1-1-2015

Bayesian Approach For Early Stage Event Prediction In Survival Data

Mahtab Jahanbani Fard
Wayne State University,

BAYESIAN APPROACH FOR EARLY STAGE EVENT PREDICTION IN SURVIVAL DATA

by

MAHTAB JAHANBANI FARD

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2015

MAJOR: COMPUTER SCIENC

Approved by:

Advisor Date

DEDICATION

To my husband for his endless encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

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 [2, 22, 32]. 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 [21]. Obtaining training data for such a time-to-event problem is a daunting task. Such studies also encounter incomplete data which occurs because of loss to follow (also known as censoring). In another words, the time to the event occurrence is not necessarily observed for all instances in the study. Thus, building event forecasting models in the presence of censored data is an important and challenging task which has significant practical value in longitudinal studies.

One of the primary challenges in the survival analysis studies is that as opposed to the standard supervised learning problems where a domain expert can provide labels in a reasonable amount of time, training data for these 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 analysis to forecast the event occurrence in future time points is an important and challenging research task.

The main objective of this work is to predict for which subject in the study event will occur at future based on few event information at the initial stages of a longitudinal study. In this thesis, we introduce a new method for handling censored data using Kaplan-Meier estimator. We also 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 extended the Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network methods based on the proposed framework, and developed three algorithms, namely, ESP-NB, ESP-TAN and ESP-BN, to effectively predict event occurrence using the training data obtained at early stage of the study. 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 methods.

This thesis is organized as follows. The rest of this chapter discusses the motivation and statement of problem along with contribution of this research. In chapter 2 we propose the Bayesian approach for survival data early stage event prediction. We develope Naive Bayes, Tree-Agumented Tree (TAN) and Bayesian Network to address this problem. Chapter 3 demonstrates the experimental results and shows the practical significance of our work using different benchmark and real-word dataset. Finally, chapter 4 concludes the discussion along with some future research directions in this area.

1.2 Motivation

It has become a common practice in many application domains to collect data over a period of time and record any interesting events that occur within this time. Survival analysis aims at finding the underlying distribution for data that measure the length of time until the occurrence of an event. In another word, the primary objective of such longitudinal studies is to determine the probability of the occurrence of a particular event of interest within a specific unseen time point. However, it cannot give an answer to the open question of *"how to forecast whether a subject will experience event by end of study having event occurrence information at early stage of survival data?"*. This problem exhibits two major challenges: 1) absence of complete information about event occurrence (censoring) and 2) availability of only a partial set of events that occurred during the initial phase of the study.

In order to have better idea, let us consider the following real-world applications which motivate the early stage time-to-event predictions.

- 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 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.

- Reliability prediction focuses on developing an accurate models that can estimate how reliable a newly released product will be. An event here corresponds to the time taken for a device to fail. In such applications, it is desirable to be able to estimate which devices will fail and if so, when they will fail. If such models can be learned using information from only a few device failures, then early warnings can be given about future failures.

- In credit score modeling applications, it is challenging to have an accurate estimation of whether a customer will default or not and if they default, when it is going to happen? If a prediction model can be accurately built using only few default individuals, then better precautions can be taken against those who will most likely default in the future.

These practical scenarios clearly emphasize the need to build algorithms that can effectively make predictions of events using the training data that contains only a few events (at an early stage). More precisely, the goal here is to predict the event occurrence for a time beyond the observation time window where only a few events have occurred.

Thus, the primary goal of this paper is to develop a method that can use only a limited amount of available information at the initial phase of a longitudinal study to forecast the event occurrence at future time points.

For a better understanding of the complexities and concerns related to this problem, let us consider an illustrative example shown in Figure 1.1. In this example, a longitudinal study is conducted on 6 subjects and the information for event occurrence until time t_c is recorded, where only subjects S2 and S5 had experienced the event. The goal of our work is to predict the event occurrence by time t_f (e.g. end of study). It should be noted that except subjects S2 and S5, all other are considered to be censored at t_c (marked by 'X'). Also, event will be occurred for subjects S1 and S6 within the time period t.

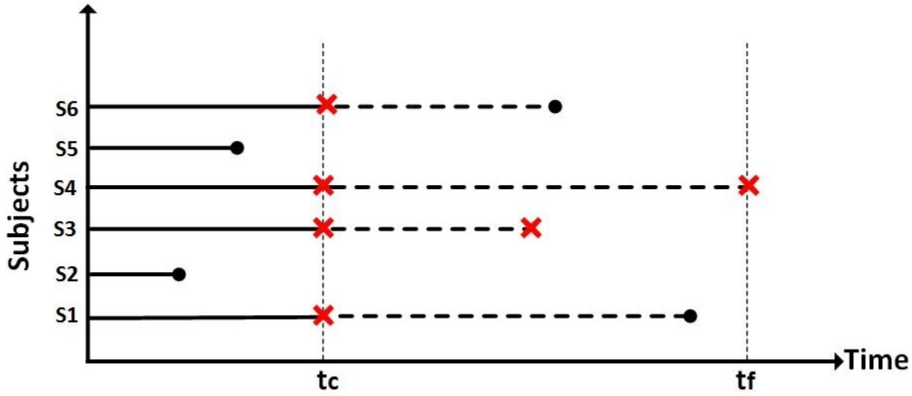


Figure 1.1: An illustration to demonstrate the problem of event forecasting at time t_f (e.g. end of study) using the information only until time t_c.

This scenario clearly motivates the need for building algorithms that can effectively forecast events using the training data at time t_c when only a few events have occurred. This problem is an important one in the domain of longitudinal studies since the only way to collect reliable data is to wait for sufficient period of time till complete information about event occurrence acquired.

The recently proposed popular variants in the machine learning field such as classification, semi-supervised learning, transfer learning, imbalance learning and multi-task learning are not suitable for tackling this problem primarily due to the fact that obtaining a labeled training set at the end of the study is not feasible since the data is available only until t_c. On the other hand, advanced statistical techniques, especially in the field of survival analysis, do not have the ability to handle the problem of predicting event occurrence for a time later than the observation time. The reason is that the probability of event provided by survival model is valid only for the specific observed time. It should be noted that this problem is completely different from the time series forecasting problem since the goal here is to predict the outcome of (binary) event occurrence for each subject for a time which is much beyond the observation time (as opposed to merely predicting the next time step value which is typically done in the standard time series forecasting models). Also such longitudinal survival data normally has missing information on events during the observation time. This incompleteness in events makes it difficult for standard machine learning methods to model such data. While ignoring this censored data will provide a suboptimal model because of neglecting the available information, treating censoring time as the actual time of event occurrence will provide an underestimate of the true performance of the model.

1.3 Thesis Contributions

In order to find an answer for the problem discussed above, we introduce an intuitive technique to handle the censoring problem in the longitudinal survival data. We also develop a Bayesian framework for early stage event prediction to tackle the problem of lack of sufficient training data on event occurrence in the initial phases (early stage) of longitudinal studies. Thus the main contributions of this thesis can be summarized as follows:

- Develop a new labelling method to handle censoredness in longitudinal studies using the Kaplan-Meier estimator.
- Propose an **E**arly **S**tage **P**rediction (ESP) framework which estimates the prob-

ability of event occurrence for a future time point using different extrapolation techniques.

- Develop a probabilistic algorithm based on Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network, called ESP-NB, ESP-TAN and ESP-BN respectively, for early-stage event prediction by adapting the posterior probability of event occurrence.

- Evaluate the proposed algorithms using several synthetic and real-world benchmark datasets and compare the effectiveness of the proposed methods with various classification and survival methods.

CHAPTER 2

PROPOSED BAYESIAN APPROACH

In this chapter we introduce our proposed Bayesian approach for handling early stage event prediction. As discussed in previous chapter predicting event occurrence at an early stage in longitudinal studies is a challenging problem. It is in contrast with the standard classification and regression problems where the labels for the data can be provided in a reasonably short period of time. Thus, for this longitudinal studies training data must be obtained only by waiting for the occurrence of sufficient number of events. On the other hand survival analysis method do not have the ability to handle the problem of predicting event occurrence for a time later than the observation time because the probability of event provided by survival model is valid only for the specific observed time. Therefore, the main objective of this chapter is to propose a framework to predict for which subject in the study event will occur at future based on few event information at the initial stages of a longitudinal study. Before we discuss the framework in detail, the related work in the areas of using machine learning techniques for survival analysis will be briefly presented.

