This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

A Bayesian Perspective on Early Stage Event Prediction in Longitudinal Data

Mahtab J. Fard, Ping Wang, Sanjay Chawla, and Chandan K. Reddy, *Senior Member, IEEE*

Abstract—Predicting event occurrence at the early stage of a longitudinal study is an important and challenging problem which has high practical value in many real-world applications. As opposed to the standard classification and regression problems where a domain expert can provide 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 a sufficient number of events. Survival analysis aims at directly predicting the time to an event of interest using the data collected in the past for a certain duration. However, it cannot give an answer to the open question of *"how to forecast whether a subject will experience an event by end of a longitudinal study using event occurrence information of other subjects at the early stage of the study?"*. The goal of this work is to predict the event occurrence at a future time point using only the information about a limited number of events that occurred at the initial stages of a longitudinal study. 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. We propose a novel Early Stage Prediction (ESP) framework for building event prediction models which are trained at the early stages of longitudinal studies. First, we develop a novel approach to address the first challenge by introducing a new method for handling censored data using Kaplan-Meier estimator. We then extend the Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network methods based on the proposed framework, and develop three algorithms, namely, ESP-NB, ESP-TAN and ESP-BN, to effectively predict event occurrence using training data obtained at an early stage of the study. More specifically, our approach effectively integrates Bayesian methods with an Accelerated Failure Time (AFT) model by adapting the prior probability of the event occurrence for future time points. 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, on an average, 20% more accurate compared to existing schemes when using only limited event information in the training data.

Keywords: Bayesian network, naive Bayes, longitudinal data, survival analysis, early stage prediction, regression, event data.

1 Introduction

It has become a common practice in many application domains to collect data over a period of time and record the occurrence of events of interest within a given period. These studies are usually called longitudinal studies, in which the subjects are followed over time for monitoring certain risks. Developing effective prediction models to estimate the outcome of a particular event of interest is a critical challenge in longitudinal studies. Such studies are ubiquitous in various real-world domains, such as healthcare, reliability, engineering, etc [1], [2], [3] and their primary goal is to build models that can accurately determine the probability of occurrence of a particular event of interest at a specific time point [4]. One of the primary challenges in these longitudinal studies is that, as opposed to the standard supervised learning problems where a domain expert can provide labels within a reasonable amount of time, training data in such tasks must be obtained only by waiting for the occurrence of a 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 problem. In addition, occurrence of the event is not necessarily observed for all the instances in the study and hence the outcome variable might be incomplete. This phenomenon is also known as '*censoring*'. Building event forecasting models in the presence of censored data is a challenging task which has a significant practical value in longitudinal studies. The main goal of this work is to answer the following open question: *"how to forecast whether a subject will experience an event by the end of a longitudinal study using event occurrence information at early stages of the study?"*. 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.

Let us consider the following real-world applications which motivate the early stage time-to-event prediction.

- In the healthcare domain, when there is a new treatment option (or drug) that is available, 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 due to treatment failure. The effectiveness of this treatment must be estimated as early as possible when there are only a few hospitalized patients [5].
- In education, early identification of students at the risk

- *M. J. Fard is with the Department of Computer Science and the Department of Industrial Engineering at Wayne State University, Detroit, MI. Email: fard@wayne.edu.*
- *P. Wang is with the Department of Computer Science at Virginia Tech, Arlington, VA. Email: ping@vt.edu.*
- *S. Chawla is with the Qatar Computing Research Institute, HBKU and the University of Sydney, Australia. Email: sanjay.chawla@sydney.edu.au.*
- *C.K. Reddy is with the Department of Computer Science at Virginia Tech, Arlington, VA. Email: reddy@cs.vt.edu.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

of dropping out of their school at the beginning of their study is crucial for improving the graduation rates. The ability to build an accurate prediction model using only the early stage data can be practically very useful [6].

- Reliability prediction focuses on developing 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 about future failures can be given.
- In credit score modeling applications, the goal is to accurately estimate whether a customer will default or not and if they default, when the default is going to happen? If a model can accurately predict using only a few default cases, 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 event predictions using training data that contains only a few events (i.e., at an early stage of a longitudinal study). More precisely, the goal here is to predict the event occurrence for a time period beyond the observation time window (when there are only a few events that have occurred in the dataset). Thus, this paper aims 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. In this example, a longitudinal study is conducted on six subjects and the information for event occurrence until time t_c is recorded, where only subjects S2 and S5 have experienced the event. The goal of our work is to predict the event occurrence by time t_f (e.g. the end of study). In other words, during the training phase, the event occurrences until the observation time t_c are the only ones available and the objective is to make predictions about the event occurrences by the end of study t_f. It should be noted that except subjects S2 and S5, all others are considered to be censored at t_c (marked by 'X'). However, an event will occur for subjects S1 and S6 within the time period t_f.

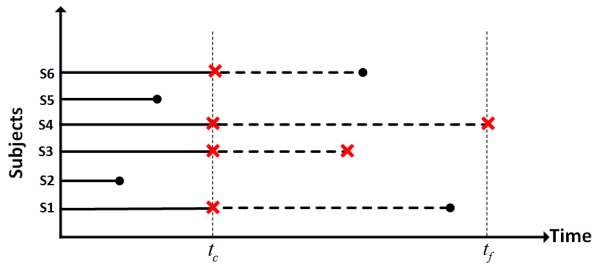


Fig. 1: An illustration to demonstrate the problem of event forecasting at time t_f (e.g. end of study) using the information available only until time t_c.

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 is an important problem in the domain of longitudinal studies since the only way to collect reliable data here is to wait for sufficient period of time until the complete information about event occurrence is acquired. In this paper, we will introduce a new method for handling censored data using Kaplan-Meier estimator. We will then develop a novel Early Stage Prediction (ESP) framework for building event prediction models which are trained at early stages of longitudinal studies. More specifically, we propose a framework based on Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network, and develop 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.

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, existing statistical techniques, especially in the field of survival analysis, do not have the ability to handle the problem of predicting event occurrence in the early stage prediction problem. The main reason is that the training and testing data are collected for the same time window in survival models, and the probability of event predictions given by any survival model is valid only for the specific observed time. The goal of this work, on the other hand, is to build model at the early stage of the study, and predict the event occurrence for the new subjects collected in the future time point. In other words, the "future" in our early stage prediction problem is different from that in the regular survival analysis methods. 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 learn from such data. There are two naive ways to handle this problem: ignoring this censored data and treating censoring time as the actual time of event occurrence. However, these methods may provide a suboptimal model because of neglecting the available information or may provide an underestimate of the true performance of the model.

To solve the problem discussed above, we introduce an intuitive method to handle the censoring problem in the longitudinal survival data. We then develop a Bayesian framework for early stage event prediction to tackle the problem of insufficient amount of training data on event occurrence in the initial phases (early stage) of longitudinal studies. More specifically, we are combining the power of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

Bayesian method(s) with the concept of parametric survival analysis to produce a solution that can be effective when there are only few events that have occurred. Thus, the main contributions of this paper 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 probability of event occurrence for a future time point using various extrapolation techniques.
- Develop probabilistic algorithms based on Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network, (we call them 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.

