Early-Stage Event Prediction for Longitudinal Data

Mahtab J. Fard[1], Sanjay Chawla[2,3], and Chandan K. Reddy[1]

[1] Computer Science Department, Wayne State University, Detroit, MI - 48202, mahtab.jahanbanifard@wayne.edu, reddy@cs.wayne.edu
[2] Qatar Computing Research Institute, HBKU, schawla@qf.org.qa
[3] University of Sydney, Sydney, NSW, Australia, sanjay.chawla@sydney.edu.au

Abstract. Predicting event occurrence at an early stage in longitudinal studies is an important problem which has high practical value. As opposed to the standard classification and regression problems where a domain expert can provide the labels for the data in a reasonably short period of time, training data in such longitudinal studies must be obtained only by waiting for the occurrence of sufficient number of events. The main objective of this work is to predict the event occurrence in the future for a particular subject in the study using the data collected at the initial stages of a longitudinal study. In this paper, we propose a novel Early Stage Prediction (ESP) framework for building event prediction models which are trained at early stages of longitudinal studies. More specifically, we develop two probabilistic algorithms based on Naive Bayes and Tree-Augmented Naive Bayes (TAN), called ESP-NB and ESP-TAN, respectively, for early stage event prediction by modifying the posterior probability of event occurrence using different extrapolations that are based on Weibull and Lognormal distributions. The proposed framework is evaluated using a wide range of synthetic and real-world benchmark datasets. Our extensive set of experiments show that the proposed ESP framework is able to more accurately predict future event occurrences using only a limited amount of training data compared to the other alternative approaches.

Keywords: Prediction, regression, longitudinal data, survival analysis.

1 Introduction

Developing effective prediction models to estimate the outcome of a particular event of interest is a critical challenge in various application domains such as healthcare, reliability, engineering, etc. [12]. In longitudinal studies, event prediction is an important area of research where the goal is to predict the event occurrence during a specific time period of interest [9]. Obtaining training data for such a time-to-event problem is a daunting task. As opposed to the standard supervised learning problems where a domain expert can provide labels in a reasonable amount of time, training data for longitudinal studies must be obtained only by waiting for the occurrence of sufficient number of events. Therefore, the ability to leverage only a limited amount of available information at early stages of longitudinal studies to forecast the event occurrence at future time points is an important and challenging research task.

Let us consider an illustrative example shown in Figure 1. In this example, a longitudinal study is conducted on 5 subjects and the information for event occurrence until time t_c is recorded, where only subjects B and E have experienced the event. The goal of our paper is to predict the event occurrence by the time t_f where t_f is much greater than t_c. It can be seen that, except subjects B and E, all the remaining subjects are considered to be censored at t_c (marked by red 'x') and the event will occur for subject A within the time period t_f. This sce-

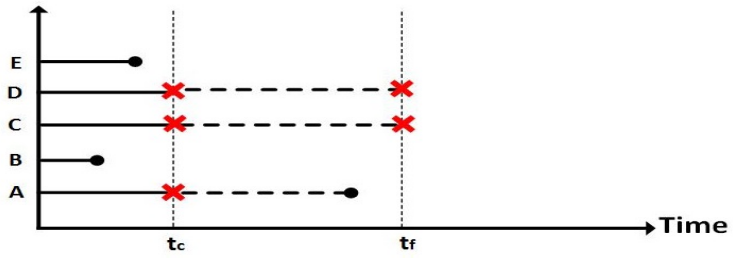


Fig. 1: An illustration to demonstrate the problem of early stage event prediction for time t_f using the information of event occurrence until time t_c.

nario is applicable for many real-world applications where it is critical to obtain early stage time-to-event predictions. For example, in the healthcare domain, let us say that there is a new treatment option (or drug) which is available and one would like to study the effect of such a treatment on a particular group of patients in order to understand the efficacy of the treatment. This patient group is monitored over a period of time and an event here corresponds to the patient being hospitalized (or occurrence of death) because the treatment has failed. The effectiveness of this treatment must be estimated as early as possible when there are only a few hospitalized patients.