2.1 Related Literature

Survival analysis is a subfield of statistics where a wide range of techniques have been proposed to model time-to-event data [37] in which the dependent variable is subject to censoring (e.g. failure, death, admission to hospital, emergence of disease etc.) [33]. The fact is Ordinary Least-Squares (OLS), the most common method for solving regression problem based on minimizing sum of squared error, does not work in the presence of censoring because it is not possible to estimate the error between the true response and the predicted response that comes from regression model [36]. However, while we do not know the ordinate in censored observation, the well-known likelihood method which finds the probability that the experiment turned out the way it did, can solve the censored regression problem [5]. Different techniques have been proposed based on Maximum Likelihood Estimation (MLE) to overcome the difficulty of handling censored data [11, 29].

There has been an increasing interest in adapting popular machine learning techniques to survival data [35]. However, longitudinal data cannot be modeled solely by traditional classification or regression approaches since certain observations have event status (or class label as event) and the rest have an unknown status up until that specific time of study. The censored observations in survival data might look similar to unlabelled samples in classification or unknown response in regression problem in the sense that status or time-to-event is not known for some observations. Such censored data have to be handled with a special care within any machine learning method in order to have an accurate prediction. Also, for censored data in survival analysis we have information up to a certain time point before censoring occurs and this information should be included in the model in order to obtain the most optimal result. Hence, the standard semi-supervised techniques [7, 50] are not directly applicable for this problem.

Several remarkable adjusted machine learning approaches have been proposed recently to address censored survival data issue. Decision trees [18, 39, 46] and Artificial Neural Networks (ANN) [4, 9, 10, 14] for censored data represent some of the earliest works in this field. Well-known Support Vector Machines (SVM) have been adopted to model survival data. Most of these methods treat the problem as regression [28, 41, 42, 46]. Other studies try to formalize the problem under the classification setting [15, 40]. However, comparison of the performance of these approaches yield no significant improvements over standard Cox model either. There are also few other studies aim at handling censored data during a preprocessing step by giving some weights to the censored observations [44, 51]. In this thesis, we tackle the problem of censoring using Kaplan-Meier method [27] to estimate the probability of event and probability of censoring for each censored subject. Such an intuitive approach can be easily applied on survival data before any further analysis is performed.

One of the popular choice in the predictive modeling literature is the Bayesian models including Naive Bayes and Bayesian Network where they have been used widely for classification [16] and successfully applied in many domains [17]. However, there has been only a little work in the literature using Bayesian methods for survival data [1, 35, 49]. Bayesian networks can visually represent all the relationships between the variables which makes it interpretable for end user. It is in contrast with simple Naive Bayes method that has the independence assumption between all features [16]. Despite the applicability of Bayesian network in the survival analysis domain, limited number of research efforts exist for tackling the censored data challenges. The authors of [34] developed a Bayesian neural network approach to model censored data. [43] gives weight to censored instances in order to learn Bayesian networks from survival data. Recently, [1] adapts a Bayesian network for survival data using an approach called inverse probability of censored weighting (IPCW) for each of the record in the dataset to handle the censoring issue.

Our work is significantly different from these previous studies since none of these works perform forecasting of event occurrence for a time beyond the observation time. They basically use the training data that is collected at the same time point as the test data. The idea is to take advantage of generative component of a Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian network to build a predictive probabilistic model [25] which will allow us to adapt the prior probability of event for different time points during forecasting. Also, it is important to note that discriminative models such as support vector machines or logistic regression are not suitable for the forecasting framework due to the lack of the prior probability component. On the other hand, for discriminative models there is no need to model the distribution of the observed variables.

Thus, they cannot be a good choice when we want to express more complex relationships between the dependent variable and other attributes [31].

2.2 Preliminaries

This section introduces some of the preliminaries required to comprehend the proposed framework. First the notations used in the study and problem formulation are described. Next, some basics about survival analysis are explained. Finally, more details are provided about Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network as the vital components of the proposed method for predicting event in survival data at an early stage of studies.

2.2.1 Problem Formulation

We begin by presenting the basic concepts and notations for survival analysis and Bayesian networks. Table 2.1 describes the notations used in this paper.

Table 2.1: Notations used in this thesis

Name	**Description**
n	number of subjects
m	number of features
$\mathbf{x_i}$	$1{\times}m$ matrix of feature vectors for subject i
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 for event 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
$F(t)$	Cumulative event probability at time t
$S(t)$	Survival probability at time t

Let us consider a longitudinal study where the data about n independent subjects are available. Let the features are represented by a m-dimensional vector $\mathbf{x_i} = \langle x_{i1}, ..., x_{im} \rangle$ 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 $\mathbf{min}(T_i, C_i)$. 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 $D = \{\mathbf{x_i}, T_i, \delta_i; i = 1, ...n\}$ 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

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

In this transformed formulation, given the training data $(\mathbf{x_i}, y_i(t_c))$, 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 $y_i(t_c) = 0$, then 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.

2.2.2 Survival Analysis

In general, survival analysis is defined as a collection of statistical methods which contains time of a particular event of interest as the outcome variable to be estimated. In many survival applications, it is common to see that the observation period of interest is incomplete for some subjects and such a data is considered to be *censored* [38]. Considering the duration to be a continuous random variable T, the survival function, $S(t)$

is the probability that the time of event occurrence is later than a certain specified time t, which is defined as

$$S(t) = \Pr(T > t) = \int_t^\infty f(u)\,du = 1 - F(t) \tag{2.1}$$

where $f(t)$ is a probability density function and $F(t)$ is a cumulative distribution function. Survival analysis involves the modelling of time-to-event data. We will use one of the popular parametric methods in survival analysis, accelerated failure time (AFT) [47] model, to adapt the probability of event using different time-to-event distributions.

2.2.3 Naive Bayes Classifier

Naive Bayes is a well-known probabilistic model in machine learning domain. Assume we have a training set in Figure 1.1 where the event occurrence information is available up to time t_c. Using binary classification transformation explained above, based on Naive Bayes algorithm the event probability can be estimated as follows:

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \le t_c\big) = \frac{P\big(y(t_c) = 1, t \le t_c\big) \prod_{j=1}^{m} P\big(x_j \mid y(t_c) = 1\big)}{P(\mathbf{x}, t \le t_c)} \tag{2.2}$$

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 and can be estimated as

$$P\big(x_j \mid y(t_c) = 1\big) = \frac{\sum_{i=1}^{n} \big(y_i(t_c) = 1, x_{ij} = x_j\big)}{\sum_{i=1}^{n} (y_i(t_c) = 1)} \tag{2.3}$$

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 jth attributes 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 [24].

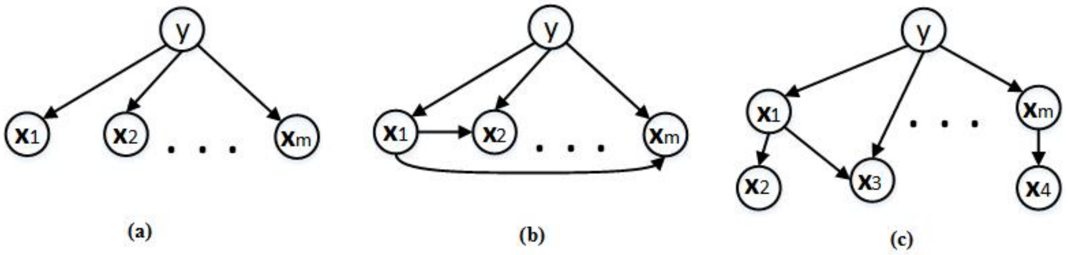


Figure 2.1: An illustration of the basic structure of (a) Naive Bayes(b) TAN and (c) Bayesian Network classifier.

2.2.4 Tree-Augmented Naive Bayes Classifier

One extension of Naive Bayes is the Tree-Augmented Naive Bayes (TAN) where the independence assumption between the attributes is relaxed [16]. The TAN algorithm imposes a tree structure on the Naive Bayes model by restricting the interaction among the variables to a single level. This method allows every attribute x_i to depend upon the class and at most one other attribute, $x_p(i)$, called the parent of x_i. Illustration of the basic structure of the dependency in Naive Bayes and TAN is shown in Figure 2.1. Given the training set $(\mathbf{x}_i, y_i(t_c))$, firstly the tree for the TAN model should be constructed based on the conditional mutual information between two attributes [16].

$$I\big(\mathbf{x_i}, \mathbf{x_j} \mid y(t_c)\big) = \sum_{x_i, x_j, y(t_c)} P\big(x_i, x_j, y(t_c)\big) \frac{P\big(x_i, x_j \mid y(t_c)\big)}{P\big(x_i \mid y(t_c)\big) P\big(x_j \mid y(t_c)\big)} \tag{2.4}$$

Then, a complete undirected graph in which the vertices correspond to the attributes x_i is constructed. Using Equation (2.4), the weight of all the edges can be computed. A maximum weighted spanning tree is built and finally 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:

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \leq t_c\big) = \frac{P\big(y(t_c) = 1, t \leq t_c\big) \prod_{j=1}^{m} P\big(x_j \mid y(t_c) = 1, x_p(j)\big)}{P(\mathbf{x}, t \leq t_c)} \tag{2.5}$$

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 maximum likelihood estimation (MLE).

