ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 6 (2011) 267–272

Procedia
Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

A new tool for survival analysis: evolutionary programming/evolutionary strategies (EP/ES) support vector regression hybrid using both censored / non-censored (event) data

Walker H. Land, Jr. [a], Xingye Qiao [b], Dan Margolis [a], Ron Gottlieb [c],

[a]*Department of Bioengineering, Binghamton University, Binghamton, NY 13902-6000, USA.*
[b]*Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA.*
[c]*Department of Radiology, University of Arizona, Tuscan, AZ 85724, USA.*

Abstract

While the role of survival analysis in medicine has continued to be increasingly essential in making treatment and other health care decisions, the common clinical methods used for performing these analyses, such as Cox Proportional Hazard models and Kaplan-Meier curves, have become antiquated. We have developed a new survival analysis technique of the Evolutionary Programming / Evolutionary Strategies Support Vector Regression Hybrid for censored and non-censored event data. This method provides the benefits of optimized statistical learning theory to be used as a replacement for or in addition to existing survival analysis protocols. The technique was tested on an artificially censored data from a well-known benchmark dataset as well as actual clinical data with encouraging results.

Keywords: SVRc; Evolutionary Programming; Statistical Learning Theory; Survival Analysis.

1. Introduction

1.1 Survival Analysis

Survival analysis is a well-known technology in the medical field related to analyzing and predicting the time until the occurrence of an event, such as death, progression of cancer, or need for a replacement joint. This technology has been applied not only in clinical research, but also in epidemiology, reliability engineering, marketing, insurance, and many other fields. In medicine, it plays a pivotal role in determining courses of treatment, developing new pharmaceuticals, preventative medicine, and improving hospital procedures. Performing meaningful survival analysis has greatly improved medicine and the health care system. However, existing methods provide only some of the potential benefit that more advanced forms of survival analysis could offer.

1877–0509 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.procs.2011.08.050

1.2 Censoring

The amount of time between a specified starting point and the occurrence of the event is specified as the survival time (or failure time). This can be measured in actual clock time or other appropriate time measures determined by the nature of the event. Thus, analysis of survival time with methods like linear regression requires all data to have a start time and a stop (event) time. Unfortunately, due to the nature of medical survival analysis, the stop time is often not the actual time the event occurred. For example, a key use is clinical trials, which often have a limited time span for data collection. This means any case in which an event would have occurred but had not by the end of the study would have to be thrown away. Furthermore, unrelated events, such as a car accident or gunshot wound may cause an event to occur but it would not be related to the original medical question for which survival analysis is being performed. Such data is referred to as censored data in survival analysis. Censored data is any data in which partial time information is known but not complete information. The data contains a stop time which is not the actual event time but the minimum possible event time. The large prevalence of such data in medicine has led to most accepted survival analysis techniques to account for censoring in some way.

1.3 Kaplan-Meier Curves

Kaplan-Meier curves are considered one of the gold standards for visualizing and comparing survival with censored data (Kaplan, 1958). It is based on a simple survival curve which divides the number of “survivors” (data for which the event has not occurred) from the total number of cases in the dataset, and plots this percentage on the y-axis against various time points on the x-axis. In the presence of censored data, the calculation of the percentage of survivors is done by first removing the number of censored cases at a time point from both the number of survivors and the total number of cases. Thus, as time increases on the x-axis, more and more censored data disappears from the calculations. Each removal of a case is represented on the curve with a small symbol (usually a “+”), providing a survival curve that takes into account censored data. In order to provide useful decision making information, cases are split into groups, such as control vs. treatment or male vs. female, and the curves generated for those different groups can be compared with relative ease visually or with a Chi-squared significance test. However, Kaplan-Meier curves cannot provide information about survival for a particular individual, only for a group as a whole. Also, the grouping is done manually by the person performing the analysis, and therefore the groups are very unlikely to represent relatively simple, linear differences between the cases rather than any nonlinear relationships.

