Article ID: WMC002167 2046-1690

Smooth Rank: A Method for Robust Risk Modeling for Smaller Samples

Corresponding Author:
Dr. Marina Sapir,
Scientist, Metapattern, 04609 - United States of America

Submitting Author:
Dr. Marina Sapir,
Scientist, Metapattern, 04609 - United States of America

Article ID: WMC002167

Article Type: Research articles

Submitted on:05-Sep-2011, 03:55:51 PM GMT **Published on:** 06-Sep-2011, 02:15:26 PM GMT

Article URL: http://www.webmedcentral.com/article_view/2167

Subject Categories:BIOSTATISTICS

Keywords: Survival Analysis, Risk Modeling, Prediction, Machine Learning, Small Samples

How to cite the article: Sapir M . Smooth Rank: A Method for Robust Risk Modeling for Smaller Samples . WebmedCentral BIOSTATISTICS 2011;2(9):WMC002167

Smooth Rank: A Method for Robust Risk Modeling for Smaller Samples

Author(s): Sapir M

Abstract

Prognosis of disease progression is necessary for development of individualized treatment, understanding of the disease. Risk modeling is a challenging problem, and too often amount of available relevant observations is not sufficient to build a quality model with traditional approaches.New method Smooth Rank for survival analysis, risk modeling is introduced here. Smooth Rank is robust against overfitting on relatively small samples. The method is compared with established risk modeling methods on 10 real life datasets. The experiments confirmed significant advantage of the proposed method on smaller samples.

Introduction

The goal of personalized medicine is to effectively match the right treatment strategy with the patient. This requires individualized prognosis of disease progression depending on the available types of treatment. The task of prognosis involves application of survival analysis to longitudinal accumulated data. Gathering data for survival analysis requires many months or years of studies, and it is practically difficult and expensive to assemble large sample of relevant observations.

Originally, survival analysis was considered as a subject of statistical exploration rather than a prognosis problem [1].

Consider a dataset, where each observations contains three components: covariate vector x, a positive survival time t and an event indicator d. The event indicator d is equal to 1 if an event (failure) occurred, and zero if the observation is (right) censored.

The commonly accepted criterion of the accuracy of the modeling is Harrell's concordance index [2] measuring agreement between the model's scores and the order of the failure times. The criterion is not directly related with any particular interpretation of the scores: any scores which correlate with the failure times will do.

The most popular method for survival analysis, Cox PH regression [1] has severe restriction on number of covariates which can be effectively used on any given sample. For example, the tutorial [2] suggests that the number of covariates should not exceed 1/10 of the number of non-censored observations. Often, this means that one either has to gather very large datasets, or exclude valuable predictors.

Most of advanced methods for prediction in survival analysis are developed to make the traditional approach more robust against overfitting on sparse data (see surveys in [3,4]). Overall, the two major directions for improvements involve (1) preliminary feature aggregation to lower problem dimensionality and (2) regularization of the Cox regression to increase method's robustness against overfitting. The shortcoming of feature aggregation is that it may produce uninterpretable decisions. Among the regularization methods, L1 -penalized Cox regression is the most attractive because it produces concise interpretable rules.

Here, we present an alternative approach to the prediction.

Methods

Let us define early failure as failure before a certain time T. The binary class function C(x) splitting the observations by the time T of failure gives rough representation of their risk. The proposed algorithm learns expectation of the class function E(C(x)).

Understanding patients' risk as risk of early failure rather than a time-dependent function, is common among medical practitioners. For example, web site of Memorial Sloan-Kettering Cancer Center has interactive ``nomograms'' to estimate of the risk of failure within 5 and 10 years for prostate cancer.

The main scheme of the algorithm can be described as two steps procedure:

(1) Independently for each feature X build a predictor f(X) and calculate its weight w;(2) Calculate a scoring function F(x) = sum(w * f(X)).

For a classification problem, there are two popular algorithms which follow this scheme. One of them is Naive Bayes classifier [5] where all weights w = 1 and each predictor is built as a log-ratio between densities of two classes.