The rest of the paper is organized as follows. In Section 2, a brief review of the related literature is provided. Section 3 introduces the notations and definitions that are necessary to comprehend our proposed algorithms. We also propose a new method to handle the censored data in this section. The proposed Bayesian approach for early stage event prediction on survival data is described in Section 4. Section 5 demonstrates the experimental results and shows the practical significance of our work using various real-world datasets. Finally, Section 6 concludes the discussion.

2 Related work

Before we discuss the early stage prediction framework in detail, the related work in the areas of using machine learning techniques for survival analysis will be briefly presented in this section.

Survival analysis is a subfield of statistics where a wide range of techniques have been proposed to model time-to-event data [7] in which the dependent variable is subject to censoring (e.g. failure, death, admission to hospital, emergence of disease, etc.) [8]. The ordinary Least-Squares, the most common method for solving regression problems, is based on minimizing sum of squared errors. It 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 obtained from the regression model [9]. Although it is challenging to know the relative rank of the event occurrences of the censored instances, the well-known likelihood method has the ability to solve the censored regression problem [10]. Different techniques have been proposed based on Maximum Likelihood Estimation (MLE) to overcome the difficulty of handling censored data [11], [12].

Similar to survival data which captures time to events of interest, time series methods deal with slightly different kind of time-centered analysis [13], [14]. Time series analysis tackles the problem of studying experimental data that have been observed at different points of time [15]. Recently, there are some efforts to address the problem of early classification in time-series data [16], [17]. Although time-series techniques have been used in many domains [18], the standard time-series methods are primarily used for discovering patterns in time-series databases or forecasting the future values for existing time-series [19], [20], [21]. In our problem, the survival estimation is used to summarize the survival times of a group of objects (e.g. patients) while the response variable in time-series methods are outcomes depending on time which is an independent variable. Hence, although these two problems appear to be similar, the problem being tackled in this paper is significantly different and cannot be solved using time-series methods. In the presence of censoring and when the goal is to predict an occurrence of an event (which is usually binary in nature), time-series methods are not applicable. The only common theme that connects our approach to time-series methods is their ability to forecast in the future based on the events that occurred until a given time point.

There has been an increasing interest in adapting popular machine learning techniques to survival data [22], [23]. 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 information about the outcome variable only until a specific time point in the study. The censored observations in survival data might look similar to unlabeled samples in classification or unknown response in regression problem in the sense that status or time-to-event is not known for some of the observations. Such censored data have to be handled with special care within any machine learning method in order to make good predictions. Also, for censored data in survival analysis, we have information until 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 [24], [25] are not directly applicable to this problem.

Several machine learning based approaches have been proposed recently to address this censored data issue. Decision trees [26], [27], [28] and Artificial Neural Networks (ANN) [29], [30], [31], [32] for censored data represent some of the earliest works in this field. Well-known Support Vector Machine (SVM) algorithms have been adopted to accommodate censored data. Most of these methods treat the problem as regression [33], [34], [35], [36], [28]. More recently, advanced machine learning methods such active learning and regularized learning have also been incorporated into survival models [37], [38]. Other studies aim at modeling the problem within classification setting [39], [40]. However, comparison of the performance of these approaches show that these methods do not yield any significant improvements over the standard Cox model. There are also few other studies which aim at handling censored data as pre-processing step by giving some weights to the censored observations [41], [42]. In this paper, we tackle the problem of censoring using Kaplan-Meier method [43] to estimate the probability of event and the probability of censoring for each censored instance. Such an intuitive approach can be easily applied to survival data before any further analysis is performed.

One of the popular choices for predictive models is the Bayesian approaches including Naive Bayes and Bayesian Network which have been used widely for classification [44] and successfully applied in many domains [45]. However, there has been only few works in the literature using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

Bayesian methods for survival data [22], [46], [47]. Bayesian networks can visually represent all the relationships between the variables which makes it interpretable for the end users. This is in contrast to the simple Naive Bayes method that makes the independence assumption between all the features [44]. Despite the applicability of Bayesian network in the survival analysis domain, only a limited number of research efforts exist for tackling the censored data challenges. The authors of [48] developed a Bayesian neural network approach to model censored data. [49] gives weight to censored instances in order to learn Bayesian networks from survival data. More recently, [47] adapts a Bayesian network for survival data using an approach called inverse probability of censored weighting for each of the record in the dataset to handle the censoring issue.

The proposed 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. Existing methods only use the training data that is collected for the same time period as the test data. However, in real-world problems it is beneficial to make forecast of the events beyond the time period available in the training data. The basic idea of our approach is to take advantage of generative component of Bayesian methods (such as Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian network) to build a probabilistic predictive model [50] 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 unavailability 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 [51]. Figure 2 positions our paper along with the related methodologies available in the literature. It gives a complete characterization and some relevant references for modeling and forecasting approaches on time-series and event data.

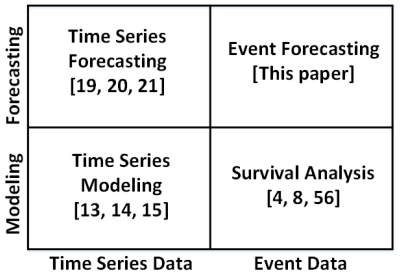


Fig. 2: Characterization of modeling and forecasting approaches on time-series and event data.

3 Preliminaries

This section introduces the preliminaries required to comprehend the proposed framework. First, the notations used in our study and our problem formulation are described. Next, details about the widely used Bayesian-based approaches such as Naive Bayes, Tree-Augmented Naive Bayes (TAN) and Bayesian Network are provided. These are the important components of the proposed method for predicting events in survival data at early stage of longitudinal studies. Finally, basic concepts of survival analysis are explained and a new method to handle censored data is introduced.

3.1 Problem Formulation

We begin by presenting the basic concepts and notations for survival analysis and Bayesian methods. Table 1 describes the notations used in this paper.

TABLE 1: Notations used in this paper

Name	Description
n	number of subjects
m	number of features
$\mathbf{x}$	n$\times m$ matrix of data
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 times
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 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 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 is not monitored). 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 is defined as $\delta_i = \mathbf{I}\{T_i \leq C_i\}$. Thus, a longitudinal dataset is 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. Suppose, among n subjects in the study, only $n(t_c)$ will experience the event at time t_c. After our data transformation, 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.

3.2 Bayesian Methods

We will now describe the basic idea of three popular Bayesian methods used in the context of prediction, namely, Naive Bayes, Tree-Augmented Naive Bayes, and Bayesian Network. All the three methods have certain commonalities in terms of using the conditional and prior probabilities. The main distinction between them is the way in which they model the dependency between the attributes and the way in which the conditional probability terms are computed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

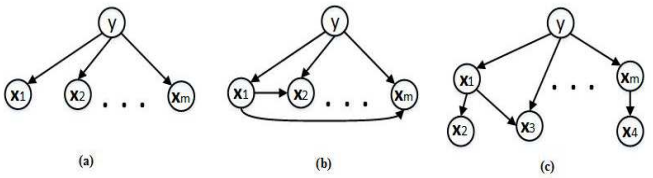


Fig. 3: An illustration of the basic structure of (a) Naive Bayes (b) TAN and (c) Bayesian Network classifiers.