1.4 Cox Proportional Hazard Models

A Cox Proportional Hazard Model (Cox PH) is the other method considered a gold standard for survival analysis with censored data in medicine (Cox, 1972). The Cox PH model is a semi-parametric linear regression model which looks at the “hazard” or risk of an event. It assumes that the hazard of an observation is proportional to an unknown “baseline” hazard common to all observations, where this proportionality is modeled as an exponential of a linear function of the covariates. From this model, a single value is created called a Cox Hazard Ratio (CHR), which supposedly represents the hazard or risk of the event occurring over time between groups. As with the Kaplan-Meier curves, the groupings are done manually by the person performing the analysis and thus neglect any non-linear relationships in the data. Also, the CHR is for the group as a whole, and once again cannot provide individualized information to a particular patient or case.

1.5 New Tool for Survival Analysis

Both Kaplan-Meier curves and Cox PH models are useful techniques that will always hold a place in survival analysis, but they lack core functionality that more advanced survival analysis techniques should be able to provide. Most significantly, there exists a need for individualized survival predictions based on a patient’s own features rather than a group survival curve or ratio based on belonging to a generalized group. The other missing functionality that a new tool will need to address is the handling of complex data of multiple types from multiple sources and/or observers, which is likely to have many linear and non-linear interactions that could be exploited for survival analysis.

2. Methods

2.1 Support Vector Regression with censoring (SVRc)

Support Vector Machine (SVM) is a widely used large-margin classification method. SVM uses a hinge loss function. In the linear case, the resulting separating hyperplane minimizes the empirical hinge loss of the training data set with a penalty term which controls the complexity of the model. The SVM optimization problem can be viewed as maximizing a notion of the margin between the two classes, while keeping the misclassification of the training data to be as small as possible.

An appealing feature of SVM, as well as other large margin classifiers, is that a so-called kernel trick can be applied, where each data vectors x_i is transformed to a functional, denoted as $\phi(\cdot,\ x_i)$, which is essentially an infinite-dimensional *vector*. The kernelized SVM then work on the functionals. The transformation is not easy to define. However, linear SVM does not work with data vector x_i's directly, but the inner product of the data vectors $x_i'x_j$. Similarly, kernelized SVM work on the inner product of two functionals in the functional space, which turns out to be $k(x_i,x_j)$, where $k(\cdot,\cdot)$ is a Mercer kernel.

SVM deals with problem where the response variables are class labels. When the response variables are (scalar) continuous, there is a nice extension of SVM, called Support Vector Regression (SVR). In contrast to the use of sum of squared error loss as in Ordinary Least Square (OLS) / Linear Regression models, SVR uses the ε-insensitive loss function,

$$|\xi_\varepsilon| := \begin{cases} 0 & \text{if } |\xi| \le \varepsilon, \\ |\xi| - \varepsilon & \text{otherwise.} \end{cases}$$

The idea behind this ε-insensitive loss function is that if the error $\xi = \hat{y} - y$ has absolute value less than the pre-specified ε, then there is no harm and no loss is imposed in the objective function. Only when $/\xi/$ becomes greater than ε, the loss starts growing linearly to $/\xi/$ and a cost constant C.

SVR is a very powerful and efficient machine in regression setting, inheriting the merits of SVM, especially when the number of the predictors is much larger than the sample size, where the burden of the computation results from the increase of the sample size rather than the number of variables. The readers can refer to Smola and Scholkopf (2004) for reference.

Despite the merits of SVR in the regression context, it is not particularly useful in the censored data setting. In the censored data setting, the true event time y^* is observable only when it is not censored, *a.k.a.*, it is event time. When the event is censored, the observed y is merely the censoring time which is earlier than the true event time, $y<y^*$. Thus, if SVR is to predict the true event time y^* by $\hat{y}$, it is very likely to have the error $\xi = \hat{y}-y>0$ for the censored observations. We call them the right errors. Similarly, we call $\xi = \hat{y}-y$ the left errors if $\xi <0$. It is perceivable that an ideal regression algorithm should tolerate more right errors than the left errors for the censored observations. This idea leads to an extension of SVR for the censored data: larger ε and smaller C for the right error in the censored data (hereinafter denoted as ε_i^* and C_i^* and will be discussed below). This is exactly the idea that was pursued in Khan and Zubek (2008), which leads to Support Vector Regression for Censored Data (SVRc). We briefly summarize the formulation and the solutions to SVRc below.