This practical problem clearly emphasizes the need to build algorithms that can effectively predict events using the training data that contains only the event information at an early stage of a longitudinal study. It should be noted that the previous research in the field of statistics mainly focuses on the prediction of survivability up to a certain specific time point. Predicting events at future timepoints using the available information at the initial phases of the study remains to be a relatively unexplored area of research. Thus, in this paper, we develop prediction models using the data collected at earlier time points in longitudinal studies. More specifically, the contributions of this paper are as follows:

- Propose an **E**arly **S**tage **P**rediction (ESP) framework which estimates the probability of event occurrence for a future timepoint using different extrapolation techniques.
- Develop a probabilistic algorithms based on Naive Bayes and Tree-Augmented Naive Bayes (TAN), called ESP-NB and ESP-TAN, respectively, for early-stage event prediction by modifying the posterior probability of event occurrence.
- Evaluate the proposed algorithms using several synthetic and real-world benchmark datasets.

This paper is organized as follows. In Section 2, we present a summary of existing works on using survival analysis and machine learning methods for longitudinal data. In Section 3, we explain the problem formulation and describe two probabilistic classifiers, namely, Naive Bayes and Tree-Augmented Naive Bayes. In Section 4, we introduce the proposed extrapolation methods and then explain our novel Early Stage Prediction (ESP) framework based on Naive Bayes and TAN algorithms. In Section 5, the results of the proposed methods along with those of the competing algorithms on various synthetic and real-world datasets are presented. In the last section, we conclude our paper with a summary of the main results of the proposed work.

2 Related Work

Survival analysis is a subfield of statistics where a wide range of techniques have been proposed to model time-to-event data (e.g. failure, death, admission to hospital, emergence of disease, etc.) [13]. For such a time-to-event prediction problem, there have also been many attempts using different machine learning methodologies that were modified and applied to this problem [19, 21]. On the other hand, longitudinal data cannot be modeled solely using traditional classification or regression approaches since certain observations have event status and the rest have an unknown status up until that specific time of study.

Several machine learning approaches have been adapted to handle the concept of censoring in survival data [15]. Modifications of decision trees [8, 17], artificial neural networks [6] and support vector machines [11,18] represent some of the works on this topic. Another popular choice in the predictive modeling literature is the Bayesian approach. However, there was only a little work in the literature using Bayesian methods for survival data [1, 14, 20].

The work that is being developed in this paper is significantly different from the above mentioned algorithms since none of the existing works perform forecasting of event occurrence at future points in the context of survival data. They basically use the training data that is collected at the same time point as the test data. The basic idea of the proposed model is to develop Naive Bayes and its extension Tree-Augmented Naive Bayes (TAN), to build a predictive probabilistic model which will allow us to adapt the prior probability of events for forecasting the event occurrence at different points of time in the future. It is important to note that discriminative models are not suitable for the forecasting framework due to the lack of the prior probability component.

3 Preliminaries

The aim of our work is to address the following question: "when will a subject in longitudinal study experience an event?" The fundamental challenge here is to determine which subject in the study will experience the event at a certain timepoint based on event occurrence information that is available only until prior points of time (usually much earlier than the timepoint used during estimation). Before describing the details of the proposed model, we formalize the problem and transform it to a binary classification task. Then, we describe two well-known probabilistic classification approaches, namely, Naive Bayes and Tree-

Augmented Naive Bayes (TAN). Table 1 describes the notations used in this paper.

Table 1: Notations used in this paper

Name	Description
n	number of samples
m	number of features
$\mathbf{x}$	n×m data matrix
T	$n \times 1$ vector of event times
C	$n \times 1$ vector of last follow-up times
O	$n \times 1$ vector of observed time which is $min(T, C)$
δ	$n \times 1$ binary vector of censored status
t_c	specified time until which information is available
t_f	desired time at which the forecast of future events is made
$y_i(t)$	event status for subject i at time t