2.2.5 Bayesian Networks Classifier

A Bayesian network is a graphical representation of a probability distribution over a set of variables. It can be consider as an extension for TAN model where features can be related to each other in different levels (Figure 2.1). It consists of two parts [19]:

1) a directed network structure in the form of a directed acyclic graph (DAG) which can be shown as $G = (V, E)$, where V denotes the set of vertices which represent variables, while E is the set of edges which show the dependence between the variables;

2) a set of the local probability distributions, one for each node variable, conditional on each value combination of its parents.

Thus, a Bayesian network can be formally defined as $BN = \big(G, P(G|D)\big)$ where $P(G|D) = \mathcal{L}\big(D|G, P(G|D)\big)$ is the networks likelihood on given data D. The Bayesian network structure in this thesis is learnt by the well-known search-and-score based Hill-climbing algorithm [20]. The weight-adapted MDL scoring (Eq. (2.6)) function is used as the criterion function to be minimised for the Hill-climbing algorithm [30].

$$MDL(BN, D) = \frac{d}{2}\log(N) - \log \mathcal{L}\big(D|G, P(G|D)\big) \tag{2.6}$$

where d is the number of free parameters of a multinomial local conditional probability distribution table. The second component of a Bayesian Network is a set of local conditional probability distributions. Together with the graph structure, these distributions are sufficient to represent the joint probability distribution of the domain. Joint probability is defined as the probability that a series of events will happen concurrently and hence it can be calculated from the product of individual probabilities of the nodes:

$$P(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \prod_{j=1}^{m} P(\mathbf{x_j} \mid Pa(\mathbf{x_j})) \tag{2.7}$$

where $Pa(\mathbf{x_j})$ is the set of parents of $\mathbf{x_j}$. Hence, given a training set, the goal of the Bayesian Network is to find the best graph structure to correctly predict the label for y given a vector of m attributes $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m})$. It can be formulated as follows:

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \leq t_c\big) = \frac{P\big(y(t_c) = 1, t \leq t_c\big) \prod_{j=1}^{m} P\big(x_j \mid y(t_c) = 1, Pa(\mathbf{x_j})\big)}{P(\mathbf{x}, t \leq t_c)} \tag{2.8}$$

In Eq. (2.8), the first element in numeraor is the prior probability of the class and the second element is the joint probability of the attributes based on the graph structure . A Bayesian Network is a generative classifier with a full probabilistic model of all variables which enable us to adapt the prior probability of event for different time points (beyond the observation time) during the forecasting.

2.3 Handling Censored Data

Two naive approaches to handle censored data are: (1) completely exclude them from the analysis which will result in losing important information (2) treat censored time as an actual event time which will induce a bias in the estimation of event time. Instead of following these approaches, our work handles censored data by dividing them into two groups [51]: *event* and *event-free*. For each censored instance, we estimate the probability of event and probability of censoring using Kaplan-Meier estimator and give a new class label based on these probability values. This approach assumes that the censoring time is independent of the event time and all the attributes X. This assumption is valid in many applications since many of the subjects are censored towards the end of the study. Let $S(t)$ be the probability that the event of interest has not occurred within the duration t. Using Kaplan-Meier estimator [27], the survival distribution is given by

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.9}$$

where d_i represents the number of events at time $t_{(i)}$ (time after ascending reordering), and n_i indicates the number of subjects who still remain in the study at time $t_{(i)}$. Thus, using Eq. (2.1) the probability of event can be estimated as $\hat{F}_e(t) = 1 - \hat{S}(t)$. On the other hand, the probability that censoring has not occurred within duration t can be defined as $G(t) = P(C > t)$ where C is censoring time, by setting "event" indicator $\delta_i^* = 1 - \delta_i$ [26]. Thus, Kaplan-Meier estimator for $G(t)$ is

$$\hat{G}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i^*}{n_i}\right) \tag{2.10}$$

where d_i^* is the number of subjects who were censored at time $t_{(i)}$, and n_i is the number of subjects at risk of censoring at time $t_{(i)}$. Let $\hat{F}_c(t)$ be the probability of censoring, then it can be estimated as $\hat{F}_c(t) = 1 - \hat{G}(t)$. We define a new label for censored data using Eq. (2.9) and (2.10). For each instance, if $\hat{F}_e(t) > \hat{F}_c(t)$, then it is labeled as *event*; otherwise, it will be labeled as *event-free* which indicates that even if there is complete followup information of that subject, there is extremely low chance of experiencing an event until the end of study (maybe even after that). Unlike other methods that handle censored data, this approach can simply solve the uncertainty with such censored data by labelling them as event or event-free based on the consistent Kaplan-Meier estimator. Even after the labeling is done, the problem of forecasting, explained in the next section, is a challenging task.

2.4 Early Stage Event Prediction (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 method on different distributions and then we will introduce ESP-NB, ESP-TAN and ESP-BN algorithms which utilize the extrapolation method.

2.4.1 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.* For this purpose, we develop the time to event estimation using the accelerated failure time model (AFT). We consider two well-known distributions, Weibull and Log-logistic, which are used widely in literature to model time-to-event [6] and the parameters of these distributions are learned from the information available until t_c. We will integrate such extrapolated values later with the proposed learning algorithms in order to make future predictions.

Weibull: When time-to-event follows Weibull distribution, the cumulative probability distribution $F(t_c)$ with shape a and scale b can be estimated as

$$\hat{F}(t_c) = 1 - e^{-(t_c/b)^a} \tag{2.11}$$

It should be noted that when the shape parameter of Weibull distribution is equal to 1, it transfers to the exponential distribution.

Log-logistic: when T_i follows log-logistic distribution with shape parameter a and scale parameter b, the prior probability distribution $F(t_c)$ can be estimated as

$$\hat{F}(t_c) = \frac{1}{1 + \left({}^{t_c}\!/_{b}\right)^{-a}} \tag{2.12}$$

Having the *cumulative probability distribution of event*, $F(t_c)$ at t_c, it can be easily extrapolated for any time t.

2.4.2 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 2.2.3, 2.2.4 and 2.2.5). We assume that the joint probability estimation from the Bayesian methods does not change over time. This is a valid assumption in survival data when that covariates do not depend on the time as the relation between feature at time $t_c$will still be the same through end of study [23]. On the other hand as time passes, the prior probability for event occurrence needs to be updated. 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:

ESP Naive Bayes (ESP-NB)

For Naive Bayes method using Eq. (2.2) and extrapolation method explained in previous section, the ESP-NB can be writen as

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big) = \frac{F(t_f) \prod_{i=1}^{m} P\big(x_i \mid y(t_c) = 1\big)}{P(\mathbf{x}, t \leq t_f)} \tag{2.13}$$

ESP Tree-Augmented Naive Bayes (ESP-TAN)

Probability of event occurrence based on TAN method for time t_f using Eq. (2.5) can be estimated as

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big) = \frac{F(t_f) \prod_{j=1}^{m} P\big(x_j \mid y(t_c) = 1, x_p(j)\big)}{P(\mathbf{x}, t \leq t_f)} \tag{2.14}$$

Algorithm 1: Early Stage Prediction (ESP) Framework

```
Require: Training data $D_n(t_c) = \big(\mathbf{x}, y(t_c), T\big)$, $t_f$
Output: Probability of event at time $t_f$
Phase 1: Conditional probability estimation at $t_c$
1: for $j = 1, ..., m$
2:     find $P\big(x_j \mid y(t_c) = 1\big)$
3: end
Phase 2: Predict probability of event occurrence at $t_f$
4: fit AFT model to $D_n(t_c)$
5: $P\big(y(t_f) = 1, t \leq t_f\big) = F(t)$
6: for $i = 1, ..., n$
7:     estimate $P\big(y_i(t_f) = 1 \mid \mathbf{x}_i, t \leq t_f\big)$
8: end
9: return $P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big)$
```

Algorithm 1 outlines the proposed ESP framework. In the first phase (lines 1-3), for each attribute j, the algorithm estimates conditional probability using the data available at time t_c. In the second phase, a probabilistic model is built to predict the event occurrence at t_f. In lines 4 and 5, the prior probability for event occurrence at time t_f is estimated using different extrapolation techniques. Then, in lines 6-9, for each subject i, we adapt the posterior probability of event occurrence at time t_f.