3.2.1 Naive Bayes Classifier

Naive Bayes is a well-known probabilistic model which is widely used in many applications. Let us say that we have a training set similar to that in Figure 1 where the event occurrence information is available until time t_c. Based on the binary classification transformation explained above, using the Naive Bayes algorithm the event probability for subject i can be estimated as follows:

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

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:

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

Thus, it is a natural estimate for the likelihood function in Naive Bayes. The estimated probability that a random variable takes a certain value is equal to the number of times the value was observed divided by the total number of observations. This formula is valid for discrete attributes; However, it can be easily adapted for continuous variables as well [52].

3.2.2 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 [44]. 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 one other attribute at most, $\mathbf{x_p}(j)$, called the parent of $\mathbf{x_j}$. Illustration of the basic structure of the dependency in Naive Bayes and TAN is shown in Figure 3. 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 [44] between two attributes as shown in Eq. (3).

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

This function measures the information that $\mathbf{x_k}$ provides about $\mathbf{x_j}$ when the value of $y(t_c)$ is known. Then, a complete undirected graph in which the vertices correspond to the attributes and the edge weights are assigned using Eq. (3). 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:

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

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 maximum likelihood estimation (MLE) [52].

3.2.3 Bayesian Network Classifier

A Bayesian network is a graphical representation of a probability distribution over a set of variables. It can be considered as an extension of the TAN model where the features can be related to each other at various levels (Figure 3). It consists of two parts [53]:

1) A directed network structure in the form of a directed acyclic graph (DAG) which can be represented as $G = (V, E)$, where V denotes the set of vertices which represent variables, while E is the set of edges which show the dependency between the variables;
2) A set of the local probability distributions, one for each node variable, conditional upon each value combination of its parents.

Thus, a Bayesian network can be formally defined as $BN = \big(G, \Theta(G|D)\big)$ where $\Theta(G|D)$ is the Maximum likelihood estimation of the set of parameters in the probability distributions estimated based on the given data D. The Bayesian network structure in this paper is learnt by the well-known search-and-score based Hill-climbing algorithm [54]. The weight-adapted minimum description length (MDL) [44] scoring (Eq. (5)) function is used as the criterion function to be minimized for the Hill-climbing algorithm [55].

$$MDL(BN|D) = \frac{d}{2}\log N - LL(BN|D) \quad (5)$$

where d is the number of free parameters of a multinomial local conditional probability distribution; $LL(BN|D)$ is the log-likelihood of BN given D and can be estimated using the joint probability distributions. 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:

$$P(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \prod_{j=1}^{m} P(\mathbf{x_j} \mid Pa(\mathbf{x_j})) \quad (6)$$

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

a vector of m attributes. It can be formulated as follows:

$$P(y(t_c) = 1 \mid \mathbf{x}, t \leq t_c) = \frac{P(y(t_c) = 1, t \leq t_c) \prod_{j=1}^{m} P(\mathbf{x_j} \mid y(t_c) = 1, Pa(\mathbf{x_j}))}{P(\mathbf{x}, t \leq t_c)} \quad (7)$$

In Eq. (7), the first element in numerator 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.

3.3 Handling Censored Data

In general, survival analysis is a statistical methodology which contains time of a particular event of interest as the outcome variable which needs 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 data is considered to be *censored* [56].

Definition 1: Survival function. Considering the duration to be a continuous random variable T, the survival function, $S(t)$, gives the probability that the time of event occurrence is later than a certain specified time t. It is defined as

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)\, du \quad (8)$$

where $f(t)$ is a probability density function. For many real-world applications, typically the survival function monotonically decreases with respect to t.

Definition 2: Cumulative death distribution function. In contrast to survival function, the cumulative death distribution function $F(t)$ represents the probability that the time to the event of interest is no later than the certain specified time t. It is defined as:

$$F(t) = \Pr(T \leq t) = 1 - S(t) \quad (9)$$

Survival analysis involves the modeling of time-to-event data. We will use one of the popular parametric methods in survival analysis, accelerated failure time (AFT) [57] model, to adapt the probability of event using different time-to-event distributions.

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 the event time. Instead of using these sub-optimal approaches, our work handles censored data by dividing them into two groups [41]: *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. $S(t)$ is the probability that the event of interest has not occurred within the duration t. Using Kaplan-Meier estimator [43], the survival distribution is given by

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right) \quad (10)$$

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. (9), the probability of event can be estimated as

$$\hat{F}_e(t) = 1 - \hat{S}(t) \quad (11)$$

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 the censoring time, by setting "event" indicator $\delta_i^* = 1 - \delta_i$ [58]. Thus, Kaplan-Meier estimator for $G(t)$ is

$$\hat{G}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i^*}{n_i}\right) \quad (12)$$

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

$$\hat{F}_c(t) = 1 - \hat{G}(t) \quad (13)$$

We define a new label for censored data using Eqs. (11) and (13). 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 follow-up information for that subject, there is extremely a low chance of experiencing an event by 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.

4 Early Stage Event Prediction (ESP) Framework

In this section, we introduce our proposed Bayesian approach for handling early stage event prediction. As discussed in previous section, 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 a sufficient number of events. While survival analysis techniques are appropriate in handling such longitudinal data, they 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 a survival model is valid only for the specific observed time [5]. Therefore, the main objective of this section is to propose a framework to predict if the event will occur in the future for each subject based on information about only a few event occurrences at the initial stages of a longitudinal study.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

distributions and then we will introduce ESP-NB, ESP-TAN and ESP-BN algorithms which utilize this extrapolation method while computing the posterior probability of event occurrence.

4.1 Prior Probability Extrapolation

In order to predict the event occurrence in longitudinal data, we develop a technique that can estimate the ratio of event occurrence beyond the original observation time window (in other words, compute the *extrapolation for prior probability of event occurrence*). To achieve this goal, we extrapolate the prior probability of event occurrence using the accelerated failure time model (AFT). We consider two well-known distributions, Weibull and Log-logistic, which are widely studied in the literature for modeling time-to-event data [59]. 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 T_i follows a Weibull distribution, the cumulative probability distribution $F(t)$ with shape parameter a and scale parameter b can be estimated using

$$\hat{F}(t) = 1 - e^{-(t/b)^a} \tag{14}$$

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

$$\hat{F}(t) = \frac{1}{1 + (t/b)^{-a}} \tag{15}$$

Having the *cumulative probability distribution of event*, $F(t)$, where the shape parameter a and scale parameter b estimated at t_c, it can be easily extrapolated for any time t much beyond t_c.

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 3.2.1, 3.2.2 and 3.2.3). Since we are already extrapolating (in some sense approximating) in the prior probability component, it is not desirable to do a similar approximation again on the likelihood component. In addition, it is not feasible to extrapolate the likelihood component due to the various complexities involved in computing that component. We assume that the joint probability estimation from the Bayesian methods does not change over time since we have data only until t_c there is no plausible way to estimate the likelihood from the data beyond t_c. This is a reasonable assumption in survival data when the covariates do not depend on the time as the relation between the features at time t_c do not significantly change until the end of the study [60], and is very effective in practice in the presence of limited data. On the other hand as time passes, the prior probability for event occurrence needs to be updated since we do not have enough data to get the exact value for joint probability at the given future time t_f. 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.

