implemented for fitting either a Cox model or an AFT model, we consider three different scenarios for the comparative study. In scenario I, all clinical variables were included, only a subset of variables selected from a Cox model is included in scenario II, and only a subset of variables selected from an AFT model is included in scenario III, respectively. For the variable selection process, 5-fold-CV is applied because the sample size is not enough for some data sets. In addition, this process was repeated 100 times and the value of C-index for the validation and test set is the average over 100 repeated samples.

3. Results

For the comparison study, we analysed four different datasets under three different scenarios described in the pervious section. Four datasets are the pancreatic cancer (Mok *et al.*, 2019), lung cancer (Kalbfleisch and Prentice, 1980), veteran's lung cancer (Loprinzi *et al.*, 1994) and melanoma (Andersen *et al.*, 1993) datasets. The pancreatic cancer data consists of 124 patients with 13 covariates, among which 73 patients were dead and 51 patients were alive at the end of study. The lung cancer data consists of 228 patients with 7 covariates, among which 165 patients were dead and 63 patients were censored. The veteran's lung cancer dataset has 137 patients with 6 covariates, among which 128 patients were dead and 9 patients were censored. Finally melanoma data consists of 205 patients, among which 57 patients were dead and 148 patients were censored.

We compare the performance of four different kernels under the three different scenarios using four datasets. Four different kernels are linear, clinical, Cox ensemble and AFT ensemble kernels. As

Table 3: Selected variables in a Cox model (Scenario II)

Data set	Variable (selection frequency among 100 repetitions)
Pancreatic	T_Stage(43), alcohol_history(36), history_of_chronic_pancreatitis(41), history_of_diabetes(45), person_neoplasm_cancer_status(87), postoperative(95), N_Stage(50), tobacco_smoking_history(54), maximum_tumor_dimension(54), radiation_therapy(52), residual_tumor(24), Age(100), Sex(100)
Veteran	karno(100), celltype(69), diagtime(43), trt(35), prior(32), Age(100)
Lung	ph.ecog(74), pat.karno(81), ph.karno(58), wt.loss(54), meal.cal(51), Age(100), Sex(100)
Melanoma	year(100), tickness(50), ulcer(17), Age(100), Sex(100)

Table 4: Selected variables in an AFT model (Scenario III)

Data set	Variable (selection frequency among 100 repetitions)
Pancreatic	T_Stage(51), alcohol_history(42), history_of_chronic_pancreatitis(42), history_of_diabetes(34), person_neoplasm_cancer_status(87), postoperative(99), N_Stage(49), tobacco_smoking_history(58), maximum_tumor_dimension(53), radiation_therapy(57), residual_tumor(26), Age(100), Sex(100)
Veteran	karno(100), celltype(78), diagtime(44), trt(46), prior(39), Age(100)
Lung	ph.ecog(76), pat.karno(76), ph.karno(46), wt.loss(46), meal.cal(52), Age(100), Sex(100)
Melanoma	year(100), tickness(47), ulcer(42), Age(100), Sex(100)

described in the previous section, all covariates were included in scenario I, only a subset of covariates were included in either scenario II or scenario III, in which only significant variables were selected under either a Cox model or an AFT model, respectively. Due to the possibly biased randomization when dividing whole data into training and test sets, we repeated this process 100 times and took an average of C-indices over 100 repetitions. Since a subset of selected variables may be different across 100 repetitions, we record how many times each variable was selected in the fitted model for scenarios II and III, in which both age and sex were always included. As shown in Table 2, the performance of both Cox and AFT ensemble kernels seems to be similar as that of clinical and linear kernels across three scenarios. When all covariates were considered in scenario I, either linear or clinical kernels always perform better than the proposed two ensemble kernels though the difference in C-index is not significantly different. In scenario II, an AFT ensemble kernel performs better than either linear or clinical kernels for the melanoma data while either linear or clinical kernels perform slightly better than the two ensemble kernels under scenario III. Since a subset of selected variables may be different across 100 repetitions, we record how many times each variable was selected in the fitted model for scenarios II and III, in which both age and sex were always included. These results are listed for a Cox model and an AFT model in Tables 3 and 4, respectively. As shown in Tables 3 and 4, the selection frequency of each covariate seems to be similar in two models except for the covariate of ‘ulcer’, the corresponding selection frequencies are 17 in a Cox model and 42 in an AFT model, respectively. In conclusion, the proposed ensemble kernels show the similar performance as the existing linear and clinical kernels and their C-index varies from 0.5255 to 0.8802 according to different datasets.

4. Discussion

We propose the two ensemble kernels, called a Cox ensemble and an AFT ensemble kernels, and compare their performance with that of the existing kernels such as linear and clinical kernels using the four real datasets. The clinical kernel considers the characteristic of clinical variables as well as their different range, and the closeness between individuals is well measured by the global kernel matrix by standardizing the effect of various clinical variables. However, the clinical kernel regards all clinical variables uniformly and uses the same weight on the global kernel matrix. On the other hand, the proposed ensemble kernels use the effect size of covariates as the weight on the global kernel matrix. Therefore, both a model fitting and a variable selection process are required for the proposed ensemble kernels. From the comparison result, both a Cox ensemble and an AFT ensemble kernels do not overpass the existing kernels except for one case but their performance is similar to the clinical kernel in terms of test C-index in most cases. However, this comparative study has limitations in that only four real data sets are analyzed. For more appropriate comparative study, an extensive simulation study is needed for further research.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Science, (2022R1F1A1074343).

References

Daemen A and Moor De B (2009). Development of a kernel function for clinical data. In *Proceedings of 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Minneapolis, MN, 5913–5917.

Daemen A, Timmerman D, Bosch T *et al.* (2012). Improved modeling of clinical data with kernel methods, *Artificial Intelligence in Medicine*, **54**, 103–114.

Harrel F, Califf R, Pryor D, Lee KL, and Rosati RA (1982). Evaluating the yield of medical tests, *Journal of American Medical Association*, **247**, 2543–2546.

Mok L, Kim Y, Lee S, Choi S, Lee S, Jang J-Y, and Park T (2019). HisCoM-PAGE: Hierarchical structural component models for pathway analysis of gene expression data, *Genes 2019*, **10**, 931.

Kalbfleisch D and Prentice R (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Loprinzi C, Laurie J, Wieand H *et al.* (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires, *North Central Cancer Treatment Group. Journal of Clinical Oncology*, **12**, 601–607.

Andersen P, Borgan O, Gill RD, and Keiding N (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Received July 03, 2024; Revised September 05, 2024; Accepted September 30, 2024