ESP Bayesian Network (ESP-BN)

For Bayesian Network, first we need to build a network using the information until t_c. We will train a Bayesian network classifier using Hill-climbing structure learning method.

Once we learn the structure of the Bayesian network, the subsequent step is to forecast the probability of event occurrence at the end of the study t_f. For this purpose we can use different extrapolation techniques as described in previous sections. Thus, the posterior probability estimation for event occurrence at time t_f can be defined as,

$$P\big(y(t_c) = 1 \mid \mathbf{x}, t \leq t_f\big) = \frac{F(t_f) \prod_{j=1}^{m} P\big(x_j \mid y(t_c) = 1, Pa(\mathbf{x_j})\big)}{P(\mathbf{x}, t \leq t_f)} \tag{2.15}$$

Algorithm 2: ESP-BN Algorithm:

```
Require: Training data $D_n(t_c)$, End of study time $t$.
Output: Probability of event at time $t_f$
Phase 1: learn Bayesian Network structure at $t_c$
1: $E_G \leftarrow \emptyset$, estimate $P\big(G|D_n(t_c)\big)$
2: $score_{final} \leftarrow \infty$ , $score = MDL\big(BN, D_n(t_c)\big)$ (Eq. (2.6))
3: while $score_{final} > score$
4:    $score_{final} \leftarrow score$
5:    for every add/remove/reverse $E_G$ on $G$
6:       estimate $P\big(G_{new}|D_n(t_c)\big)$
7:       $score_{new} = MDL\big(BN_{new}, D_n(t_c)\big)$
8:    select network structure with minimum $score_{new}$
9:    if $score > score_{new}$
10:        $score \leftarrow score_{new}$ , $G \leftarrow G_{new}$
Phase 2: Forecasting event occurrence at $t_f$
11: fit AFT model to $D_n(t_c)$
12: $P\big(y(t_f) = 1, t \leq t_f\big) = F(t)$
13: for all $i$ in $D_n(t)$
14:    estimate $P(\delta_i(t)|\mathbf{X}_i)$
15:        Weibull using Eqs. (2.8), (2.11) and (2.15)
16:        Log-logistic using Eqs. (2.8), (2.12) and (2.15)
17: return $P\big(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f\big)$
```

Algorithm 2 outlines the proposed ESP-BN model. Lines 1-10 describe the first stage where a Bayesian network structure is learnt using Hill-climbing method for training data until t_c. After the initial set up to build a network (lines 1-2), the Hill-climbing algorithm will find a network with the minimum MDL based on the score given in Eq. (2.6). In the second phase, a probabilistic model is built to forecast event occurrence at t. In line 11, the AFT model is built on $D_n(t_c)$ using various distributions. Then, in lines 13-17, we adapt the posterior probability of event occurrence at time t. This phase has the time complexity of $O(n)$. The time complexity of the ESP algorithm follows the time complexity of 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 it is $O(mn)$, for ESP-TAN it is $O(m^2n)$, where n is total number of subjects and m is the number of features in dataset and for ESP-BN $O(m^kn)$, where k is maximum number of parents (in our study we test different values of k to get the best performance which is in the range of 2 to 5) [45].

CHAPTER 3

EXPERIMENTAL RESULTS

In this chapter, we will implement our proposed ESP method on extensive dataset and and provide comparisons with various baseline prediction methods. First we explain real-world datasets as well as synthetic data that have been used in this thesis. We also discuss evaluation method that have been used to check the performance of the proposed method. Finally the experimental result will be provided and practical implications of the ESP framework in survival studies will be discussed.

3.1 Dataset Description

We evaluated the performance of the models using both synthetic and real-world benchmark survival datasets which are summarized in Table 3.1.

Synthetic Datasets: We generated synthetic dataset in which the feature vectors $\vec{x}$ 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, T can be generated using the method described in [3]. Given the observed covariates $\vec{x}_i$ for observation i, the failure time can be generated by

$$T_i = -\left(\frac{log(Unif(0,1))}{\lambda exp(\beta' \vec{x}_i)}\right)^{\nu} \tag{3.1}$$

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

Real-world Survival Datasets: There are several real-world survival benchmark datasets that we used in our experiments. Primary biliary cirrhosis (PBC), breast and colon cancer which are widely used in evaluating longitudinal studies are available in the survival data repository [1]. We also used Framingham heart study dataset which is publicly available [12].

In addition, we also used two proprietary datasets. One is the electronic health record

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

Table 3.1: 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 and C_{50} shows the number of censoring before T_{50}.

Dataset	#Features	#Instances	#Events	C_50	T_50	T_100
Syn1	5	100	50	5	1014	3808
Syn2	20	1000	602	87	943	7723
Breast	8	673	298	37	646	2659
Colon	13	888	445	8	394	3329
PBC	17	276	110	15	1191	4456
Framingham	16	5209	1990	0	1991	5029
EHR	77	4417	3479	0	50	4172
Kickstarter	54	4175	1961	162	21	60

(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 [2], 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.

3.2 Performance Evaluation

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

- ***Accuracy*** is expressed in the percentage of subjects in the test set that were classified correctly.

[2] www.kickspy.com

- ***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.

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

- ***AUC*** is the area under the receiver operating characteristic (ROC); the curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) by varying the threshold value.

In terms of our implementation, the joint probability for Naive Bayes and TAN is learnt using *e1071* package available in the R programming language [13]. Bayesian network structure for the proposed ESP-BN method is learned using a Hill-climbing algorithm that is available in open source Weka software [48], while the proposed model is implemented using the R programming language. The *coxph* function and *survreg* in the survival package are employed to train the Cox and AFT models, respectively. The Breslow's method was used to handle tied observations and the censored handling methods are also implemented in R using the survival package.

3.3 Results and Discussion

Tables 3.2, 3.3 and 3.4 provide the performance of different extrapolation methods using AUC, Accuracy and F-measure evaluation metrics. Models are trained at time when only 50% of events have occurred and the event forecasting is done at the end of study. For evaluation, we used stratified 10-fold cross-validation and average values (along with the standard deviations) of the results on all 10-folds. The result shows that Weibull distribution gives a better performance compare to log-logistic in most of survival data. This align with the time-to-event characteristic in survival data that fit perfectly with Weibull distribution. The choice of the particular distribution will depend on the nature of the dataset being considered, particularly the distribution that the event occurrence follows. However, our results indicate that for almost all of the datasets, Weibull distribution will provide much better results.

Table 3.2: Comparison of AUC different extrapolation methods used in ESP-NB, ESP-TAN and ESP-BN (with standard deviation values).

	ESP-NB		**ESP-TAN**		**ESP-BN**	
Dataset	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**
Syn1	**0.865** **(0.004)**	0.841 (0.003)	**0.869** **(0.001)**	0.849 (0.001)	**0.867** **(0.002)**	0.843 (0.002)
Syn2	**0.823** **(0.002)**	0.812 (0.003)	**0.825** **(0.003)**	0.821 (0.002)	**0.833** **(0.001)**	0.822 (0.002)
Breast	**0.669** **(0.001)**	0.643 (0.003)	**0.678** **(0.007)**	0.653 (0.005)	**0.673** **(0.001)**	0.649 (0.003)
Colon	**0.639** **(0.013)**	0.622 (0.014)	**0.642** **(0.009)**	0.631 (0.011)	**0.659** **(0.009)**	0.644 (0.01)
PBC	**0.767** **(0.001)**	0.744 (0.004)	**0.772** **(0.003)**	0.758 (0.001)	**0.786** **(0.003)**	0.775 (0.001)
Framingham	0.954 (0.007)	**0.971** **(0.003)**	0.969 (0.004)	**0.973** **(0.002)**	0.964 (0.003)	**0.979** **(0.001)**
EHR	**0.656** **(0.018)**	0.628 (0.021)	**0.657** **(0.011)**	0.63 (0.026)	**0.667** **(0.012)**	0.664 (0.018)
Kickstarter	0.822 (0.024)	**0.829** **(0.023)**	0.827 (0.019)	**0.833** **(0.018)**	0.845 (0.023)	**0.847** **(0.021)**

For performance benchmarking, we compare the proposed ESP-NB, ESP-TAN and ESP-BN algorithms using the best distributions from previous tables as extrapolation techniques with Cox, Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Tree-Augmented Naive Bayes (TAN) and Bayesian Network (BN)classification methods which trained at time when only 50% of events have occurred and the event prediction is done at the end of study. Tables 3.5, 3.6 and 3.7 summarize the comparison result for AUC, Accuracy and F-measure evaluation metrics. For all of the datasets, our results evidently show that the proposed ESP-based methods will provide significantly better prediction results compared to other methods.

