Survival Analysis Meets Data Stream Mining

Mark Last and Hezi Halpert

Department of Information Systems Engineering
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel
{mlast, halpertc}@bgu.ac.il

Abstract. Survival analysis deals with monitoring entities over their lifetime. The definition of "birth" and "death" events depends on the nature of a given entity. When we observe an infinite stream of birth and death events, at each point in time some of the monitored entities are "right-censored", i.e. we know the time elapsed since their birth event, but their death event has not occurred yet and we do not know when it will occur in the future. Often, the snapshots of partially censored observations keep arriving over time in the form of a data stream. Given each snapshot, we may be interested to predict the timing of death events for all live entities or, alternatively, to predict their label ("survived" or "failed") as a function of time. In this research, our intention is to modify standard classification algorithms, such as decision trees, so that they can seamlessly handle a snapshot stream of both censored and non-censored data. The objective is to provide reasonably accurate predictions after observing relatively few snapshots of the data stream and to improve the classification model with additional information obtained from each new snapshot.

Keywords: Data stream mining; survival analysis; censored data; classification; probability estimation.

1 Extended Abstract

1.1 Motivation

Survival analysis [2] deals with monitoring entities over their lifetime. The definition of "birth" and "death" events depends on the nature of a given entity. In the health care domain, the "birth event" may represent the biological birth of a patient as well as some medical procedure (e.g., surgery), whereas the "death event" may indicate the actual patient's death or some critical change in the patient condition (e.g., cancer recurrence).

Survival analysis is applicable to many additional domains. Thus, in Customer Relationship Management (CRM), the entities are customers who are "born" when they subscribe to a service and "die" when they churn. The service provider is usually interested to estimate the customer retention ("survival") probability within a predefined period (e.g., one year). In the equipment maintenance domain, an item (e.g., a vehicle) is born when it is manufactured or shipped to a customer and it dies when there is a failure. Car manufacturers are particularly interested in estimating the failure probabilities during the warranty period.

When we observe an infinite stream of birth and death events, at each point in time some of the monitored entities are "right-censored", i.e. we know the time elapsed since their birth event (called "observed lifetime"), but their death event has not occurred yet and we do not know when it will occur in the future. Often, the snapshots of partially censored observations keep arriving over time in the form of a data stream. Given each snapshot, we may be interested to predict the timing of death events for all live entities or, alternatively, to predict their label ("survived" or "dead") as a function of time.

1.2 Proposed Methodology

The traditional approach to estimating the survival probability is to use the entity age as the only predictive feature. The Kaplan-Meier method (also called the product-limit estimator) is used to estimate the survival function, the probability to survive beyond a specific time from incomplete observations. It was first proposed by Böhmer in 1912 and rediscovered by Kaplan and Meier in 1958 [1]. The survival probability is estimated separately for each observed lifetime in the dataset. It is assumed that the probability of surviving each lifetime is statistically independent of every other lifetime. The Kaplan–Meier estimate of the survival curve *S(t)* is then:

$$S(T) \approx \prod_{t_i < T} \frac{n(t_i) - d_i}{n(t_i)} \tag{1}$$

where $n(t_i)$ is the number of objects “at risk” (not lost through censoring or failure) at time t_i and d_i is the number of objects which failed at time t_i. Usually d_i = 1, since t_i is chosen as the observed time to failure.

The purpose of the algorithm proposed here is to provide an accurate probability estimation model for each incoming data snapshot, with special consideration to censored records. The goal is to estimate the probability of an event (such as a disease recurrence in a patient or a technical failure in a car) within a pre-defined “follow-up” period based on multiple predictive features (such as patient demographic data or vehicle usage data). In each snapshot, the model induction process is done in a similar way to the approach of [4] by estimating the probability of the different outcomes of censored records based on the survival curve estimation. However, unlike the study of [4], which is aimed at inducing an accurate classification model from a single snapshot of survival data, here we deal with a stream of snapshots that keep arriving over time, whereas we are interested to estimate the failure probability for records in each new snapshot.

In each snapshot, there is a record for every monitored entity. If the record has a class label (it has a failure or its follow-up period has expired) then this record can be considered as uncensored. Otherwise, this record will be considered as censored and it will be stored along with its age (observed lifetime) value – the time, which passed since its creation ("birth"). The age value is used later to calculate the record weight based on the Kaplan-Meier estimation. The algorithm also compares the current and the previous data snapshots to determine for each record, whether this record became uncensored for the first time in this snapshot or not.