Another example of an algorithm with the same scheme is "weighted voting'' [6]. There are several ways the general scheme can be implemented in the

context of survival analysis. The exact description of the algorithm is presented in Fig 1.

Smooth Rank is compared with two methods within traditional approach on ten real life medical survival analysis datasets.

Cox PH regression is used a baseline method, because it is virtually the only method applied in most of medical publications about risk modeling. We use the methods implementation from the R package survival.L1-penalized Cox Path regression [7] builds models for several values of the method's parameter lambda. The function implemented in the R package glmpath by the method's authors is used here. For each model, the function outputs values of three criteria: AIC, BIC, loglik. The criterion AIC was chosen to select the best model for the given training set. We used default values of the parameters of the coxpath procedure. The method does not work with missing values.For Smooth Rank, the parameters of the kernel approximation and LOESS were selected to ensure maximal smoothness. Thus, unlike most of other advanced risk modeling methods, Smooth Rank does not have parameters to tune up. Kernel approximations is done on equally spaced K points, with default K = 512 using cosine kernel.

Results

The next datasets were selected for methods comparison.

1.BMT: The dataset represents data on 137 bone marrow transplant patients [8] . The data allow to model several outcomes. Here, the models are built for disease free survival time. The first feature is diagnosis, which has three values: ALL; AML Low Risk; AML High Risk. Other features characterize demographics of the patient and donor, hospital, time of waiting for transplant, and some characteristics of the treatment. There are 11 features overall, among them two are nominal.

2.Colon: These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There is possibility to model two outcome: recurrence and death. The data can be found in R package survival. The features include treatment (with three options: Observation, Levamisole, Levamisole+5-FU); properties of the tumor, number of lymph nodes. There are total 11 features and 929 observations.

3.Lung1: Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group [9]. Performance scores rate how well the patient can perform usual daily activities. Other features characterize calories intake and weight loss. The dataset has 228 records with 7 features

4.Lung2: The dataset from [10] Along with the patients' performance scores, the features include cell type (squamous, small cell, adeno, and large), type of treatment and prior treatment.

5.Breast cancer dataset [11]. It contains 7 tumor characteristics in 97 records of patients.

6.PBC: This data is from the Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984 [12]. Patients are characterized by standard description of the disease conditions. The dataset has 17 features and 228 observations.

7.Al: The data [13] of the 40 patients with diffuse largeB-cell lymphoma contain information about 148 gene expressions associated with cell proliferation from lympohichip microarray data. Since there are more features than the observations, the Cox regression could not be applied on the data.

8.The dataset from [14] contains information about 240 patients with lymphoma. Using hierarchical cluster analysis on whole dataset and expert knowledge about factors associated with disease progression, the authors identified relevant four clusters and a single gene out of the 7399 genes on the lymphochip. Along with gene expressions, the data include two features for histological grouping of the patients. The authors aggregated gene expressions in each selected cluster to create a signatures of the clusters. The signatures, rather than gene expressions themselves were used for modeling. The dataset with aggregated data has 7 features.

9,10 Ro03g, Ro03s: the data [15] of 92 lymphoma patients. The input variables include data from lymphochip as well as results of some other tests. The Ro03s data contain averaged values of the gene expressions related with cell proliferation (proliferation signature). The Ro03g dataset includes the values of the gene expressions included in the proliferation cluster, instead of their average. Thus, the Ro03s dataset contains 6 features, and the dataset Ro03g contains 26 features.

The results are presented in the Fig 2. The ratio N/M is included as a measure of the dataset "sparsity": the smaller is the ratio, the less representative (more sparse) is the dataset. For all datasets, except Colon, the the table contains average CI on the test data over 50 random splits on training and test in proportion 2 to 1. For the largest data, Colon 30 splits were conducted.