The results show that by incorporating the time-to-event extrapolation method within the ESP framework, we are able to adapt the prior probability in Bayesian methods com-

Table 3.3: Comparison of Accuracy for different extrapolation methods used in ESP-NB, ESP-TAN and ESP-BN (with standard deviation values).

	ESP-NB		ESP-TAN		ESP-BN	
Dataset	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**
Syn1	**0.779** **(0.023)**	0.771 (0.017)	**0.792** **(0.02)**	0.782 (0.024)	**0.787** **(0.019)**	0.785 (0.021)
Syn2	**0.777** **(0.023)**	0.771 (0.029)	**0.785** **(0.025)**	0.779 (0.027)	**0.789** **(0.021)**	0.782 (0.023)
Breast	**0.738** **(0.027)**	0.725 (0.022)	**0.805** **(0.022)**	0.738 (0.027)	**0.754** **(0.019)**	0.791 (0.015)
Colon	**0.615** **(0.155)**	0.611 (0.141)	**0.619** **(0.148)**	0.614 (0.165)	**0.622** **(0.12)**	0.617 (0.145)
PBC	**0.719** **(0.116)**	0.705 (0.119)	**0.731** **(0.118)**	0.714 (0.114)	**0.748** **(0.11)**	0.732 (0.101)
Framingham	0.827 (0.093)	**0.859** **(0.103)**	0.853 (0.089)	**0.865** **(0.096)**	0.879 (0.106)	**0.892** **(0.096)**
EHR	**0.771** **(0.126)**	0.745 (0.119)	**0.785** **(0.156)**	0.764 (0.123)	**0.815** **(0.112)**	0.789 (0.116)
Kickstarter	0.739 (0.043)	**0.756** **(0.059)**	0.745 (0.048)	**0.769** **(0.042)**	0.767 (0.048)	**0.785** **(0.052)**

Table 3.4: Comparison of F-measure for different extrapolation methods used in ESP-NB, ESP-TAN and ESP-BN (with standard deviation values).

	ESP-NB		ESP-TAN		ESP-BN	
Dataset	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**	**Weibull**	**Log-Logistic**
Syn1	**0.776** **(0.022)**	0.778 (0.022)	**0.789** **(0.019)**	0.783 (0.023)	**0.785** **(0.017)**	0.783 (0.02)
Syn2	**0.774** **(0.023)**	0.769 (0.029)	**0.779** **(0.02)**	0.769 (0.021)	**0.783** **(0.026)**	0.776 (0.021)
Breast	**0.749** **(0.036)**	0.721 (0.042)	**0.796** **(0.032)**	0.748 (0.039)	**0.761** **(0.042)**	0.743 (0.038)
Colon	**0.621** **(0.145)**	0.611 (0.151)	**0.626** **(0.148)**	0.617 (0.15)	**0.629** **(0.18)**	0.622 (0.15)
PBC	**0.712** **(0.11)**	0.687 (0.109)	**0.715** **(0.099)**	0.698 (0.114)	**0.725** **(0.098)**	0.721 (0.11)
Framingham	0.875 (0.073)	**0.883** **(0.083)**	0.894 (0.059)	**0.908** **(0.066)**	0.902 (0.076)	**0.925** **(0.066)**
EHR	**0.787** **(0.126)**	0.765 (0.206)	**0.798** **(0.16)**	0.804 (0.14)	**0.826** **(0.16)**	0.811 (0.12)
Kickstarter	0.753 (0.037)	**0.758** **(0.053)**	0.765 (0.048)	**0.779** **(0.032)**	0.782 (0.058)	**0.797** **(0.042)**

Table 3.5: Comparison of AUC values for Cox, LR, RF, NB, TAN and BN with proposed ESP-NB, ESP-TAN and ESP-BN methods using best method of extrapolation methods (with standard deviation values).

Dataset	**Cox**	**LR**	**RF**	**NB**	**TAN**	**BN**	**ESP-NB**	**ESP-TAN**	**ESP-BN**
Syn1	0.717 (0.004)	0.725 (0.005)	0.712 (0.006)	0.715 (0.007)	0.722 (0.002)	0.718 (0.005)	0.865 (0.004)	**0.869** **(0.001)**	0.867 (0.002)
Syn2	0.71 (0.004)	0.729 (0.004)	0.714 (0.002)	0.713 (0.007)	0.718 (0.005)	0.721 (0.006)	0.823 (0.002)	0.825 (0.003)	**0.833** **(0.001)**
Breast	0.619 (0.01)	0.658 (0.007)	0.647 (0.004)	0.629 (0.009)	0.662 (0.004)	0.635 (0.002)	0.669 (0.001)	**0.678** **(0.007)**	0.673 (0.001)
Colon	0.61 (0.024)	0.618 (0.011)	0.621 (0.014)	0.627 (0.011)	0.629 (0.014)	0.633 (0.01)	0.639 (0.013)	0.642 (0.009)	**0.659** **(0.009)**
PBC	0.698 (0.009)	0.665 (0.005)	0.72 (0.003)	0.687 (0.003)	0.693 (0.01)	0.731 (0.004)	0.767 (0.001)	0.772 (0.003)	**0.786** **(0.003)**
Framingham	0.879 (0.007)	0.935 (0.002)	0.929 (0.005)	0.957 (0.002)	0.963 (0.005)	0.969 (0.004)	0.971 (0.007)	0.973 (0.004)	**0.979** **(0.001)**
EHR	0.616 (0.023)	0.637 (0.017)	0.65 (0.025)	0.642 (0.019)	0.645 (0.025)	0.651 (0.026)	0.656 (0.018)	0.657 (0.011)	**0.667** **(0.012)**
Kickstarter	0.823 (0.019)	0.842 (0.019)	0.845 (0.027)	0.815 (0.022)	0.819 (0.025)	0.844 (0.023)	0.822 (0.024)	0.827 (0.019)	**0.847** **(0.021)**

Table 3.6: Comparison of Accuracy values for Cox, LR, RF, NB, TAN and BN with proposed ESP-NB, ESP-TAN and ESP-BN methods using best method of extrapolation methods (with standard deviation values).

Dataset	**Cox**	**LR**	**RF**	**NB**	**TAN**	**BN**	**ESP-NB**	**ESP-TAN**	**ESP-BN**
Syn1	0.658 (0.022)	0.649 (0.024)	0.675 (0.019	0.642 (0.018)	0.681 (0.021)	0.673 (0.022)	0.779 (0.023)	**0.792** **(0.02)**	0.787 (0.019)
Syn2	0.657 (0.021)	0.609 (0.026)	0.669 (0.025)	0.665 (0.027)	0.673 (0.029)	0.677 (0.024)	0.777 (0.023)	0.785 (0.025)	**0.789** **(0.021)**
Breast	0.632 (0.017)	0.557 (0.013)	0.622 (0.016)	0.613 (0.023)	0.657 (0.014)	0.628 (0.021)	0.738 (0.027)	**0.805** **(0.022)**	0.754 (0.019)
Colon	0.49 (0.133)	0.487 (0.167)	0.562 (0.18)	0.526 (0.159)	0.531 (0.174)	0.552 (0.15)	0.615 (0.155)	0.619 (0.148)	**0.622** **(0.12)**
PBC	0.657 (0.111)	0.578 (0.123)	0.658 (0.132)	0.599 (0.125)	0.638 (0.115)	0.633 (0.119)	0.719 (0.116)	0.731 (0.118)	**0.748** **(0.11)**
Framingham	0.745 (0.085)	0.77 (0.093)	0.732 (0.085)	0.761 (0.099)	0.782 (0.107)	0.804 (0.087)	0.827 (0.093)	0.853 (0.089)	**0.892** **(0.096)**
EHR	0.651 (0.121)	0.586 (0.132)	0.619 (0.173)	0.642 (0.156)	0.659 (0.182)	0.691 (0.191)	0.771 (0.126)	0.785 (0.156)	**0.815** **(0.112)**
Kickstarter	0.656 (0.049)	0.698 (0.039)	0.709 (0.052)	0.691 (0.068)	0.736 (0.051)	0.746 (0.046)	0.739 (0.043)	0.745 (0.048)	**0.785** **(0.052)**

Table 3.7: Comparison of F-measure values for Cox, LR, RF, NB, TAN and BN with proposed ESP-NB, ESP-TAN and ESP-BN methods using best method of extrapolation methods (with standard deviation values).