3.1 Problem Formulation

Let us consider a longitudinal study where the data about n independent subjects are available. Let the feature vector for sample i be represented by $\mathbf{x_i} = \langle x_{i1}, ..., x_{im} \rangle$ where x_{ij} is the j^{th} feature for subject i. For each subject i, we can define T_i as the event time, and C_i as the last follow-up time or censoring time (the time after which the subject has left the study). For all the subjects $i = \{1, ..., n\}$, O_i denotes the observed time which is defined as $min(T_i, C_i)$. Then, the event status can be defined as $\delta_i = \mathbf{I}\{T_i \leq C_i\}$. Thus, a longitudinal dataset can be represented as $(\mathbf{x_i}, T_i, \delta_i)$ where $\mathbf{x_i} \in \mathbf{R}^m$, $T_i \in \mathbf{R}^+$, $\delta_i \in \{0, 1\}$.

It should be noted that we only have the information for few events until the time t_c. Our aim is to predict the event status at time t_f where $t_f > t_c$. Let us define $y_i(t_c)$ as event status for subject i at time t_c. We consider t_c to be less than the observation time since we aim to forecast the event occurrence at early stage of the study. Suppose, among n subjects in the study, only $n(t_c)$ will experience the event at time t_c. For each subject i we can define

$$y_i(t_c) = \begin{cases} 1 & \text{if } O_i \leq t_c \text{ and } \delta_i = 1, \\ 0 & \text{otherwise} \end{cases}$$

In this transformed formulation, given the training data $(x_i, y_i(t_c))$, we can build a binary classifier using $y_i(t_c)$ as the class label. If $y_i(t_c) = 1$, then the event has occurred for subject i and if $y_i(t_c) = 0$, the event has not occurred. It should be noted that a new classifier will have to be built to estimate the probability of event occurrence at t_f based on the training data that is available at t_c.

3.2 Naive Bayes Method

Naive Bayes is a well-known probabilistic model in the machine learning domain. Assume we have a training set in Figure 1 where the event occurrence information is available up to time t_c. Based on the binary classification transformation explained above, using Naive Bayes algorithm, the event probability can be estimated as follows:

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \leq t_c\big) = \frac{P\big(y(t_c) = 1, t \leq t_c\big) \prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1\big)}{P(\mathbf{x}, t \leq t_c)} \qquad (1)$$

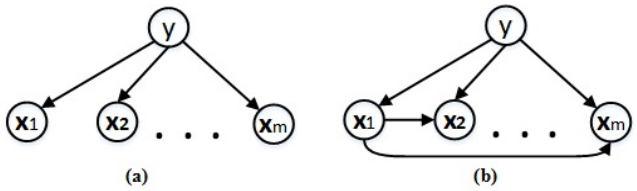


Fig. 2: An illustration of the basic structure of (a) Naive Bayes and (b) TAN classifier.

The first component of the numerator is the prior probability of the event occurrence at time t_c. The second component is a conditional probability distribution which can be estimated as follows:

$$P\big(\mathbf{x_j} \mid y(t_c) = 1\big) = \frac{\sum_{i=1}^{n} \big(y_i(t_c) = 1, x_{ij} = \mathbf{x_j}\big)}{\sum_{i=1}^{n} (y_i(t_c) = 1)} \tag{2}$$

where x_{ij} is the value of attribute j for subject i. Thus, it is a natural estimate for the likelihood function in Naive Bayes to count the number of times that event occurred at time t_c in conjunction with j^{th} attribute that takes a value of x_j. Then we count the number of times the event occurred at time t_c in total and finally take the ratio of these two terms. This formula is valid for discrete attributes; However, it can be easily adapted for continues variables as well [10].

3.3 Tree-Augmented Naive Bayes Method

A prominent extension of Naive Bayes is the Tree-Augmented Naive Bayes (TAN) where the independence assumption between the attributes is relaxed [7]. The TAN algorithm imposes a tree structure on the Naive Bayes model by restricting the interaction between the variables to a single level. This method allows every attribute $\mathbf{x_j}$ to depend upon the class as well as at most one other attribute, $\mathbf{x_p}(j)$, called the parent of $\mathbf{x_j}$. Illustration of the basic structure of the dependency in Naive Bayes and TAN is shown in Figure 2. Given the training set $(\mathbf{x}, y(t_c))$, firstly the tree for the TAN model should be constructed based on the conditional mutual information between two attributes [7].