Discussion

The results allow one to make the next observations:
(1) In 8 out of 10 cases the Smooth Rank has the best results. The two other cases are the least sparse, they have the highest N/M ratio. It is interesting that Cox regression turned out to be the best method in these cases.
(2) In the three cases with the lowest ratio of N/M (lines 1,7,10) the advantage of Smooth Rank is the most prominent. Its performance is higher than performance of other methods by 10% - 12%.
(3) Datasets R003s, Ro03g contain information on the same patients. Dataset Ro03g contains original values of gene expression, and Ro03s includes aggregated features, ``signatures''. Smooth Rank has better (the best) results without aggregation, while other methods require preliminary feature aggregation for comparable performance.
Overall, the results demonstrate large advantage of the Smooth Rank on the smaller datasets. On the first glance, the success of the new method seems counter-intuitive: it does not take into account large part of the available information (early censored observations do not participate in training, the outcome is reduced to two classes, information about multivariate relationship between features and the outcome is ignored), while the competing methods use all this information for modeling. But let us recall that Naive Bayes classifier, which has the same general scheme as Smooth Rank, often outperforms more contemporary and sophisticated classification algorithms on smaller datasets. Smooth Rank, as an algorithm for survival analysis, inherits this important property of the Naive Bayes.

Conclusion(s)

We presented a new survival prediction algorithm Smooth Rank. The method uses smoothing techniques and aggregation of univariate predictors to reduce variance associated with small samples. This makes the method more robust, and leads to significant performance improvements on sparse datasets comparing with the popular methods within the traditional approach. As we demonstrated in special experiments, Smooth rank requires much lest data to build a model with comparable quality. The method can work with missing data, it does not require preliminary data aggregation or turning parameters. Smooth Rank produces models in linear form, where each predictor corresponds to a single input feature and weights are related with performance of each predictor. This makes the model convenient for interpretation and can help in medical understanding of the prognosis.

References

1. Cox D R. Regression Models and Life-Tables. Journal of Royal Statistical Society. Series B (Methological)1972; 34: 187220.
2. Harrell F E Jr, Lee K L, Mark D B. Tutorial in Biostatistics. Multivariate Prognostic Models. Statistics in Medicine1996; 15: 361387
3. Segal M R. Microarray gene expression data with linked survival phenotypes. Biostatistics 2006; 7: 2, 268285.
4. Wieringen W, Kun D, Hampel R, Boulesteix A-L. Survival prediction using gene expression data: a review and comparison. Computational Statisticsand Data Analysis 2009; 53: 15901603.
5. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning 2001; Springer, NewYork.
6. Golub TR et al. Molecular Classification o fCancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science1999; 286: 531537.
7. Park M Y, Hasie T. L1-regularization path algorithm for generalized linear models .J. R. Statist. Soc. B. 2007; 69: 659-677.
8. Klein J P , Moeschberger M L. Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition. 2003; Springer, NewYork.
9. Loprinzi C L. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group.Journal of Clinical Oncology 1994; 12, 601607.
10. Kalbfleisch J,Prentice R. TheStatistical Analysis of Failure Time Data 2002; J. Wiley, Hoboken N .J.
11. vant Veer L J et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.
12. Therneau T, Grambsch P.Modeling Survival Data: Extending the Cox Model. 2000; Springer-Verlag, NewYork.
13. Alizadeh A et al. Distinct types of diffuse large-b-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503511.
14. Rosenwald A et al. The use of molecular profiling to predict survival after chemotherapy for defuse large -b-cell lymphoma. New England journal of Medicine 2002; 346: 19371947.
15. Rosenwald A et al. The proliferation gene expression signature is a quantitative predictor of oncogenic events that predict survival in mantle cell