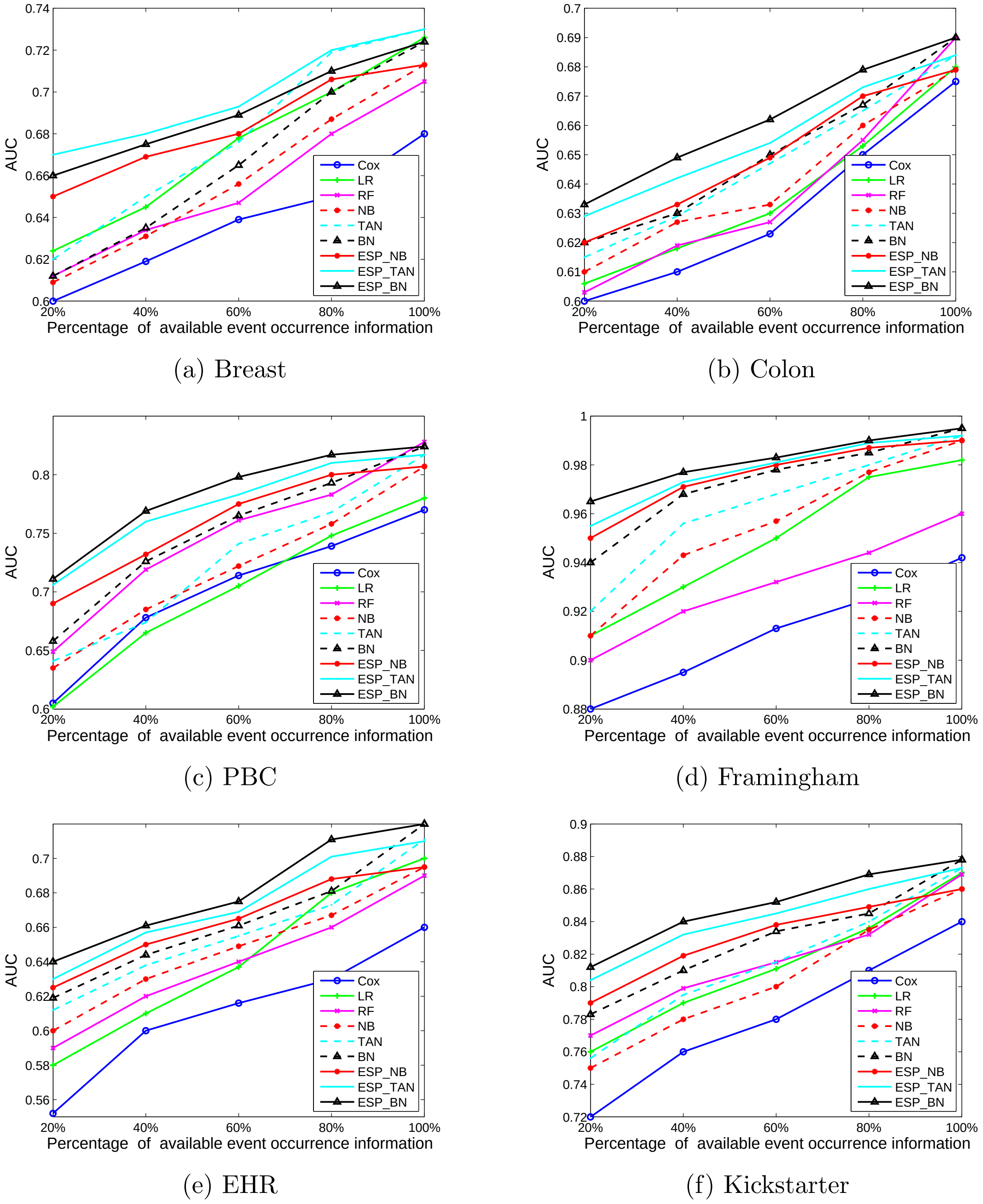
Dataset	**Cox**	**LR**	**RF**	**NB**	**TAN**	**BN**	**ESP-NB**	**ESP-TAN**	**ESP-BN**
Syn1	0.651 (0.021)	0.645 (0.025)	0.667 (0.022)	0.762 (0.021)	0.778 (0.023)	0.773 (0.021)	0.776 (0.022)	**0.789** **(0.019)**	0.785 (0.017)
Syn2	0.647 (0.023)	0.599 (0.025)	0.659 (0.027)	0.655 (0.029)	0.663 (0.024)	0.671 (0.023)	0.774 (0.023)	0.779 (0.02)	**0.783** **(0.026)**
Breast	0.648 (0.035)	0.573 (0.063)	0.642 (0.033)	0.623 (0.053)	0.672 (0.034)	0.638 (0.031)	0.749 (0.036)	**0.796** **(0.032)**	0.761 (0.042)
Colon	0.512 (0.161)	0.487 (0.17)	0.578 (0.194)	0.543 (0.169)	0.549 (0.184)	0.562 (0.19)	0.621 (0.145)	0.626 (0.148)	**0.629** **(0.18)**
PBC	0.61 (0.141)	0.529 (0.13)	0.613 (0.12)	0.541 (0.121)	0.562 (0.15)	0.575 (0.14)	0.712 (0.11)	0.715 (0.099)	**0.725** **(0.098)**
Framingham	0.769 (0.078)	0.735 (0.093)	0.792 (0.085)	0.794 (0.075)	0.809 (0.073)	0.845 (0.083)	0.875 (0.073)	0.894 (0.059)	**0.925** **(0.066)**
EHR	0.681 (0.11)	0.584 (0.166)	0.617 (0.188)	0.684 (0.156)	0.708 (0.198)	0.715 (0.21)	0.787 (0.126)	0.798 (0.16)	**0.826** **(0.16)**
Kickstarter	0.689 (0.084)	0.711 (0.048)	0.737 (0.067)	0.721 (0.058)	0.726 (0.061)	0.743 (0.054)	0.753 (0.037)	0.765 (0.048)	**0.797** **(0.042)**

putations. Thus, it clearly indicates that the ESP-based method outperforms the other methods in building an accurate forecasting model. Furthermore, ESP-NB build on independence assumption between attributes which does not hold in many clinical survival applications. Thus, the introduced ESP-TAN and ESP-BN weakened this assumption which leads to increase in AUC, accuracy and F-measure in almost all of results. Also in almost all the cases ESP-BN gives the better results. This is due to the fact that Bayesian netwrok can model more complex data specially when we have more features compare to TAN however it has higher time complexity [8].

Comparing the result in Table 3.5 with Tables 3.6 or 3.7 one can conclude that improvement in the accuracy and F-measure is more significant than improvement in AUC. The reason is that the area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. It measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class, but without actually assigning patterns to classes. In our method we adapt the prior probability using time-to-event information. This update the probability of event for all subjects in the study with some rate. Thus, the overall ranking for probability of event occurrence after using ESP framework will change slightly compared to baseline classifiers. On the other hand, the overall accuracy also depends on the ability of the classifier to rank patterns, but also on its ability to select a threshold in the ranking used to assign patterns to the positive and negative class. Using the same threshold, ESP method result in better confusion matrix which cause both accuracy and F-measure change significantly compare to other methods that do not have the ability to extrapolate event occurrence.

This result supports our claim that probabilistic models can provide an accurate forecasting of event occurrences beyond the observation time. From our experiments, we can conclude that our model can 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. The proposed prediction model 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.

In Figures 3.1, 3.2 and 3.3, we present the prediction performance of different methods by varying the percentage of event occurrence information that is available to train the model for all real-world datasets. For example, 20% on the x-axis corresponds to the training data obtained when only 20% of events have occurred and prediction of the event occurrences was made for the end of study period. From this plot we can see that the evaluation metric values improve when there is more information on the event occurrence in the training data. For all the cases, our proposed ESP-based method gives the better prediction performance compared to other techniques. This behaviour is similar across all the benchmark datasets. Furthermore, it should be noted that the improvements of the proposed methods are more significant over the baseline methods