The goal of SVRc is to find out the coefficient vector *w* and the bias term *b* in the prediction function $\hat{y} = f(x) = x'\omega + b$. They are found by the primal optimization problem,

$$\min_{\{\omega,b,\xi_i,\xi_i^*\}} L_P := \frac{1}{2} \parallel \omega \parallel^2 + \textstyle\sum_i (C_i \xi_i + C_i^* \xi_i^*)$$

subject to,

$$y_i - x_i'\omega - b \le \varepsilon_i + \xi_i,$$
$$x_i'\omega + b - y \le \varepsilon_i^* + \xi_i^*,$$
$$\xi_i, \xi_i^* \ge 0.$$

The values of ε_i, ε_i^*, C_i, C_i^* depend on the censoring status and whether they are corresponding to left or right errors. For example, ε_i^* and C_i^* are the level of tolerance and the cost for the right errors respectively. We should have a larger ε_i^* and smaller C_i^* in the censored cases than otherwise. In the empirical section, we use the following specifications of these parameters, which are found and validated by the EP/ES process discussed in Section 2.2

C	$i\in$Non-censored	$i\in$Censored
Left - C_i	4.79	.0967
Right - C_i^*	.121	4.78

ε	$i\in$Non-censored	$i\in$Censored
Left - ε_i	0.227562	67.0792
Right - ε_i^*	0.428331	2.23837

In order to find out the Lagrangian of this problem, we introduce Lagrange multipliers $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$.

$$L = \frac{1}{2} \parallel \omega \parallel^2 + \sum_i (C_i\xi_i + C_i^*\xi_i^*) - \sum_i \alpha_i(\varepsilon_i + \xi_i - y_i + x_i'\omega + b)$$

$$-\sum_i \alpha_i^*(\varepsilon_i^* + \xi_i^* + y_i - x_i'\omega - b) \; - \sum_i \eta_i\xi_i - \sum_i \eta_i^*\xi_i^*$$

The partial derivatives of the Lagrangian L with respect to the primal variables ($w, \; b, \; \xi_i, \; \xi_i^*$) have to vanish for optimality (KKT conditions).

- $\frac{\partial L}{\partial \omega} = \omega - \sum_i (\alpha_i - \alpha_i^*)x_i = 0.$
- $\frac{\partial L}{\partial b} = \sum_i (\alpha_i - \alpha_i^*) = 0.$
- $\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0.$
- $\frac{\partial L}{\partial \xi_i^*} = C^* - \alpha_i^* - \eta_i^* = 0.$

Substituting these conditions into the Lagrangian yields the dual optimization problem.

$$\max_{\{\alpha_i, \alpha_i^*\}} -\frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)x_i'x_j - \sum_i (\alpha_i\varepsilon_i + \alpha_i^*\varepsilon_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

subject to,

$$\sum_i (\alpha_i - \alpha_i^*) = 0,$$

$$0 \leq \alpha_i \leq C_i,$$

$$0 \leq \alpha_i^* \leq C_i^*.$$

The dual problem of SVRc above can be translated to a standard Quadratic Programming. We use the solver in LibSVM to calculate the solutions to SVRc. Once α_i, α_i^* are found, the coefficient vector w can be written as,

$$\omega = \sum_i (\alpha_i - \alpha_i^*)x_i$$

We use the following rules to find out the bias term b:

1) If there exists i so that $0 < \alpha_i < C_i$, then use:

$$\mathrm{b} = -\varepsilon_i + y_i - x_i'\omega$$

2) If there exists i so that $0 < \alpha_i^* < C_i^*$, then use:

$$\mathrm{b} = \varepsilon_i^* + y_i - x_i'\omega$$

Note that these two cases do not happen at the same time.

Finally, the predicted value for x is:

$$\hat{y} = f(x) = x'\omega + b.$$

As an extension to SVR, which further is a sibling to SVM, SVRc is also subject to kernel trick. The next section will discuss various issues about the hybrid of SVRc and the kernels built by evolutionary programming/evolutionary strategies (EP/ES). In general, one only needs to replace $x_i'x_j$ by $k(x_i,x_j)$ in the dual problem of SVRc to apply the kernel trick. The prediction function then becomes

$$\hat{y} = f(x) = \sum_i (\alpha_i - \alpha_i^*)k(x_i, x) + b,$$

where b is found by $-\varepsilon_i + f(x_i)$ or $\varepsilon_i^* + f(x_i)$ in the similar way as before.