4.2.1 ESP Naive Bayes (ESP-NB)

For Naive Bayes method using Eq. (1) and extrapolation method explained in previous section, the ESP-NB can be written as follows:

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \le t_f\big) = \frac{F(t_f)\prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1\big)}{P(\mathbf{x}, t \le t_f)} \tag{16}$$

4.2.2 ESP Tree-Augmented Naive Bayes (ESP-TAN)

Probability of event occurrence based on TAN method for time t_f using Eq. (4) can be estimated as follows:

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \le t_f\big) = \frac{F(t_f)\prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1, \mathbf{x_p}(j)\big)}{P(\mathbf{x}, t \le t_f)} \tag{17}$$

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: $\quad P\big(\mathbf{x_j} \mid y(t_c) = 1\big)$
3: **end**
Phase 2: Predict probability of event occurrence at t_f
4: fit AFT model to $D_n(t_c)$
5: $P\big(y(t_f) = 1, t \le t_f\big) = F(t)$
6: **for** $i = 1, ..., n$
7: $\quad$ estimate $P\big(y_i(t_f) = 1 \mid \mathbf{x}_i, t \le t_f\big)$
8: **end**
9: return $P\big(y(t_f) = 1 \mid \mathbf{x}, t \le t_f\big)$

Algorithm 1 outlines the proposed ESP framework. In the first phase (lines 1-3), for each attribute j, the algorithm estimates the 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.

4.2.3 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 earlier. Thus, the posterior probability estimation for event occurrence at time t_f can be defined as,

$$P\big(y(t_f) = 1 \mid \mathbf{x}, t \le t_f\big) = \frac{F(t_f)\prod_{j=1}^{m} P\big(\mathbf{x_j} \mid y(t_c) = 1, Pa(\mathbf{x_j})\big)}{P(\mathbf{x}, t \le t_f)} \tag{18}$$

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

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 $P(G|D_n(t_c))$
2: $score_{final} \leftarrow \infty$, $score = MDL(BN, D_n(t_c))$ (Eq. (5))
3: **while** $score_{final} > score$
4: $score_{final} \leftarrow score$
5: for every add/remove/reverse E_G on G
6: estimate $P(G_{new}|D_n(t_c))$
7: $score_{new} = MDL(BN_{new}, D_n(t_c))$
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: $P(y(t_f) = 1, t \leq t_f) = F(t)$
13: **for** all i in $D_n(t)$
14: estimate $P(y_i(t)|\mathbf{x}_i)$
15: Weibull using Eqs. (7), (14) and (18)
16: Log-logistic using Eqs. (7), (15) and (18)
17: **endfor**
18: return $P(y(t_f) = 1 \mid \mathbf{x}, t \leq t_f)$

1-2), the Hill-climbing algorithm will find a network with the minimum MDL based on the score function given in Eq. (5). 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 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 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 the data and for ESP-BN it is $O(m^kn)$, where k is the maximum number of parents (in our study we test different values of k to get the best performance within the range of 2 - 5) [61]. This means that ESP improves the prediction performance without increasing the complexity compared to its base models.

5 Experimental Results

In this section, we will show the results of our proposed ESP method on a wide range of datasets and provide comparisons with various baseline prediction methods. First, we explain the synthetic as well as real-world datasets that are used in our experiments. We also discuss the metrics that are used to quantitatively evaluate the performance of the proposed method. Finally, we will provide our experimental results and the practical implications of the ESP framework in survival studies will also be discussed.

5.1 Dataset Description

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

(i) Synthetic Datasets: We generated synthetic dataset in which the feature vectors $\mathbf{x}$ are created using a normal distribution $N(0, 1)$. Covariate coefficient vector, shown as β, is generated based on a uniform distribution $Unif(0, 1)$. Given the observed covariates $\mathbf{x}_i$ for observation i, the failure time, T can be generated by the procedure described in [62] as follows:

$$T_i = -\left(\frac{log(Unif(0,1))}{\lambda exp(\beta' \mathbf{x}_i)}\right)^{\nu} \quad (19)$$

In our experiments, we set $\lambda = 0.01$, $\nu = 2$ and generate two sets of synthetic data, namely, Syn1 with 5 features and 100 instances and Syn2 with 20 features and 1000 instances, where the time to event of interest follows a Weibull distribution.

(ii) Real-world Survival Datasets: Several real-world survival benchmark datasets are used in our experiments. Primary biliary cirrhosis (PBC), breast and colon cancer which are widely used in evaluating longitudinal studies and are available in the survival data repository[1]. We also used Framingham heart study dataset which is also publicly available [63].

In addition, we also used two in-house proprietary datasets. The first 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 [37]. The second dataset was obtained from Kickstarter[2], a popular crowdfunding platform. Each project was tracked for a specific period of time. If the project reaches the desired funding goal before its goal date, then it is considered to be a success (or the event has occurred). On the other hand, the project is considered to be censored if it fails to reach its goal amount within the goal date [64]. All the datasets (except the EHR) used in our work are made publicly available at https://github.com/MLSurvival/ESP.

5.2 Performance Evaluation

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

- ***Accuracy*** is expressed as the percentage of instances in the test set that are classified correctly.
- ***F-measure*** is defined as the harmonic mean of precision and recall. A high value of F-measure indicates that both precision and recall are reasonably high.

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

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

For our implementation, the joint probability for Naive Bayes and TAN is learned using *e1071* package [65] available in the R programming language. Bayesian network structure for the proposed ESP-BN method is learned using

1. http://cran.rproject.org/web/packages/survival/
2. www.kickstarter.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

TABLE 2: Statistics of the datasets used in our experiments. T_{50} and T_{100} correspond to the time taken (in days) for the occurrence of 50% and 100% of the events, respectively. $C_{50}\%$ and $C_{100}\%$ give the percentage of censored instances at T_{50} and T_{100}, respectively.