(a) Breast

(b) Colon

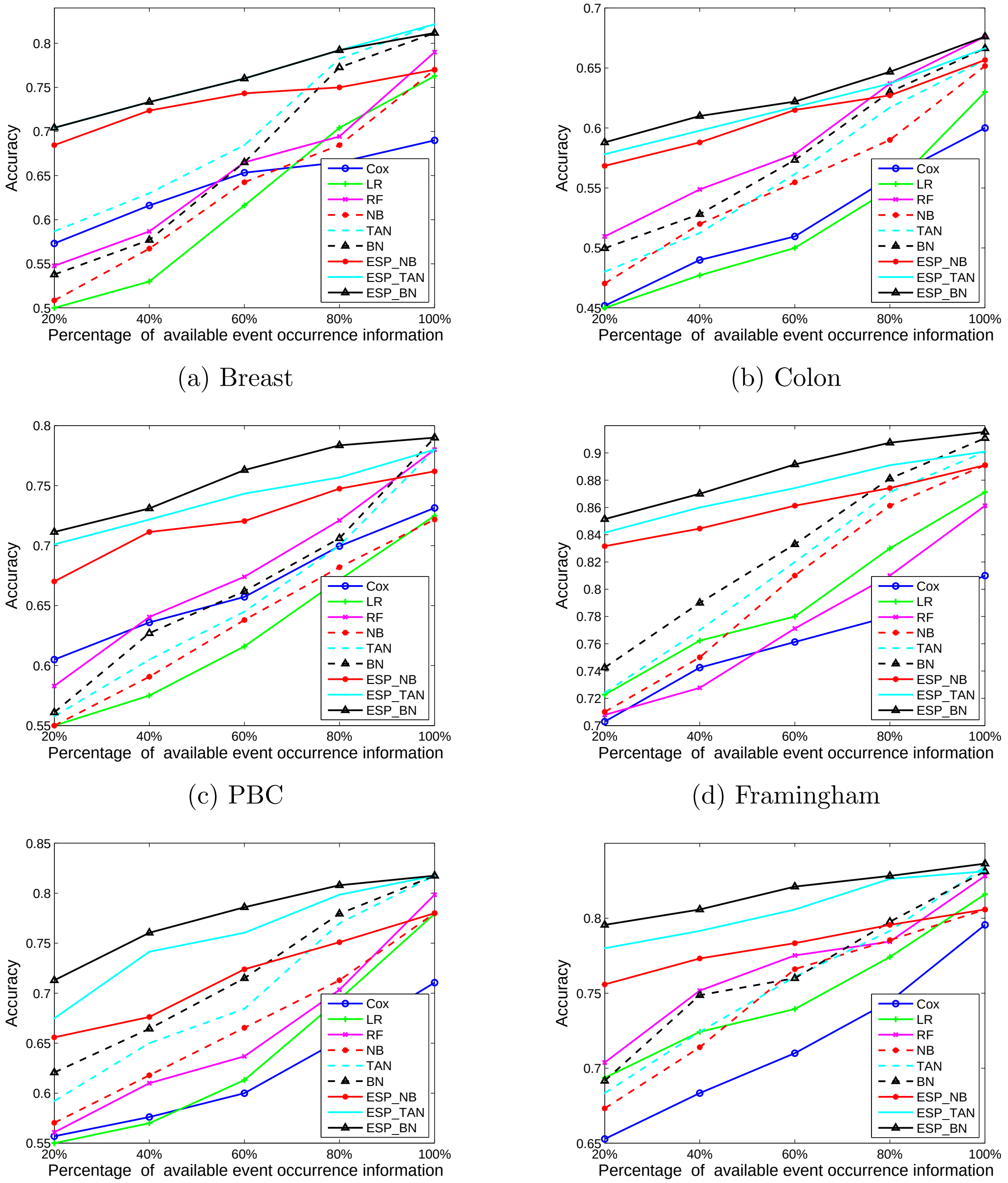
(c) PBC

(d) Framingham

(e) EHR

(f) Kickstarter

Figure 3.1: AUC values of different methods obtained by varying the percentage of event occurrence information.



(a) Breast

(b) Colon

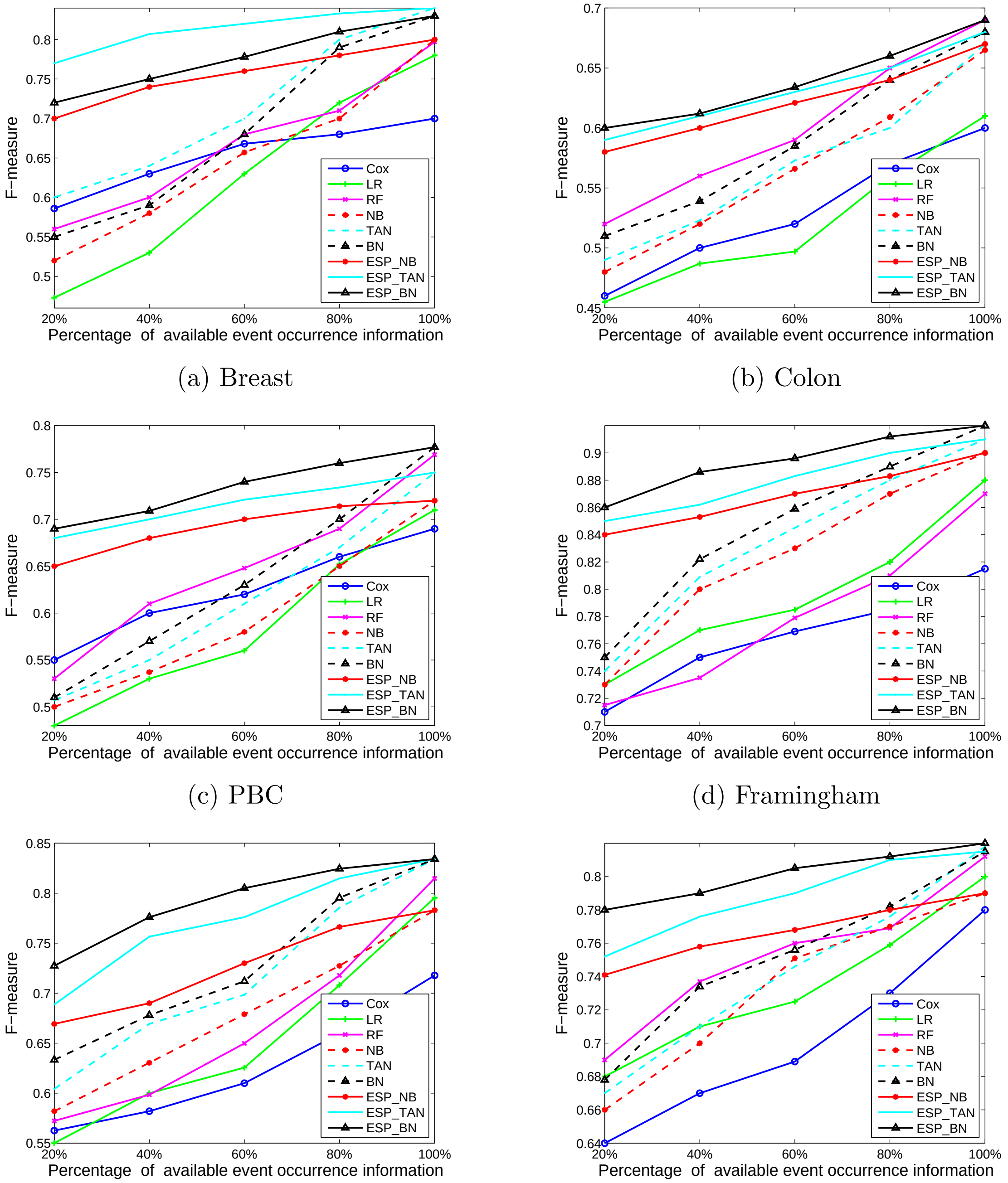
(c) PBC

(d) Framingham

(e) EHR

(f) Kickstarter

Figure 3.2: Accuracy values of different methods obtained by varying the percentage of event occurrence information for the pbc dataset.



(a) Breast

(b) Colon

(c) PBC

(d) Framingham

(e) EHR

(f) Kickstarter

Figure 3.3: F-measure values of different methods obtained by varying the percentage of event occurrence information for the pbc dataset.

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 Bayesian Network method 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.

CHAPTER 4

CONCLUSION AND FUTURE WORK

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 of longitudinal studies. In this thesis, we developed an early stage event prediction framework by extending Bayesian methods through fitting a statistical distribution to time-to-event data with fewer available events at the early stages. Instead of excluding the censored data, we develop a new mechanism to handle censored data by estimating the probability of event and the probability of censoring using Kaplan-Meier estimator. One of the main objectives of this paper is to demonstrate that more accurate predictions can be made when the prior probability at end of study time is appropriately estimated using the current information of event occurrence. This is extremely important in such longitudinal survival studies since accumulating enough training data about the event occurrence is a time-consuming process.

The proposed ESP-based model adapts prior probability of event occurrence by fitting time-to-event information using Weibull and Log-logistic distributions. 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 or other classification methods in forecasting events at future time points. 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.

REFERENCES

[1] Bandyopadhyay, S., Wolfson, e.: 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] Bellaachia, A., Guven, E., Dc, W.: Predicting Breast Cancer Survivability Using Data Mining Techniques. In: Society for Industrial and Applied Mathematics (SIAM) (2006)

[3] Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate Cox proportional hazards models by Ralf Bender, Thomas Augustin and Maria Blettner, Statistics in Medicine 2005; 24:1713-1723. Statistics in medicine **25**, 1978–1979 (2006). DOI 10.1002/sim.