$$I\big(\mathbf{x_j}, \mathbf{x_k} \mid y(t_c)\big) = \sum_{\mathbf{x_j}, \mathbf{x_k}, y(t_c)} P\big(\mathbf{x_j}, \mathbf{x_k}, y(t_c)\big) \frac{P\big(\mathbf{x_j}, \mathbf{x_k} \mid y(t_c)\big)}{P\big(\mathbf{x_j} \mid y(t_c)\big) P\big(\mathbf{x_k} \mid y(t_c)\big)} \tag{3}$$

Then, a complete undirected graph in which the vertices correspond to the attributes $\mathbf{x_j}$ is constructed. Using Equation (3), the weight of all the edges can be computed. A maximum weighted spanning tree is built and finally, an undirected tree is transformed into a directed one by randomly choosing a root variable and setting the direction of all the edges outward from the root. After the construction of the tree, the conditional probability of each attribute on its parent and the class label is calculated and stored. Hence, the probability of event at time t_c, can be defined as follows:

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \le t_c\big) = \frac{P\big(y(t_c) = 1, t \le t_c\big) \prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1, \mathbf{x_p}(j)\big)}{P(\mathbf{x}, t \le t_c)} \tag{4}$$

The numerator consists of two components; the prior probability of the event occurrence at time t_c and the conditional probability distributions which can be estimated using the maximum likelihood estimation (MLE).

4 The Proposed ESP Framework

In this section, we describe the proposed **E**arly **S**tage **P**rediction (ESP) framework. First, we describe our proposed prior probability extrapolation based method using different distributions and then we will introduce ESP-NB and ESP-TAN algorithms which utilize the extrapolation method.

4.1 Prior Probability Extrapolation

In order to predict event occurrence in longitudinal data, we develop a technique that can estimate the ratio of event occurrence beyond the original observation range or in other words, compute the *extrapolation for prior probability of event occurrence*. This extrapolation approach will be based on Weibull and Lognormal distributions which are used widely in the literature for modeling the time-to-event data [3, 16]. We will integrate such extrapolated values later with the proposed learning algorithms in order to make predictions at future timepoints.

Weibull: We estimate the shape and scale parameters, α_{t_c} and β_{t_c}, in Weibull distribution, by fitting the distribution to data obtained until t_c and then making the following extrapolation

$$p(t_f) = \frac{{t_f}^{\alpha-1}}{\beta^{\alpha}} exp\big(-({}^{t_f}/_{\beta})^{\alpha}\big) \tag{5}$$

Lognormal: We can also assume that the time to event follows a log-normal distribution, and then we can estimate μ_{t_c} and σ_{t_c}, mean and standard deviation of log-normal distribution, from the training data. The extrapolation is given as follows:

$$p(t_f) = \frac{1}{\sqrt{2\pi}\sigma_{t_c} t_f} exp - {}^{(log(t_f) - \mu_{t_c})^2}/_{2\sigma^2_{t_c}} \tag{6}$$

4.2 The ESP Algorithm

We will now describe the ESP Algorithm which consists of two phases. In the first phase, the conditional probability distribution is estimated using training data which is obtained until time t_c (see sections 3.2 and 3.3). In the second phase, we extrapolate the prior probability of event occurrence for time t_f which is beyond the observed time using different extrapolation techniques as follows:

$$P\big(y(t_f) = 1, t \leq t_f\big) = p(t_f) \tag{7}$$

It should be noted that the Eq. (7) can be estimated using Eqs. (5) and (6). Thus, the posterior probability for event occurrences at time t_f can be estimated as:

ESP-NB:

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big) = \frac{p(t_f) \prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1\big)}{P(\mathbf{x}, t \leq t_f)} \tag{8}$$