Dataset	#Features	#Instances	#Events	$\mathbf{C_{50}}\%$	$\mathbf{C_{100}}\%$	$\mathbf{T_{50}}$	$\mathbf{T_{100}}$
Syn1	5	100	50	20%	50%	1014	3808
Syn2	20	1000	602	29%	40%	943	7723
Breast	8	673	298	25%	56%	646	2659
Colon	13	888	445	4%	50%	394	3329
PBC	17	276	110	27%	60%	1191	4456
Framingham	16	5209	1990	0%	62%	1991	5029
EHR	77	4417	3479	0%	21%	50	4172
Kickstarter	54	4175	1961	17%	53%	21	60

TABLE 3: Comparison of Accuracy values for Cox, LR, RF, NB, TAN and BN along with the proposed ESP-NB, ESP-TAN and ESP-BN methods (and their 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 4: Comparison of F-measure values for Cox, LR, RF, NB, TAN and BN along with the proposed ESP-NB, ESP-TAN and ESP-BN methods (and their 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 (0.170)	0.578 (0.194)	0.543 (0.169)	0.549 (0.184)	0.562 (0.190)	0.621 (0.145)	0.627 (0.148)	**0.630 (0.180)**
PBC	0.61 (0.141)	0.529 (0.130)	0.613 (0.120)	0.541 (0.121)	0.562 (0.150)	0.575 (0.140)	0.712 (0.110)	0.719 (0.099)	**0.725 (0.098)**
Framingham	0.755 (0.078)	0.735 (0.093)	0.792 (0.085)	0.787 (0.075)	0.798 (0.073)	0.845 (0.083)	0.873 (0.073)	0.905 (0.059)	**0.925 (0.066)**
EHR	0.672 (0.110)	0.584 (0.166)	0.617 (0.188)	0.684 (0.156)	0.708 (0.198)	0.715 (0.210)	0.781 (0.126)	0.798 (0.160)	**0.826 (0.160)**
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)**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

TABLE 5: Comparison of AUC values for Cox, LR, RF, NB, TAN and BN along with the proposed ESP-NB, ESP-TAN and ESP-BN methods (and their 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)	0.720 (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	0.863 (0.007)	0.935 (0.002)	0.929 (0.005)	0.945 (0.002)	0.953 (0.005)	0.959 (0.004)	0.971 (0.007)	0.973 (0.004)	**0.979 (0.001)**
EHR	0.612 (0.023)	0.637 (0.017)	0.650 (0.025)	0.633 (0.019)	0.638 (0.025)	0.651 (0.026)	0.654 (0.018)	0.649 (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)**

a hill-climbing algorithm that is available in the open-source Weka software [66], while the proposed model is implemented using the R programming language. The *coxph* and *survreg* functions 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 method is also implemented in R using the survival package. The source code of the proposed algorithms in R programming environment is available at https://github.com/MLSurvival/ESP.

5.3 Results and Discussion

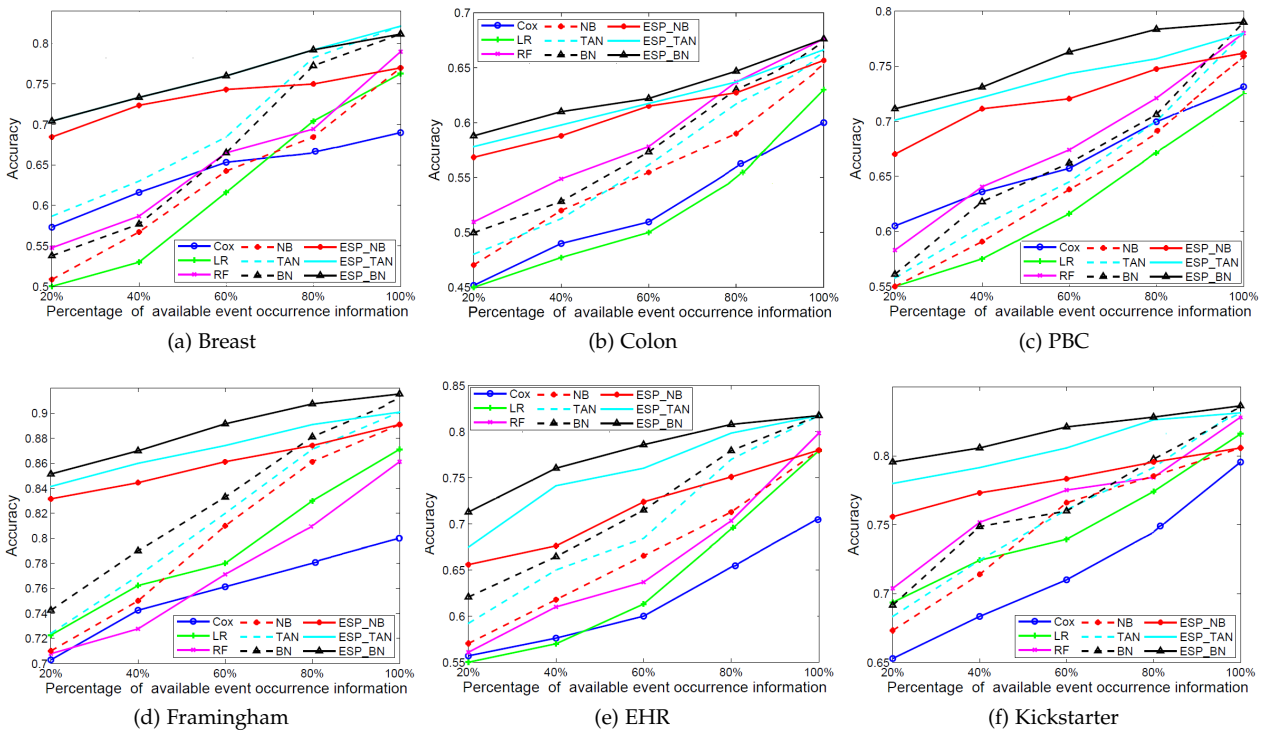
Tables 3, 4, and 5 summarize the performance comparison results for Accuracy, F-measure, and AUC, respectively. We compared the proposed ESP-NB, ESP-TAN and ESP-BN algorithms using the best performed distributions from extrapolation techniques with Cox, Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Tree-Augmented Naive Bayes (TAN) and Bayesian Network (BN) classification methods. All the models are trained using the data collected at the time point where only 50% of events have occurred (T_{50}) and the event forecasting is done using the data at the end of study (T_{100}). We used stratified 10-fold cross-validation and average values (along with the standard deviations) of the results on all 10-folds are reported. For the ESP based methods, we extrapolated using both Weibull and log-logistic distributions and best results are being reported. It should be noted that in most of the cases Weibull distribution provided better results.

For all of the datasets, our results evidently show that the proposed ESP-based method is, on an average, 20% more accurate compared to existing methods using only a limited amount of training data. These results confirm the fact that by incorporating the time-to-event extrapolation method within the ESP framework, forecasting can be done more accurately compared to the standard methods. It is important to note that the choice of the best algorithm will depend on the nature of the dataset. For instance, ESP-NB builds on independence assumption between attributes which does not hold in many survival applications. Thus, the introduced ESP-TAN and ESP-BN relaxed this assumption and thus yielding better performance in almost all of the datasets. Upon further analysis of our results, we can observe that, in most of the cases, ESP-BN has higher accuracy compared to its other Bayesian counterparts. This is due to the fact that Bayesian network can model more complex data especially in the presence of feature dependencies [67].

In Figures 4, 5 and 6, we present the prediction performance of different methods by varying the percentage of event occurrence information that is available to train the model in the real-world datasets. For example, 20% on the x-axis corresponds to the training data obtained when only 20% of the events have occurred and the prediction of the event occurrences was made on the data at the end of the study period. From these plots, we can see that the performance of the ESP algorithm improves when there is more information on the event occurrence in the training data. For all the cases, our proposed ESP-based methods provide more accurate predictions compared to other techniques and the improvements are consistent across all the benchmark datasets. 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.