[4] Bennett, K.P., Demiriz, A.: Semi-supervised support vector machines. In: Advances in Neural Information Processing Systems, pp. 368–374. MIT Press (1998)

[5] Buckley, J., James, I.: Linear Regression with Censored Data. Biometrics Trust **66**(3), 429–436 (1979)

[6] Carroll, K.J.: On the use and utility of the weibull model in the analysis of survival data. Controlled clinical trials **24**(6), 682–701 (2003)

[7] Chapelle, O., Schölkopf, B., Zien, A., et al.: Semi-supervised learning, vol. 2. MIT press Cambridge (2006)

[8] Cheng, J., Greiner, R.: Comparing bayesian network classifiers. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 101–108. Morgan Kaufmann Publishers Inc. (1999)

[9] Chi, C.l., Street, W.N., Wolberg, W.H.: Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. In: AMIA Annual Symposium, pp. 130–134 (2007)

[10] Cordon-cardo, C., Kotsianti, A., Verbel, D.A., et. al.: Improved prediction of

prostate cancer recurrence through systems pathology. Journal of clinical investigation **117**(7), 1876–1883 (2007). DOI 10.1172/JCI31399DS1

[11] Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society **34**(2), 187–220 (1972)

[12] Dawber, T.R., Kannel, W.B., Lyell, L.P.: An approach to longitudinal studies in a community: the framingham study. Annals of the New York Academy of Sciences **107**(2), 539–556 (1963)

[13] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., Leisch, M.F.: Package e1071. R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html (2009)

[14] Donovan, M.J., Donovan, M.J., Hamann, S., Clayton, M., 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 (2008)

[15] Evers, L., Messow, C.M.: Sparse kernel methods for high-dimensional survival data. Bioinformatics (Oxford, England) **24**(14), 1632–8 (2008)

[16] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine learning **29**(2-3), 131–163 (1997)

[17] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis, vol. 2. Taylor & Francis (2014)

[18] Gordon, L., Plshen, R.: Tree-structured survival analysis. Cancer Treat Reports **69**(10), 1065–1074 (1985)

[19] Heckerman, D.: A tutorial on learning with Bayesian networks. Springer (1998)

[20] Heckerman, D., Geiger, D., Chickering, D.M.: Learning bayesian networks: The combination of knowledge and statistical data. Machine learning **20**(3), 197–243 (1995)

[21] Hosmer, D.W., Lemeshow, S.: Applied survival analysis: regression modeling of time to event data. Wiley, New York (1999)

[22] Jahanbani Fard, M., Ameri, S., Zeinal Hamadani, A.: Bayesian approach for early stage reliability prediction of evolutionary products. In: Proceedings of the International Conference on Operations Excellence and Service Engineering (2015)

[23] Jiang, X., Xue, D., Brufsky, A., Khan, S., Neapolitan, R.: A new method for predicting patient survivorship using efficient bayesian network learning. Cancer informatics **13**, 47 (2014)

[24] John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (1995)

[25] Jordan, A.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems **14**, 841 (2002)

[26] Kalbfleisch, J.D., Prentice, R.L.: The statistical analysis of failure time data. John Wiley & Sons (2002)

[27] Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)

[28] Khan, F.M., Zubek, V.B.: Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. 2008 Eighth IEEE International Conference on Data Mining pp. 863–868 (2008)

[29] Koul, H., Susarla, V., Van Ryzin, J.: Regression Analysis with Randomly Righe-Censored. The Annals of Statistics **9**(6), 1276–1288 (1981)

[30] Lam, W., Bacchus, F.: Learning bayesian belief networks: An approach based on the mdl principle. Computational intelligence **10**(3), 269–293 (1994)

[31] Lasserre, J., Bishop, C.M., Minka, T.P., et al.: Principled hybrids of generative and

discriminative models. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1, pp. 87–94. IEEE (2006)

[32] Lavrac, N.: Selected techniques for data mining in medicine. Artificial Intelligence in Medicine **16**, 3–23 (1999)

[33] Lee, E.T., Wang, J.: Statistical methods for survival data analysis, vol. 476. John Wiley & Sons (2003)

[34] Lisboa, P.J., Wong, H., Harris, P., Swindell, R.: A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artificial intelligence in medicine **28**(1), 1–25 (2003)

[35] Lucas, P.J.F., van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial intelligence in medicine **30**(3), 201–14 (2004)

[36] Miller, R.G.: Least squares Regression with Censored Data. Biometrics Trust **63**(3), 449–464 (1976)

[37] Miller, R.G., Halpern, J.: Regression with Censored Data. Biometrika Trust **69**(3), 521–531 (1982)

[38] Reddy, C.K., Li, Y.: A review of clinical prediction models. In: C.K. Reddy, C.C. Aggarwal (eds.) Healthcare Data Analytics. Chapman and Hall/CRC Press (2015)

[39] Segal, M.R.: Regression Trees for Censored Data. Biometrics **44**(1), 35–47 (1988)

[40] Shiao, H.T., Cherkassky, V.: Learning using privileged information (LUPI) for modeling survival data. 2014 International Joint Conference on Neural Networks (IJCNN) pp. 1042–1049 (2014)

[41] Shim, J., Hwang, C.: Support vector censored quantile regression under random censoring. Computational Statistics & Data Analysis **53**(4), 912–919 (2009)

[42] Shivaswamy, P.K., Chu, W., Jansche, M.: A Support Vector Approach to Censored Targets. Seventh IEEE International Conference on Data Mining (ICDM 2007) pp. 655–660 (2007)

[43] Štajduhar, I., Dalbelo-Bašić, B.: Learning bayesian networks from survival data using weighting censored instances. Journal of biomedical informatics **43**(4), 613–622 (2010)

[44] Štajduhar, I., Dalbelo-Bašić, B.: Uncensoring censored data for machine learning: A likelihood-based approach. Expert Systems with Applications **39**(8), 7226–7234 (2012)

[45] Su, J., Zhang, H.: Full bayesian network classifiers. In: Proceedings of the 23rd international conference on Machine learning, pp. 897–904. ACM (2006)

[46] Van Belle, V., Pelckmans, K., Van Huffel, S., Suykens, J.a.K.: Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial intelligence in medicine **53**(2), 107–18 (2011)

[47] Wei, L.: The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. Statistics in medicine **11**(14-15), 1871–1879 (1992)

[48] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)

[49] Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D.M., Musgrove, D., Adomavicius, G., Johnson, P.E., O'Connor, P.J.: A naive bayes machine learning approach to risk prediction using censored, time-to-event data. Statistics in medicine pp. 105–112 (2015)

[50] Zhou, Z.H., Li, M.: Semi-supervised regression with co-training. In: IJCAI, pp. 908–916 (2005)

[51] Zupan, B., DemšAr, J., Kattan, M.W., Beck, J.R., Bratko, I.: Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial intelligence in medicine **20**(1), 59–75 (2000)

ABSTRACT

BAYESIAN APPROACH FOR EARLY STAGE EVENT PREDICTION IN SURVIVAL DATA

by

MAHTAB JAHANBANI FARD

December 2015

Advisor: Dr. Chandan Reddy

Major: Computer Scienc

Degree: Master of Science

Predicting event occurrence at an early stage in longitudinal studies is an important and challenging 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. On the other hand, survival analysis aims at finding the underlying distribution for data that measure the length of time until the occurrence of an event. However, it cannot give an answer to the open question of *"how to forecast whether a subject will experience event by end of study having event occurrence information at early stage of survival data?"*. This problem exhibits two major challenges: 1) absence of complete information about event occurrence (censoring) and 2) availability of only a partial set of events that occurred during the initial phase of the study. Thus, the main objective of this work is to predict for which subject in the study event will occur at future based on few event information at the initial stages of a longitudinal study.

In this thesis, we propose a novel approach to address the first challenge by introducing a new method for handling censored data using Kaplan-Meier estimator. The second challenge is tackled by effectively integrating Bayesian methods with an Accelerated

Failure Time (AFT) model by adapting the prior probability of the event occurrence for future time points. In another word, 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 extended the Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network methods based on the proposed framework, and developed three algorithms, namely, ESP-NB, ESP-TAN and ESP-BN, to effectively predict event occurrence using the training data obtained at early stage of the study. 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 prediction methods.

AUTOBIOGRAPHICAL STATEMENT

Mahtab Jahanbani Fard

Mahtab Jahanbani Fard is a Ph.D. candidate at Industrial & System Engineering and M.Sc. student of Computer Science at Wayne State University. She received her bachelor's of science in Physics and M.Sc. in Industrial Engineering from Isfahan University of Technology, Iran, in 2007 and 2009. Her main areas of research are data mining, machine learning, survival analysis and healthcare.