2.2 Evolutionary Programming / Evolutionary Strategies (EP/ES)

Determining the best parameters for the SVRc is a very difficult and complex problem. The nature and complexity of the problem requires a powerful stochastic method that can explore the large problem space of potential solutions to find a global minimum without becoming stuck in local minima. Significant previous experience with the problem of determining the best parameters for a Support Vector Machine (SVM) led to the creation of an Evolutionary Programming / Evolutionary Strategies SVM Hybrid (EP-SVM) that was successfully utilized many times on different types of data. The same stochastic process has been modified for use with the SVRc. The chromosome includes kernel type, kernel parameters (gamma, coeff, and degree), cost function, and epsilons, though many of these can be locked by the user if desired. An initial population of SVRc chromosomes are created, mutated, and then selected using tournament selection with or without replacement for a number of generations. In medical classification problems, such as those solved by the EP-SVM, Receiver Operating Characteristic curves are used to measure performance. However, SVRc is a regression technique, and thus mean squared error or the concordance index is used to measure fitness of a chromosome. The EP/ES Hybrid method allows the user to find an optimal set of SVRc parameters quickly but without premature convergence. This combined technique will be referred to as the EP-SVRc.

3. Results

Two datasets were used to perform a preliminary test of the EP-SVRc. The first was the Auto-MPG dataset taken from Carnegie Mellon University's StabLib library, which is a commonly used regression analysis dataset (Zhou *et al.*, 2000; Birattari *et al.*, 1998, Greig *et al.*, 1997). The data contained 392 samples of vehicle related features and corresponding actual miles per gallon. The dataset was artificially censored by reducing the miles per gallon value randomly for ~10% of the data, which was also chosen at random. Figure 1 shows the prediction values (blue) in order and the corresponding actual values (red) overlaid. The mean squared error for this set of predictions was 6.68. The second dataset was the Mayo Clinic's Primary Biliary Cirrhosis (PBC) survival dataset, which was a real dataset with survival times and censored data in the form of either a patient still alive or lost contact at the end of the study, or a patient who received a liver transplant. This dataset had 276 samples of which 165 were censored (Murtaugh, 1994 and Fleming and Harrington 1991). Figure 2 shows the prediction values (blue) in order and the corresponding actual values (red) overlaid. The outcomes were scaled by taking the log and then normalizing to a 0 to 1 scale. This dataset had a mean squared error of .023.

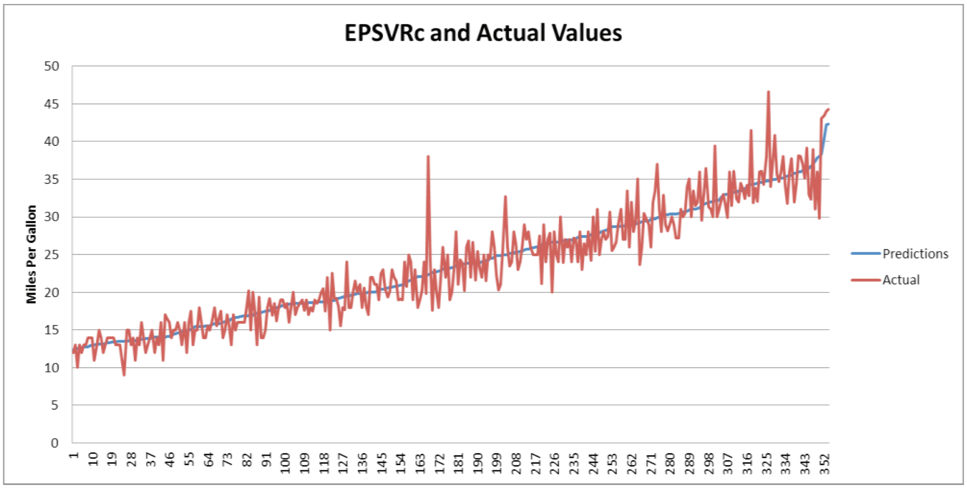

Figure 1: Plot of EP-SVRc prediction values (blue) with corresponding actual values overlaid (red) for the Auto-MPG dataset. Artificial censoring was performed and the points were sorted in order of prediction value, and show a good fit to the actual data.

