Computational Intelligence in Survival Analysis

Malgorzata Kretowska
Bialystok University of Technology, Poland

INTRODUCTION

Survival analysis, often called time-to-event analysis, is a set of methods focused on failure time prediction. Such methods are developed in a variety of research domains like medicine, sociology, economics or engineering, and are used to analyze different types of failures. For example, in business applications the failure may denote bankruptcy or loan default.

One of the most important characteristics of survival data is censoring. Censored observations contain incomplete information of failure occurrence. It means that for a number of companies we do not observe the event of interest, we only know that the failure did not occur before some specified time. Since the percentage of censored observation is usually high, they should be taken into account in the prediction process.

Considering statistical methods for analysis of survival data, the most common approach is Cox's proportional hazards model (Cox, 1972). This semi-parametric model is usually used for determining risk factors – variables, that influence the risk of failure. Its application is limited by additional assumptions concerning proportional hazards and known functional form of independent variables effect. Other methods, accelerated failure time models (Marubini & Valsecci, 1995), may be used only when the failure time distribution is specified. Since the assumptions are often difficult to fulfill, other assumption-free models are developed. Many of them are based on computational intelligence techniques, mainly on artificial neural networks and survival trees.

In the chapter I would like to introduce the problem of survival analysis and to describe the possibilities of the use of computational intelligence techniques for analysis of survival data. Although the term "computational intelligence" covers many different techniques, we will narrow the examples of their applications to the most common approaches - artificial neural networks, survival trees, and ensembles of tree-based models.

BACKGROUND

In this section short introduction to survival data, as well as to artificial neural networks and tree based models is provided.

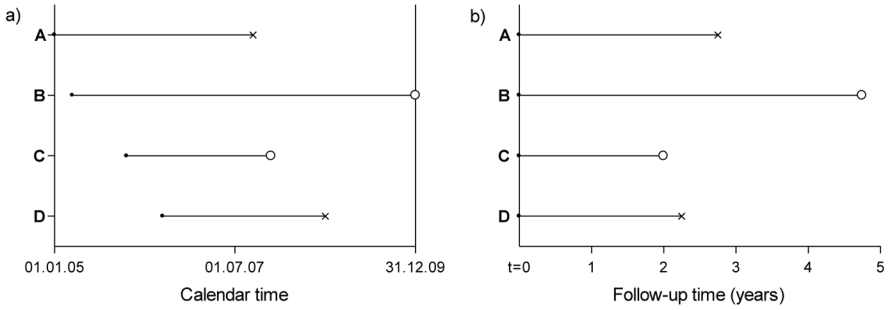
Survival Data

Survival analysis techniques are developed to predict the time of occurrence of a specified event, called a failure. For a given company, the failure may denote e.g. bankruptcy. If for all investigated companies the failure occurred, the classical regression models may be used. The problem arises when collected data do not contain the exact values of failure times. The lack of knowledge of exact event times is caused, on the one hand, by other, not investigated accidents, on the other hand, by the end of follow-up time.

Figure 1 presents two described above situations. Assuming that the follow-up time is a five-year interval, from 01.01.2005 to 31.12.2009, the observations (e.g. companies) are included into the study just after a certain starting event has occurred. The time of starting event is a starting point of their follow-up $t = 0$ (Figure 1b). As we can see in Figure 1a), for the observations A and D the failure has occurred during the follow-up

DOI: 10.4018/978-1-4666-5202-6.ch044

Figure 1. Observations in two points of reference: a) calendar time, b) follow-up time; x indicates uncensored observation, o - censored observation



time, the observation C was lost to follow-up before 31.12.2009, and the sample B was observed to the end of follow-up time and during this time the failure did not occur. Therefore, the observations B and C are censored - the exact failure time is unknown for them, we only know that the failure time is not less than their follow-up time.

Assuming, that we observe n independent companies, the data consist of n independent samples $L=(\mathbf{x}_i, t_i, \delta_i)$, $i=1,\ldots,n$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i - survival time and δ_i - failure indicator, which is equal to 0 for censored cases and 1 for uncensored ones. To describe the distribution of failure time (a random variable T), we can use one of the fallowing functions:

- Survival function $S(t) = P(T > t)$ gives the probability of survival (event-free) up to time t;
- Cumulative distribution function $F(t) = P(T \leq t) = 1\text{-}S(t)$ calculates the probability of failure before or at time t;
- Hazard function (risk of failure) is a conditional probability of failure at time t and is calculated as $h(t)=\lim_{(\Delta t->\infty)} P(t \leq T<t+\Delta t|T \geq t)/\Delta t$;
- Probability density function $f(t)=\lim_{\Delta t->\infty} P(t \leq T<t+\Delta t|T \geq t)$.

To estimate the functions presented above we can use statistical methods which are divided into non-parametric, semi-parametric, and parametric models (Marubini & Valsecci, 1995; Kalbfleish & Prentice, 1980). The first group covers two estimators of survival function: the Kaplan-Meier estimator and the life table (actuarial) method. They are calculated on the base of t_i, and δ_i (without taking into account the value of covariates vector x_i) and might be specified for the whole data or for groups formed by different values of covariates (e.g. sex: male, female). To compare survival functions between two or more groups the log-rank test (and its extensions: the log-rank test for trend and stratified log-rank test) is widely used. Cox's proportional hazards model (Cox, 1972) belongs to the second group of models: semi-parametric ones. It has a form: $h(t)= h_0(t)\exp(b_i x_i)$, i=1,…N, where $h_o(t)$ is a baseline hazard and usually remains unknown and b_i are coefficients. Its application requires a strong assumption that the effects of variables on survival are constant over time and additive. Cox's proportional hazard model is used mainly to discover the influence of covariates on the hazard but does not estimate the shape of this function. Parametric models (accelerated failure time models) require specification of a particular type of distribution. The most common parametric models are: Weibull, lognormal, and loglogistic models. The use of parametric models is not very frequent, because the survival time distribution quite rarely takes a required shape.