For every censored record, instead of a single outcome, two class labels are considered: survival and failure. Given the age T_x of a censored record *x* and the follow-up period T_f ($T_x \leq T_f$), the weight of the record survival outcome is calculated by Eq. 2.

$$W_S(x) = \frac{P_S(T_f)}{P_S(T_x)} \tag{2}$$

Where $P_S(t)$ is the Kaplan – Meier estimation for survival probability at a time *t* based on Eq. 1. The weight of the failure outcome is: $W_F(x) = 1 - W_S(x)$.

To avoid dealing with record "portions", two record copies are created: one labeled by S and weighted with the W_S value and the other one labeled by F and weighted with the W_F value. Censored and non-censored records can be easily combined into a single dataset by assigning each non-censored record *x* the weight of $W(x) = 1.0$. The probability of each class label C (survival S or failure F) in the resulting dataset is then calculated by Eq. 3.

$$P(C) = \frac{\sum_{x \in C} W(x)}{\sum_x W(x)} \tag{3}$$

Where the values of $W(x)$ are calculated by the Kaplan – Meier estimation for censored records and are equal to 1.0 for non-censored records.

1.3 Preliminary Results

The accuracy evaluation of the algorithm is done in every snapshot by applying the model induced from the previous snapshot to the new uncensored records in the current snapshot. We assume that the focus on those records as a testing set reflects well the real-world scenario where at each point in time we are interested to estimate the failure probability for censored (unlabeled) records only. The model accuracy is evaluated by building an ROC curve and calculating the Area under ROC curve (AUC). If the accuracy is significantly below a certain value defined explicitly by the user then the model is considered inaccurate implying that it should be replaced by a new model based on the new information. Each new model is based on all the data available in a given snapshot, including censored records. This is done by estimating the distribution of the different classes for each censored record according to the survival curve estimation as described in [4].

In warranty data and survival data, we are interested in failure prediction model for a specific pre-defined prediction period, also called "follow-up period" [4]. For example, in warranty data, one of the natural prediction periods would be the warranty period. Usually the period starts with a special 'birth' event which denotes the event when the 'clock starts ticking' towards a failure ("death") event. In warranty data, such an event may be associated with the product delivery to the customer. The follow-up period can also be defined as any usage interval in warranty data (e.g., car mileage between 1,000 and 2,000 miles).

We are currently evaluating our algorithm for mining censored data streams on several real-world datasets from the following domains: vehicle warranty data, patient re-infection data, and customer retention data. According to our preliminary results, the proposed algorithm may increase the predictive accuracy by as much as 20% vs. the baseline approach, which ignores the censored data.

1.4 Future Research

For simplicity, our algorithm refers to situations where only two outcomes are possible for each entity: "success" (survival of the follow-up period) and "failure". This is a typical case in warranty data and in survival analysis. An extension to the general case of censored data where multiple outcomes are possible is straightforward and requires minor adaptations in all steps of the algorithm.

Although the algorithm proposed here produces a model for a classification problem, it may easily be adapted to the case of probability estimation problem. For example, instead of a simple decision tree model, the algorithm may use the probability estimation tree (PET) model [3], which can reflect different probabilities for each outcome. The decision whether to produce a classification model or a probability estimation model in a given domain is subject to the user preferences. In some domains, such as in warranty data of high quality products, there are very few records of actual failures. In such cases, a classification model is useless – it will always predict a "success". A probability estimation model would be an obvious alternative in such cases. In general, any classification model can be adapted to the proposed algorithm as long as it can process weighted data instances. The induced models can be evaluated using various measures such as classification accuracy, classification sensitivity, AUC (as in the case of the warranty data example), etc.

2 References

[1] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457-481.

[2] Miller Jr, R. G. (2011). Survival analysis (Vol. 66). John Wiley & Sons.

[3] Provost, F., & Domingos, P. (2003). Tree Induction for Probability-Based Ranking. Machine Learning, 52 (3), 199-215.

[4] Zupan, B., Demsar, J., Kattan, M. W., Beck, R., & Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial Intelligence in Medicine, 20(1), 59–75.