Algorithm 1: Early Stage Prediction (ESP) Framework

Require: Training data $D(t_c) = \big(\mathbf{x}, y(t_c), T\big)$, t_f
Output: Probability of event at time t_f
Phase 1: Conditional probability estimation at t_c
1. **for** $j = 1, ..., m$
2. Naive Bayes: $P\big(\mathbf{x_j} \mid y(t_c) = 1\big)$ (Eq. (2))
3. TAN: $P\big(\mathbf{x_j} \mid y(t_c) = 1, x_p(j)\big)$ (Eq. (3))
4. **end**
Phase 2: Predict probability of event occurrence at t_f
5. Estimate $P\big(y(t_f) = 1, t \leq t_f\big)$
6. Weibull: ${t_f}^{\alpha-1}/\beta^{\alpha} exp\big(-({t_f}/\beta)^{\alpha}\big)$ (Eq. (5))
7. Lognormal: $1/\sqrt{2\pi}\sigma_{t_c} t_f exp - (log(t_f) - \mu_{t_c})^2/2\sigma_{t_c}^2$ (Eq. (6))
8. **for** $i = 1, ..., n$
9. Estimate $P\big(y_i(t_f) = 1 \mid \mathbf{x}_i, t \leq t_f\big)$
10. ESP-NB: Eq. (8)
11. ESP-TAN: Eq. (9)
12. **end**
13: return $P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big)$

ESP-TAN:

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big) = \frac{p(t_f) \prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1, \mathbf{x_p}(j)\big)}{P(\mathbf{x}, t \leq t_f)} \tag{9}$$

Algorithm 1 outlines the proposed ESP method. In the first phase (lines 1-4), for each attribute j, the algorithm estimates the conditional probability using the data available until time t_c. In the second phase, a probabilistic model is built to predict the event occurrence at t_f. In lines 5-7, the prior probability for event occurrence at time t_f is estimated using different extrapolation techniques. Then, in lines 8-12, for each subject i, we adapt the posterior probability of event occurrence at time t_f. The time complexity of the ESP algorithm follows the time complexity of the learning method that is chosen. It should be noted that the complexity of the extrapolation component is a constant and does not depend on either m or n. Hence, for ESP-NB, the overall complexity is $O(mn)$ and for ESP-TAN, it is $O(m^2 n)$, where n is the total number of subjects and m is the number of features in the dataset.

5 Experimental Results

In this section, we will describe the datasets that are used for evaluating the proposed methods along with the comparisons of the proposed algorithms with various baseline prediction methods.

5.1 Dataset Description

We evaluated the performance of the models using both synthetic and real-world survival datasets which are summarized in Table 2.

Synthetic Datasets: We generated synthetic dataset in which the feature vectors $\mathbf{x}$ are generated based on a normal distribution $N(0, 1)$. Covariate coefficient vector β is generated based on a uniform distribution $Unif(0, 1)$. Thus,

Table 2: Number of features, instances and events. T_{50} and T_{100} corresponds to the time taken for the occurrence of 50% and 100% of the events, respectively.

Dataset	#Features	#Instances	#Events	$\mathbf{T_{50}}$	$\mathbf{T_{100}}$
Syn1	5	100	50	1014	3808
Syn2	20	1000	602	943	7723
Breast	8	673	298	646	2659
Colon	13	888	445	394	3329
PBC	17	276	110	1191	4456
Framingham	16	5209	1990	1991	5029
EHR	77	4417	3479	50	4172
Kickstarter	54	4175	1961	21	60

T can be generated using the method described in [2]. Given the observed covariates $\mathbf{x}_i$ for observation i, the failure time can be generated by

$$T_i = -\left(\frac{log(Unif(0,1))}{\lambda \dot{e}xp(\beta' \mathbf{x}_i)}\right)^{\nu} \tag{10}$$

In our experiments, we set $\lambda = 0.01$ and $\nu = 2$.

Real-world Survival Datasets: Several real-world survival benchmark datasets were used in our experiments. We used primary biliary cirrhosis (PBC), breast and colon cancer datasets (available in the survival data repository [4]) which are widely used in evaluating longitudinal studies. We also used Framingham heart study dataset which is publicly available [4]. In addition, we also used two in-house proprietary datasets. One is the electronic health record (EHR) data from heart failure patients collected at the Henry Ford Health System in Detroit, Michigan. This data contains patient's clinical information such as procedures, medications, lab results and demographics and the goal here is to predict the number of days for the next readmission after the patient is discharged from the hospital. Another dataset was obtained from Kickstarter, a popular crowdfunding platform. Each project has been tracked for a specific period of time. If the project reaches the desired funding goal within deadline date then it is considered to be a success (or event occurred). On the other hand, the project is considered to be censored if it fails to reach its goal within the deadline date.