When 100% of the training data is available, the performance of the proposed ESP methods will converge to that of the original baseline methods since the prior probabilities in both scenarios will be the same and fitting a distribution (and extrapolating it) will not have any impact when evaluated at the end of the study since there is effectively no extrapolation that is done. We should also mention that in our experiments the percentage of censoring in each dataset is different. Therefore, it is hard to measure how the amount of censored data affects the results. However, since the amount of censored and event data are closely related, one can measure the effect of censored data using the number of

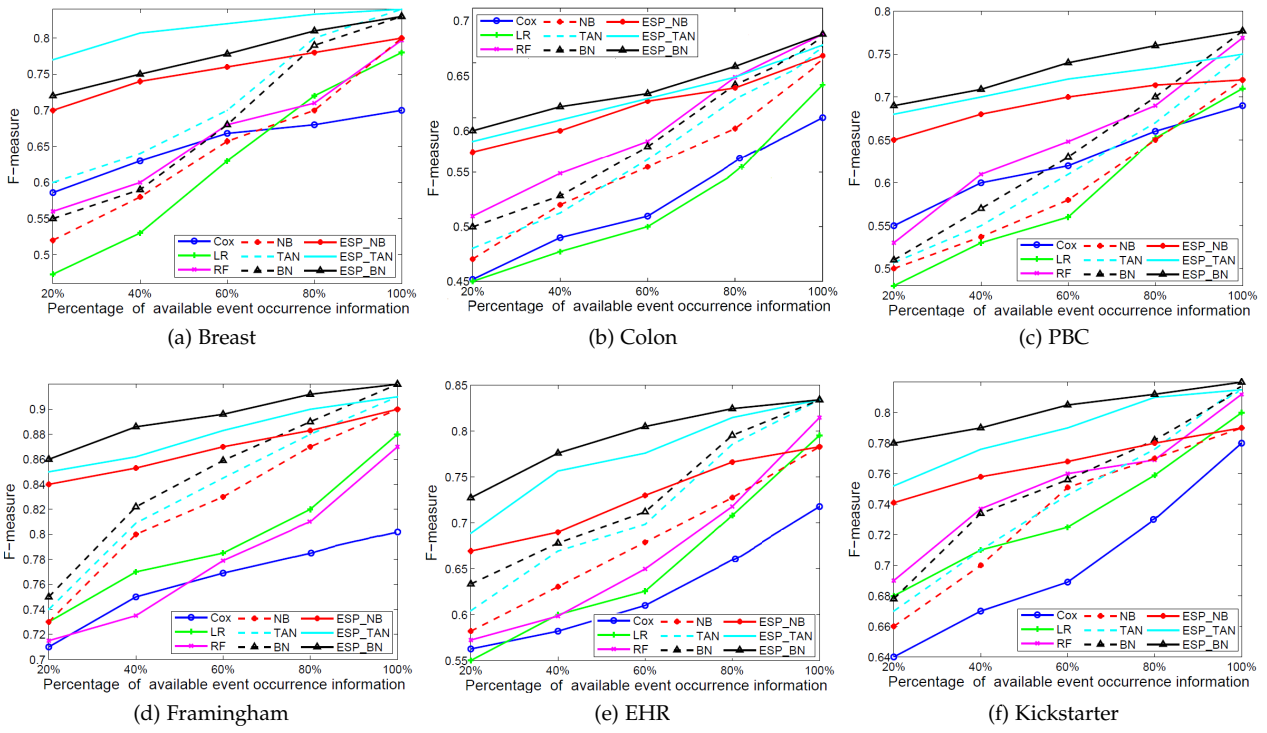
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering



(a) Breast (b) Colon (c) PBC

(d) Framingham (e) EHR (f) Kickstarter

Fig. 4: Accuracy values of different methods obtained by varying the percentage of event occurrence information for various datasets.



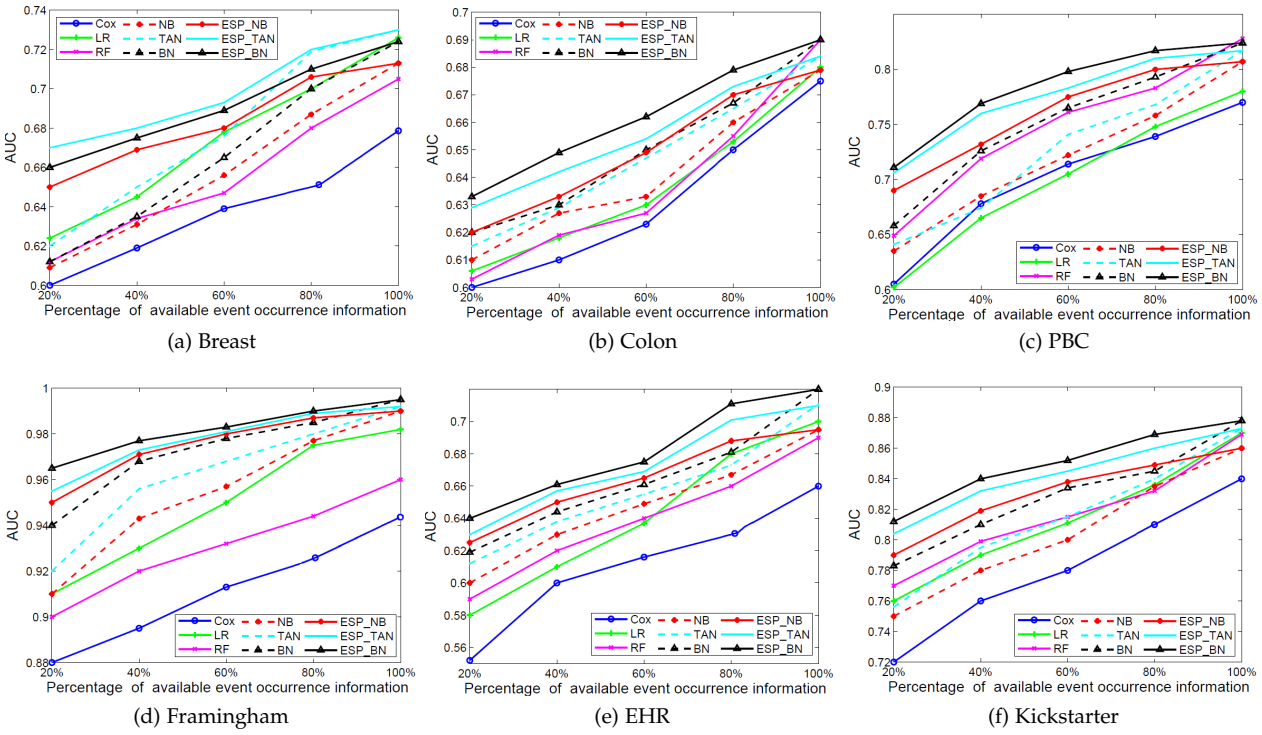
(a) Breast (b) Colon (c) PBC

(d) Framingham (e) EHR (f) Kickstarter

Fig. 5: F-measure values of different methods obtained by varying the percentage of event occurrence information for various datasets.