lymphoma3; Cancer Cell 2003; 185197.1. Cox D R. Regression Models and Life-Tables. Journal of Royal Statistical Society. Series B (Methological)1972; 34: 187220.
16. Harrell F E Jr, Lee K L, Mark D B. Tutorial in Biostatistics. Multivariate Prognostic Models. Statistics in Medicine1996; 15: 361387
17. Segal M R. Microarray gene expression data with linked survival phenotypes. Biostatistics 2006; 7: 2, 268285.
18. Wieringen W, Kun D, Hampel R, Boulesteix A-L. Survival prediction using gene expression data: a review and comparison. Computational Statisticsand Data Analysis 2009; 53: 15901603.
19.Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning 2001; Springer, NewYork.
20. Golub TR et al. Molecular Classification o fCancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science1999; 286: 531537.
21. Park M Y, Hasie T. L1-regularization path algorithm for generalized linear models .J. R. Statist. Soc. B. 2007; 69: 659-677.
22. Klein J P , Moeschberger M L. Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition. 2003; Springer, NewYork.
23. Loprinzi C L. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group.Journal of Clinical Oncology 1994; 12, 601607.
24. Kalbfleisch J,Prentice R. TheStatistical Analysis of Failure Time Data 2002; J. Wiley, Hoboken N .J.
25. vant Veer L J et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.
26. Therneau T, Grambsch P.Modeling Survival Data: Extending the Cox Model. 2000; Springer-Verlag, NewYork.
27. Alizadeh A et al. Distinct types of diffuse large-b-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503511.
28. Rosenwald A et al. The use of molecular profiling to predict survival after chemotherapy for defuse large -b-cell lymphoma. New England journal of Medicine 2002; 346: 19371947.
29. Rosenwald A et al. The proliferation gene expression signature is a quantitative predictor of oncogenic events that predict survival in mantle cell lymphoma3; Cancer Cell 2003; 185197.

Illustrations

Illustration 1

The Smooth Rank Algorithm

Algorithm Smooth Rank

1. Split observations on two classes $C \in \{-1, 1\}$ by median survival time T;
2. For each feature x^i:
   (a) Build kernel approximations g^i_{-1}, g^i_1 of the density of each class on the set $R^i \in dom(x^i)$;
   (b) For each point $r \in R^i$ calculate

$$q_i(r) = \frac{g^i_1(r) - g^i_{-1}(r)}{g^i_1(r) + g^i_{-1}(r)}$$

   (c) Build LOESS approximation $f_i(x)$ of the function $q_i(x)$
   (d) Calculate $w_i = CI(f_i(x^i)) - 0.5$
3. Calculate scoring function

$$F(x) = \sum_{i:\, x^i \neq NA} w_i \cdot f_i(x^i) / \sum_{i:\, x^i \neq NA} w_i$$

Illustration 2

The Method's Comparison

#	Data	N/M	Cox	Cox Path	Smooth Rank
1	BMT	12.4	0.58 ± 1.02E-02	0.57 ± 9.70E-03	**0.69** ± 7.73E-03
2	Colon	84.4	**0.66** ± 2.88E-03	0.65 ± 3.50E-03	0.65 ± 4.53E-03
3	Lung1	32.6	0.61 ± 6.32E-03	0.61 ± 7.30E-03	**0.63** ± 5.70E-03
4	Lung2	22.8	0.69 ± 5.92E-03	0.70 ± 6.07E-03	**0.72** ± 5.32E-03
5	BCW	13.9	0.70 ± 8.2E-03	0.71 ± 7.3E-03	**0.72** ± 5.32E-03
6	PBC	24.6	0.81 ± 4.56E-03	**0.82** ± 4.59E-03	**0.82** ± 4.02E-03
7	Al	0.27	—	0.52 ± 2.16E-02	**0.64** ± 1.01E-02
8	Ro02s	34.3	**0.73** ± 4.00E-03	0.72 ± 4.40E-03	0.69 ± 4.01E-03
9	Ro03s	15.3	0.74 ± 8.67E-03	**0.75** ± 8.48E-03	**0.75** ± 8.04E-03
10	Ro03g	3.54	0.58 ± 2.0E-02	0.67 ± 1.9E-02	**0.77** ± 7.65E-03

Every cell contains mean value of CI and its 95% confidence interval on the test data for all the random splits for each method. The highest value(s) in each row are marked by bold font.

Disclaimer