5.2 Performance Evaluation

The performance of the proposed models is measured using following metrics,

- AUC is the area under the receiver operating characteristic (ROC) curve. The curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) by varying the threshold value.
- *F-measure* is defined as a harmonic mean of precision and recall. A high value of F-measure indicates that both precision and recall are reasonably high.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

[4] `http://cran.rproject.org/web/packages/survival/`

Table 3: Comparison of AUC values for Cox, NB and TAN with proposed ESP-NB and ESP-TAN methods using Weibull (W) and Lognormal (L) extrapolation methods (with standard deviation values).

Data	AUC						
	Cox	**NB**	**TAN**	**ESP-NB(W)**	**ESP-NB(L)**	**ESP-TAN(W)**	**ESP-TAN(L)**
Syn1	0.697 (0.004)	0.702 (0.007)	0.713 (0.002)	0.865 (0.003)	0.841 (0.003)	**0.865 (0.001)**	0.849 (0.001)
Syn2	0.703 (0.003)	0.699 (0.009)	0.705 (0.005)	0.818 (0.002)	0.811 (0.003)	**0.821 (0.002)**	0.817 (0.002)
Breast	0.612 (0.011)	0.621 (0.009)	0.632 (0.004)	0.655 (0.001)	0.633 (0.003)	**0.662 (0.007)**	0.635 (0.005)
Colon	0.601 (0.024)	0.615 (0.011)	0.617 (0.014)	0.621 (0.013)	0.617 (0.014)	**0.627 (0.009)**	0.619 (0.011)
PBC	0.665 (0.009)	0.643 (0.003)	0.679 (0.01)	0.765 (0.001)	0.761 (0.004)	**0.768 (0.003)**	0.763 (0.001)
Framingham	0.863 (0.006)	0.945 (0.002)	0.953 (0.005)	0.953 (0.007)	0.959 (0.003)	0.961 (0.004)	**0.971 (0.002)**
EHR	0.612 (0.022)	0.633 (0.019)	0.638 (0.025)	0.654 (0.018)	0.624 (0.021)	**0.649 (0.011)**	0.628 (0.026)
Kickstarter	0.761 (0.018)	0.811 (0.022)	0.816 (0.025)	0.821 (0.024)	0.825 (0.023)	0.822 (0.019)	**0.831 (0.018)**

Implementation Details: The proposed ESP-NB and ESP-TAN methods are implemented using *e1071* package available in the R programming language [5]. The same package used for comparison results from Naive Bayes and TAN classification model. The *coxph* model in the survival package is employed to train the Cox model. The source code of the proposed algorithms in R programming environment is available at `http://dmkd.cs.wayne.edu/codes/ESP` .

5.3 Results and Discussion

For performance benchmarking, we compare the proposed ESP-NB and ESP-TAN algorithms using Weibull and Lognormal distributions as extrapolation techniques with Cox regression, Naive Bayes (NB) and Tree-Augmented Naive Bayes (TAN) classification methods which are trained at time when only 50% of events have occurred and the event prediction is done at the end of study. Table 3 and Table 4 summarize the comparison result in AUC and F-measure evaluation metrics, respectively. We used stratified 10-fold cross-validation and average values (along with the standard deviations) of the results on all the ten folds are being reported. For all of the datasets, our results evidently show that the proposed ESP-based methods using either Weibull or lognormal distribution will provide significantly better prediction results compared to the other methods. The choice of the optimal distribution will depend on the nature of the dataset being considered, in particular, the distribution that the event occurrence follows. Furthermore, ESP-NB build on independence assumption between the attributes which does not hold in many survival applications. Thus, the introduced ESP-TAN relaxed the independence assumption which leads to improved AUC and F-measure values in almost all of the results.

