Medical Survival Analysis
Through Transduction of Semi-Supervised Regression Targets

Faisal M. Khan, Rutgers University, USA
Qiuhua Liu, Schlumberger Limited, USA

ABSTRACT

A crucial challenge in predictive modeling for survival analysis applications such as medical prognosis is the accounting of censored observations in the data. While these time-to-event predictions inherently represent a regression problem, traditional regression approaches are challenged by the censored characteristics of the data. In such problems the true target times of a majority of instances are unknown; what is known is a censored target representing some indeterminate time before the true target time. While censored samples can be considered as semi-supervised targets, the current limited efforts in semi-supervised regression do not take into account the partial nature of unsupervised information; samples are treated as either fully labeled or unlabelled. This paper presents a novel semi-supervised learning approach where the true target times are approximated from the censored times through transduction. The method can be employed to transform traditional regression methods for survival analysis, or can be employed to enhance existing state-of-the-art survival analysis methods for improved predictive performance. The proposed approach represents one of the first applications of semi-supervised regression to survival analysis and yields a significant improvement in performance over the state-of-the-art in prostate and breast cancer prognosis applications.

Keywords: Cancer Prognosis, Regression, Semi-Supervised, Support Vector, Survival Analysis, Transduction

INTRODUCTION

The employment of predictive time-to-event modeling in medical survival analysis usually falls into two broad categories. The first is prognostic, developing models for how a certain disease will progress. The purpose of such models includes understanding disease processes and prediction of how new patients will behave in the context of existing data. Examples include predicting which prostate cancer patients will recur so that therapy can be initiated early (Donovan et al., 2009) or identifying which group of patients will benefit more from a certain therapy. The second purpose is factor analysis; to analyze disease processes and explore interaction affects between disease factors. An example is understanding the inter-
action of whether a potentially significant gene will continue to be relevant when combined with other factors in a multivariate setting (Donovan et al., 2009) in order to possibly prioritize and identify candidate genes for targeted therapeutic drug development.

While time-to-event prediction is inherently a regression problem, it challenges computational modeling approaches due to the fact that healthcare data in such settings is characterized by censored and non-censored (event) observations. Healthcare data used in such prognostic modeling is usually obtained from tracking patients over the course of a well designed study, perhaps lasting years. Contrary to traditional regression problems, the information for most observations is incomplete and only known “up-to-a-point.” Patients who have experienced the endpoint of interest (cancer remission, recurrence, etc.) during their follow-up are considered as non-censored or events. Patients that did not experience the endpoint during study or were lost to follow-up for any cause (i.e., the patient moved during a multi-year study) are considered censored. All that is known about them is that they were disease free up to a certain point, but what subsequently occurred is unknown. For a $d$-dimensional vector $x_i \in \mathbb{R}^d$ the observed time $S_i$ is called the censoring time. For such individuals, it is only known that they survived for at least time $S_i$. The actual target $T_i$ is unknown for censored cases, thus $S_i < T_i$. An important assumption is that $T_i$ and $S_i$ are independent conditional on $x_i$, i.e., the cause for censoring is independent of the survival time. With an indicator function $\delta_i$ which is 0 if an event occurred and 1 if the observation is censored, the available training data can be summarized for $N$ patients as $D = \{ T_i, x_i, \delta_i \}_{i=1}^N$ (Raykar et al., 2008).

Censored observations contribute incomplete information as the event of interest may occur after they were lost to follow-up. Simply omitting the censored observations (Burke et al., 1997; Shivaswamy, Chu, & Janasche, 2007) or treating them as non-recurring samples in a classifier (Snow, Smith, & Catalona, 1997) both bias the resulting model and should be avoided.

Additionally, in the field of healthcare diagnostics, due to the costs involved in identifying acceptable patients who will provide consent for inclusion in research, and then actively tracking them over a significant period of time, the sample size is often small, in the tens or hundreds. Since most of samples may be censored, e.g., 91% in prostate cancer (Donovan et al., 2008), 76% in breast cancer (Mangasarian, Street, & Wolberg, 1994) dropping such patients is a further unattractive option and accounting for them is of crucial importance for a model. Such samples (event-free and lost to follow-up) are considered right-censored; their information on the right-hand side of a timeline is unknown. The problem is further confounded by the fact that non-censored patients may experience the event-of-interest prior to their recorded time ($S_i > T_i$); for example a cancer patient may visit a doctor every six months, so if recurrence is observed, it happened somewhere in the six months between his last visit and the visit when the disease was detected. The term left-censoring describes this phenomenon where even the status of event patients is not completely known. The incomplete nature of the outcome targets in time-to-event prediction thus challenges traditional regression techniques and precludes their use. Instead, methods which can correctly account for censored observations are crucial for analyzing time-to-event problems.

A heretofore unrealized notion of great interest is the fact that the censored samples prevalent in time-to-event problems can be considered as semi-supervised targets. While there has been significant work in semi-supervised classification approaches, particularly for SVMs (Bennet & Demiriz, 1999; Chapelle, Sindhwami, & Keerthi, 2008; Chen, Wang, & Dong, 2003; Fung & Mangasarian, 2001, Gamerman, Vovk, & Vapnik, 1998; Kemp, Griffiths, Stromsten, & Tenenbaum, 2004; Seeger, 2006), there has been limited work in semi-supervised regression, especially SVR (Belkin, Niyogi, & Sindhwami, 2006; Cortes & Mohri, 2007; Rwebangira & Lafferty, 2008; Szummer & Jaakkola 2000; Zhou & Li, 2005). The limited work thus far in semi-supervised
12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/article/medical-survival-analysis-through-transduction/63617?camid=4v1

This title is available in InfoSci-Journals, InfoSci-Journal Disciplines Medicine, Healthcare, and Life Science.
Recommend this product to your librarian:
www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Data Mining and Meta-Analysis on DNA Microarray Data
www.igi-global.com/chapter/data-mining-meta-analysis-dna/76115?camid=4v1a

Efficient Mining Frequent Closed Discriminative Biclusters by Sample-Growth: The FDCluster Approach
www.igi-global.com/article/efficient-mining-frequent-closed-discriminative/49550?camid=4v1a

Unsupervised Feature Selection
www.igi-global.com/chapter/unsupervised-feature-selection/53905?camid=4v1a
Incorporating Network Topology Improves Prediction of Protein Interaction Networks from Transcriptomic Data


[www.igi-global.com/article/incorporating-network-topology-improves-prediction/47093?camid=4v1a](www.igi-global.com/article/incorporating-network-topology-improves-prediction/47093?camid=4v1a)