events which is shown in Figures 4-6. In general, we observe that the less censored data we have, the higher the accuracy we could achieve. In order to measure the improvements made by handing censored data, we compared the results in Tables 3-5 with those provided in [5]. The results support our claim that the proposed Bayesian models can provide an accurate forecasting for event occurrence beyond the observation time. From our experiments, we can conclude that our model obtains useful practical results at the initial phase of a longitudinal study and can provide good predictions about the event occurrence at the end of the study using only a limited information. The proposed prediction model

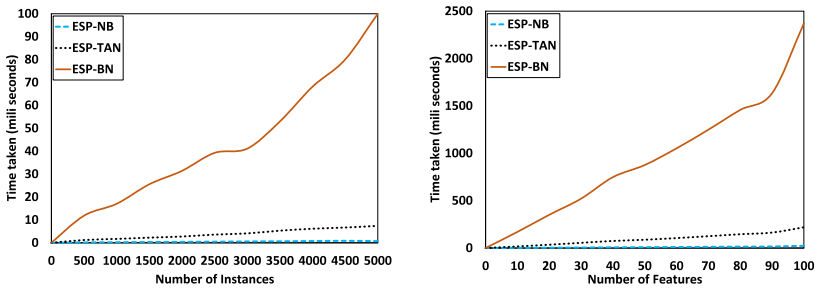
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering



(a) Breast (b) Colon (c) PBC

(d) Framingham (e) EHR (f) Kickstarter

Fig. 6: AUC values of different methods obtained by varying the percentage of event occurrence information for various datasets.



(a) Scalability w.r.t. the number of instances (b) Scalability w.r.t. the number of features

Fig. 7: Assessing the scalability of ESP-NB, ESP-TAN and ESP-BN with different number of instances and features.

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.

5.4 Scalability Experiments

As mentioned earlier (Section 4.2), the time complexity of the extrapolation component of the model is constant ($O(n)$) and does not depend on the number of features or instances. Therefore, time complexity of the ESP-based algorithms follows that of the corresponding base learning method that is chosen. In other words, the ESP-NB, ESP-TAN and ESP-BN have the same time complexity as NB, TAN and BN, respectively. This means that ESP framework improves the prediction performance without increasing the time complexity. In this section, we study the scalability of our proposed ESP-based algorithms when the number of instances or features in the dataset are varied by random selection. We randomly sampled different number of features or instances from the original dataset and estimated the average running time of each of the proposed ESP based algorithms (average of 100 runs).

In Figure 7, we provide the scalability plots for ESP-NB, ESP-TAN and ESP-BN. To obtain these plots we sampled different set of instances and features in an increasing order and obtained the time required to build our proposed ESP-based algorithms. The x-axis represents the selected number of instances (in Figure 7(a)) and features (in Figure 7(b)) and the y-axis represents the time taken in milliseconds. These plots indicate that ESP-NB is relatively faster even when the number of instances and features is large. This is because the complexity of ESP-NB is linear with respect to instances and features. As number of instances increase, the time taken for ESP-TAN and ESP-BN is also increased. However, ESP-TAN has quadratic and ESP-BN has trinomial runtime complexity (if k, the number of parents for each features, is 3), but it tends to build more effective models. Hence, there is a trade-off between complexity and performance. It is clear that, in the presence of high-dimensional data, ESP-NB will be the optimal choice. However, if there are many dependencies

between features or data has a high dimension, ESP-TAN is a better choice. ESP-BN would be recommended to use only when the data consists of lots of complex dependencies and at the same time has only a reasonable dimensionality. For high-dimensional data, it is recommended to use unsupervised dimensionality reduction methods before applying our proposed early stage prediction algorithms.

6 Conclusion

In many real-world application domains, it is important to forecast the occurrence of future events by only using the data collected at early stages of longitudinal studies. In this paper, we developed new early stage event prediction framework through fitting a statistical distribution to time-to-event data with fewer available events at the early stages. One of the common characteristic of longitudinal data is the presence of censored instances where the outcome is not known after a certain time period during the study. Instead of excluding such censored data, we developed a new mechanism to handle this data by estimating the probability of event and the probability of being censored using the 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 estimated using the current (limited) information of event occurrence. This is extremely important in 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. Using this approach, we developed three new Bayesian algorithms to effectively predict the event occurrence for future time points using the training data obtained at early stage of the study. Our extensive experiments using both synthetic and real datasets demonstrate that the proposed ESP-based algorithms are more effective in forecasting events at future time points compared to the widely used Cox model and other popular classification methods.

Acknowledgments

This work was supported in part by the US National Science Foundation grants IIS-1231742, IIS-1527827 and IIS-1646881.

References

[1] N. Lavrač, "Selected techniques for data mining in medicine," *Artificial intelligence in medicine*, vol. 16, no. 1, pp. 3–23, 1999.

[2] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," *Age*, vol. 58, no. 13, pp. 10–110, 2006.

[3] M. Jahanbani Fard, S. Ameri, and A. Zeinal Hamadani, "Bayesian approach for early stage reliability prediction of evolutionary products," in *Proceedings of the International Conference on Operations Excellence and Service Engineering*, 2015, pp. 361–371.

[4] D. W. Hosmer, S. Lemeshow, and S. May, *Applied survival analysis: regression modeling of time to event data*. New York: Wiley, 1999.

[5] M. J. Fard, S. Chawla, and C. K. Reddy, "Early-stage event prediction for longitudinal data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2016, pp. 139–151.

[6] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, "Survival analysis based framework for early prediction of student dropouts," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016.

[7] R. G. Miller and J. Halpern, "Regression with Censored Data," *Biometrika Trust*, vol. 69, no. 3, pp. 521–531, 1982.

[8] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[9] R. G. Miller, "Least squares Regression with Censored Data," *Biometrics Trust*, vol. 63, no. 3, pp. 449–464, 1976.

[10] J. Buckley and I. James, "Linear Regression with Censored Data," *Biometrics Trust*, vol. 66, no. 3, pp. 429–436, 1979.

[11] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, vol. 34, no. 2, pp. 187–220, 1972.

[12] H. Koul, V. Susarla, and J. Van Ryzin, "Regression Analysis with Randomly Right-Censored," *The Annals of Statistics*, vol. 9, no. 6, pp. 1276–1288, 1981.

[13] A. C. Harvey, *Time series models*. New York: Harvester Wheatsheaf, 1993, vol. 2.

[14] T. W. Liao, "Clustering of time series data: a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[15] J. D. Hamilton, *Time series analysis*. Princeton: Princeton university press, 1994, vol. 2.

[16] Z. Xing, J. Pei, and S. Y. Philip, "Early classification on time series," *Knowledge and information systems*, vol. 31, no. 1, pp. 105–127, 2012.

[17] G. He, Y. Duan, R. Peng, X. Jing, T. Qian, and L. Wang, "Early classification on multivariate time series," *Neurocomputing*, vol. 149, pp. 777–787, 2015.

[18] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2010.

[19] C. Chatfield, *Time-Series Forecasting*. CRC Press, 2000.

[20] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015, vol. 5.

[21] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[22] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and healthcare." *Artificial intelligence in medicine*, vol. 30, no. 3, pp. 201–14, Mar. 2004.