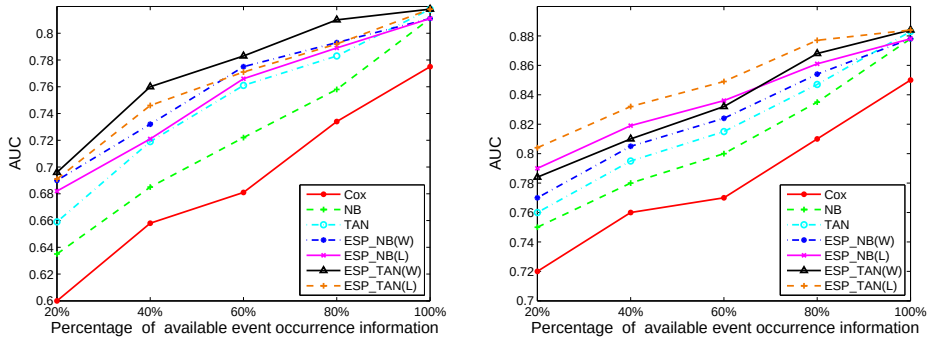
The results clearly show that our models can obtain practically useful results using hte data collected at an early stage of the study. This is due to the fact that classification methods do not have the ability to predict the event occurrence for a time beyond the observation time. Also, in the Cox regression model,

Table 4: Comparison of F-measure values for Cox, NB and TAN with proposed ESP-NB and ESP-TAN methods using Weibull (W) and Lognormal (L) extrapolation methods (with standard deviation values).

Data	F-measure						
	Cox	**NB**	**TAN**	**ESP-NB(W)**	**ESP-NB(L)**	**ESP-TAN(W)**	**ESP-TAN(L)**
Syn1	0.632 (0.023)	0.753 (0.021)	0.762 (0.026)	0.775 (0.021)	0.771 (0.022)	**0.785 (0.019)**	0.785 (0.023)
Syn2	0.629 (0.025)	0.638 (0.034)	0.647 (0.023)	0.764 (0.025)	0.763 (0.029)	**0.777 (0.02)**	0.769 (0.021)
Breast	0.628 (0.031)	0.543 (0.053)	0.555 (0.034)	0.712 (0.039)	0.653 (0.042)	**0.723 (0.039)**	0.679 (0.039)
Colon	0.496 (0.163)	0.523 (0.169)	0.529 (0.184)	0.619 (0.145)	0.606 (0.151)	**0.626 (0.148)**	0.623 (0.15)
PBC	0.603 (0.141)	0.529 (0.121)	0.535 (0.11)	0.709 (0.11)	0.664 (0.109)	**0.715 (0.098)**	0.698 (0.114)
Framingham	0.755 (0.079)	0.787 (0.085)	0.798 (0.073)	0.865 (0.073)	0.873 (0.093)	0.894 (0.069)	**0.905 (0.056)**
EHR	0.672 (0.125)	0.616 (0.156)	0.623 (0.198)	0.781 (0.126)	0.750 (0.206)	**0.798 (0.16)**	0.781 (0.12)
Kickstarter	0.672 (0.084)	0.713 (0.058)	0.719 (0.067)	0.747 (0.034)	0.742 (0.054)	0.762 (0.048)	**0.775 (0.032)**

the baseline hazard is undefined after the observation time t_c. Thus, from our experiments, we can conclude that the proposed framework is able to obtain practically useful results at the initial phases of a longitudinal study and can provide good insights about the event occurrence by the end of the study.