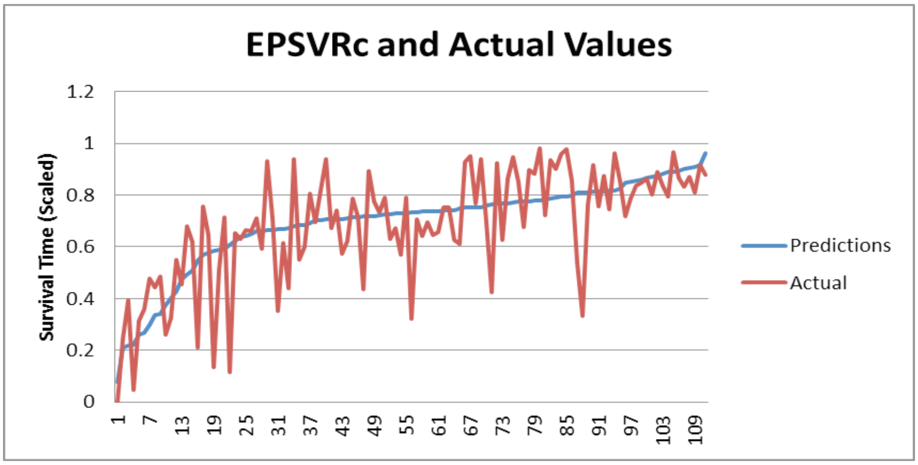

Figure 2: Plot of EP-SVRc prediction values (blue) with corresponding actual values overlaid (red) for the Mayo Clinic's Primary Biliary Cirrhosis survival dataset. While a larger dataset in number of features and number of samples is needed, the EP-SVRc was able to begin to fit to the actual data.

4. Discussion

The EP-SVRc performed quite well on the Auto MPG dataset, which was a larger and less noisy dataset than the PBC dataset. It proved that the EP-SVRc is very capable of handling traditional regression tasks with impartial (censored) information. However, the Auto MPG dataset does not represent the type of problem that the EP-SVRc has been developed for, which the PBC dataset does. The PBC dataset represented a classic survival analysis problem, with widely varying survival times and lots of censored information. These preliminary results are encouraging as they show the EP-SVRc managing to gain some information from this very noisy and small dataset which has not gone through any serious preprocessing or feature selection.

5. Preliminary Conclusions

The EP-SVRc is a promising new application of stochastic technology, but additional research and testing are required. Preliminary tests show the fundamental theory and technology works as expected and that it has the potential to perform effective survival analysis. However, a larger dataset including modern features such as epigenetics, gene microarrays, and CT/PET imaging based information are expected greatly improve system performance. Further theoretical work is being implemented that includes more complex chromosome kernels as well as feature selection as an integral part of the EP-SVRc.

6. References

Birattari, B., Bontempi, G., Bersini, H. (1998) “Lazy Learning Meets the Recursive Least Squares Algorithm”. NIPS.

Chang, C. and Lin, C. (2011). “LIBSVM: a library for support vector machines”. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cox, D. (1972). “Regression Models and Life-Tables”. *Journal of the Royal Statistical Society. Series B* 34 (2):187–220.

Fleming T and Harrington D. (1991), “Counting Processes and Survival Analysis”, Wiley, New York.

Greig, D., Siegelmann, T., Zibulevsky, M. (1997). “A New Class of Sigmoid Activation Functions That Don't Saturate”.

Kaplan, E. and Meier, P. (1958). “Nonparametric estimation from incomplete observations”. *J. Amer. Stat. Assoc.* 53:457–481.

Khan, F. and Zubek, V. (2008). “Support vector regression for censored data (SVRc): a novel tool for survival analysis.” *Eighth IEEE International Conference on Data Mining*, 863—868.

Mann, N. R., *et al.* (1975). Methods for Statistical Analysis of Reliability and Life Data. John Wiley & Sons.

Murtaugh, P., Dickson, E., Van Dam, *et al.* (1994). “Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits”. *Hepatology*. 20(1.1):126-34.

Smola, A. and Scholkopf, B. (2004). “A tutorial on support vector regression”. *Statistics and Computing*. 14, 3, 199-222.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Zhou, Z., Chen, S., Chen, Z. (2000). “A Statistics Based Approach for Extracting Priority Rules from Trained Neural Networks”. IJCNN (3).