[23] Y. Li, K. Xu, and C. K. Reddy, "Regularized parametric regression for high-dimensional survival analysis," in *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2016.

[24] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*. MIT press Cambridge, 2006, vol. 2.

[25] Z. Zhou and M. Li, "Semi-supervised regression with co-training." in *IJCAI*, 2005, pp. 908–916.

[26] L. Gordon and R. Plshen, "Tree-structured survival analysis," *Cancer Treat Reports*, vol. 69, no. 10, pp. 1065–1074, 1985.

[27] M. R. Segal, "Regression Trees for Censored Data," *Biometrics*, vol. 44, no. 1, pp. 35–47, 1988.

[28] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. Suykens, "Support vector methods for survival analysis: a comparison between ranking and regression approaches," *Artificial intelligence in medicine*, vol. 53, no. 2, pp. 107–18, Oct. 2011.

[29] C. Chi, W. N. Street, and W. H. Wolberg, "Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets," in *AMIA Annual Symposium*, 2007, pp. 130–134.

[30] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems*. MIT Press, 1998, pp. 368–374.

[31] C. Cordon-Cardo, A. Kotsianti, D. A. Verbel, M. Teverovskiy *et al.*, "Improved prediction of prostate cancer recurrence through systems pathology," *Journal of clinical investigation*, vol. 117, no. 7, pp. 1876–1883, 2007.

[32] M. J. Donovan, S. Hamann, M. Clayton, F. M. Khan *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*, vol. 26, no. 24, pp. 3923–3929, Aug. 2008.

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

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

[35] Y. Li, B. Vinzamuri, and C. K. Reddy, "Regularized weighted linear regression for high-dimensional censored data," in *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2016.

[36] J. Shim and C. Hwang, "Support vector censored quantile regression under random censoring," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 912–919, Feb. 2009.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2608347, IEEE Transactions on Knowledge and Data Engineering

[37] B. Vinzamuri and C. K. Reddy, "Cox regression with correlation based regularization for electronic health records," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 757–766.

[38] B. Vinzamuri, Y. Li, and C. K. Reddy, "Active learning based survival regression for censored data," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 241–250.

[39] L. Evers and C. Messow, "Sparse kernel methods for high-dimensional survival data." *Bioinformatics (Oxford, England)*, vol. 24, no. 14, pp. 1632–8, Jul. 2008.

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

[41] 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*, vol. 20, no. 1, pp. 59–75, 2000.

[42] I. Stajduhar and B. Dalbelo-Basic, "Uncensoring censored data for machine learning: A likelihood-based approach," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7226–7234, Jun. 2012.

[43] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[44] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[45] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Taylor & Francis, 2014, vol. 2.

[46] J. Wolfson, S. Bandyopadhyay, M. Elidrisi, G. Vazquez-Benitez *et al.*, "A naive bayes machine learning approach to risk prediction using censored, time-to-event data," *Statistics in Medicine*, vol. 34, no. 21, pp. 2941–2957, 2015.

[47] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez *et al.*, "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*, vol. 29, no. 4, pp. 1033–1069, 2015.

[48] P. J. Lisboa, H. Wong, P. Harris, and R. Swindell, "A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artificial intelligence in medicine*, vol. 28, no. 1, pp. 1–25, 2003.

[49] I. Štajduhar and B. Dalbelo-Bašić, "Learning bayesian networks from survival data using weighting censored instances," *Journal of biomedical informatics*, vol. 43, no. 4, pp. 613–622, 2010.

[50] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, pp. 841–848, 2002.

[51] J. A. Lasserre, C. M. Bishop, and T. P. Minka, "Principled hybrids of generative and discriminative models," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 87–94.

[52] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[53] D. Heckerman, "A tutorial on learning with bayesian networks," *Learning in graphical models*, pp. 301–354, 1998.

[54] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[55] W. Lam and F. Bacchus, "Learning bayesian belief networks: An approach based on the mdl principle," *Computational intelligence*, vol. 10, no. 3, pp. 269–293, 1994.

[56] C. K. Reddy and Y. Li, "A review of clinical prediction models," in *Healthcare Data Analytics*, C. K. Reddy and C. C. Aggarwal, Eds. Chapman and Hall/CRC Press, 2015.

[57] L. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[58] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

[59] K. J. Carroll, "On the use and utility of the weibull model in the analysis of survival data," *Controlled clinical trials*, vol. 24, no. 6, pp. 682–701, 2003.

[60] X. Jiang, D. Xue, A. Brufsky, S. Khan, and R. Neapolitan, "A new method for predicting patient survivorship using efficient bayesian network learning," *Cancer informatics*, vol. 13, no. 13, pp. 47–57, 2014.

[61] J. Su and H. Zhang, "Full bayesian network classifiers," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 897–904.

[62] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate cox proportional hazards models," *Statistics in Medicine*, vol. 24, no. 11, pp. 1713–1723, 2005.

[63] 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*, vol. 107, no. 2, pp. 539–556, 1963.

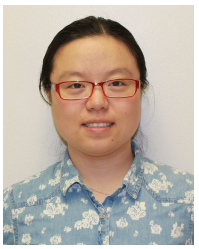
[64] V. Rakesh, J. Choo, and C. K. Reddy, "Project recommendation using heterogeneous traits in crowdfunding," in *Ninth International AAAI Conference on Web and Social Media*, 2015, pp. 257–266.

[65] 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.

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

[67] J. Cheng and R. Greiner, "Comparing bayesian network classifiers," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 101–108.

Mahtab J. Ford received her Ph.D. in Industrial & System Engineering and M.Sc. in Computer Science from Wayne State University. Her main areas of research are data mining, machine learning, survival analysis, healthcare and robotics. She is a student member of the INFORMS and the IEEE.



Ping Wang is a Ph.D. student in the Department of Computer Science at Virginia Tech. She received her MS in Computer Science from Wayne State University and BS in Statistics from Civil Aviation University of China. Her research interests include data mining, machine learning, survival analysis and healthcare.

Sanjay Chawla is a Professor in the School of Information Technologies, University of Sydney. His research work has appeared in leading data mining journals and conferences including ACM TKDD, Machine Learning, IEEE TKDE, DMKD, ACM SIGKDD, IEEE ICDM, SDM, and PAKDD. He is an associate editor for IEEE TKDE and serves on the editorial board of Data Mining and Knowledge Discovery. He served as a Program Co-Chair of PAKDD 2012. He received his PhD in 1995 from the University of Tennessee, Knoxville, USA under Professor Suzanne Lenhart.

Chandan K. Reddy is an Associate Professor in the Department of Computer Science at Virginia Tech. He received his PhD from Cornell University and MS from Michigan State University. His primary research interests are in the areas of data mining and machine learning with applications to healthcare, bioinformatics, and social network analysis. His research is funded by NSF, NIH, DOT, Susan G. Komen for the Cure Foundation. He has published over 75 peer-reviewed articles in leading conferences and journals. He received the Best Application Paper Award at the ACM SIGKDD conference in 2010 and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a senior member of the IEEE and a life member of the ACM.