In Figure 3, we present the prediction performance of different methods by varying the percentage of event occurrence information that is available to train the model for the PBC dataset. For example, 20% on the x-axis corresponds to the training data obtained when only 20% of the events have occurred and prediction of the event occurrences was made for the end of the study period. From this plot we can see that the AUC values improve when there is more information on the event occurrence in the training data. For all the cases, our proposed ESP framework gives better prediction performance compared to other techniques. Furthermore, it should be noted that the improvements of the proposed methods are more significant over the baseline methods when there is only a limited amount (20% or 40%) of training data. Also, when 100% of the training data is available, the performance of the proposed methods will converge to that of the standard Naive Bayes and TAN methods since the prior probabilities in both scenarios will be the same and fitting a distribution will not have any impact when evaluated at the end of the study. The proposed prediction framework is an extremely useful tool for domains where one has to wait for a significant period of time to collect sufficient amount of training data. The practical implication of this result is the fact that using the proposed models, one can obtain an approximate result and gain insights about the problem within the early stage of the study. Thus, it is not needed to wait until the end of the study to obtain the model performance. Also, we can observe that, in many real-world datasets, 50% of the events typically occur within 25% of the total study time. Such an early stage model building is an extremely useful tool for domains where one has to wait for longer time periods to collect the required training data.



(a) PBC **(b)** Kickstarter

Fig. 3: AUC values of different methods obtained by varying the percentage of event occurrence information for the PBC and Kickstarter dataset.

6 Conclusion

In many real-world application domains, it is important to be able to forecast the occurrence of future events by only using the data collected at early stages in longitudinal studies. In this paper, we developed event prediction algorithms by extending Bayesian methods through fitting a statistical distribution to time-to-event data with fewer available events at the early stages. This enables us to have a reliable prediction of event occurrence for future time points. Our extensive experiments using both synthetic and real datasets demonstrate that the proposed ESP-based algorithms are more effective than Cox model and other classification methods in forecasting events at future time points. Also, we investigated different kinds of extrapolation approaches by fitting various distributions such as Weibull and log-normal. Though motivated by biomedical and healthcare application scenarios (primarily for estimating survival), the proposed algorithms are also applicable to various other domains where one needs to predict event occurrences at early stage of analysis when there are only a relatively fewer set of events that have occurred until a certain time point.

Acknowledgements

This work was supported in part by the National Science Foundation grants IIS-1527827 and IIS-1231742.

References

1. S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. OConnor. Data mining for censored time-to-event data: A bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29(4):1033–1069, 2015.
2. R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 25:1978–1979, 2006.
3. K. J. Carroll. On the use and utility of the weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6):682–701, 2003.

4. T. R. Dawber, W. B. Kannel, and L. P. Lyell. An approach to longitudinal studies in a community: the framingham study. *Annals of the New York Academy of Sciences*, 107(2):539–556, 1963.
5. E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M. F. Leisch. Package e1071. *R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html*, 2009.
6. M. J. Donovan, M. J. Donovan, S. Hamann, M. Clayton, and et. al. Systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(24):3923–9, Aug. 2008.
7. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.
8. L. Gordon and R. Plshen. Tree-structured survival analysis. *Cancer Treat Reports*, 69(10):1065–1074, 1985.
9. D. W. Hosmer and S. Lemeshow. *Applied survival analysis: regression modeling of time to event data*. Wiley, New York, 1999.
10. G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
11. F. M. Khan and V. B. Zubek. Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. In *Eighth IEEE International Conference on Data Mining*, pages 863–868, 2008.
12. N. Lavrac. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 16:3–23, 1999.
13. E. T. Lee and J. Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.
14. P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, 30(3):201–14, Mar. 2004.
15. C. K. Reddy and Y. Li. A review of clinical prediction models. In C. K. Reddy and C. C. Aggarwal, editors, *Healthcare Data Analytics*. Chapman and Hall/CRC Press, 2015.
16. P. Royston. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55(1):89–104, 2001.
17. M. R. Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988.
18. H.-T. Shiao and V. Cherkassky. Learning using privileged information (LUPI) for modeling survival data. *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1042–1049, July 2014.
19. I. Štajduhar and B. Dalbelo-Bašić. Uncensoring censored data for machine learning: A likelihood-based approach. *Expert Systems with Applications*, 39(8):7226–7234, 2012.
20. J. Wolfson, S. Bandyopadhyay, M. Elidrisi, G. Vazquez-Benitez, D. M. Vock, D. Musgrove, G. Adomavicius, P. E. Johnson, and P. J. O'Connor. A naive bayes machine learning approach to risk prediction using censored, time-to-event data. *Statistics in medicine*, 34(21), 2015.
21. B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.