JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Bayesian approach for Predicting in longitudinal data at prima stage

[1] Bankal Pavankumar, [2] Vasavi Ravuri , [3]Dr.K.Srinivas
[1] Department Of CSE, VNRVJIET, HYDERABAD- 500090, Telangana, India
[2] Assistant Professor, Department Of CSE, VNRVJIET, HYDERABAD- 500090, Telangana, India
[3]Assistant Professor, Department Of CSE, VNRVJIET, HYDERABAD- 500090, Telangana, India

study is a significant and difficult problem with considerable practical significance in many real-world applications. In contrast to typical classification and regression problems where a domain expert may provide labels for the data quickly, training data in these longitudinal studies must only be acquired by waiting for the occurrence of a significant number of events. Utilizing data gathered in the past over a predetermined period of time, survival analysis seeks to directly anticipate the time to an event of interest. It cannot, however, provide a response to the unanswered query 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 study is to predict an event's recurrence at a future time point using only information from a limited sample of events that occurred at the beginning of a longitudinal study. Due to the censoring of data on event occurrence and the availability of just a small number of data on events that occurred during the initial phase of the inquiry, this issue presents two important challenges. In order to create event prediction models that are trained early on in longitudinal research, we offer a novel Early-Stage Prediction (ESP) framework. First, using the Kaplan-Meier estimator, we create a new technique for dealing with censored data in order to address the first obstacle. We next build “three algorithms, namely, ESP-NB, ESP-TAN, and ESPBN, to efficiently forecast event occurrence utilizing training data obtained at an early stage of the investigation. These algorithms enhance the Naive Bayes, Tree-Augmented Naive Bayes (TAN), and Bayesian Network approaches based on the proposed framework”. More particularly, by modifying the prior probability of the event's occurrence for future time points, our method successfully combines Bayesian methodologies with an Accelerated Failure Time (AFT) model. With the aid of numerous synthetic and actual benchmark datasets, the suggested framework is assessed. Our extensive collection of trials shows that the suggested ESP framework is, on average, 20% more accurate than existing systems even using only a little quantity of event information in the training data.

1. INTRODUCTION

In many application industries, it has been normal practice to collect data over time and maintain track of the occurrence of notable events over a predetermined time period. These studies, which track people over time to monitor particular hazards, are typically referred to as longitudinal studies. A significant issue in longitudinal studies is creating accurate prediction models to assess the outcome of a given event of interest. Such studies are common in many real-world industries, including engineering, healthcare, and reliability [1, 2, 3]. Their major objective is to create models that can precisely calculate the likelihood that an important event will occur at a given time.[8-10]

One of the main issues in these longitudinal studies is the fact that training data in these tasks can only be obtained by waiting for the occurrence of a sufficient number of events, as opposed to the standard supervised learning problems where labels can be provided by a domain expert in a timely manner. The inability of early longitudinal studies to predict the occurrence of events at future time points using the data at hand is thus a serious problem.[11-14] Additionally, not all cases in the study will necessarily have the event occur, so the outcome variable may not be full.

'Censoring' is another name for this occurrence. Building event predicting models while dealing with censored data is a difficult task that has great application in longitudinal investigations. [4-7]

This study aims to address 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 issue displays two significant difficulties: 1. the filtering (lack of full information) and 2. The fact that only a portion of the activities that occurred during the study's initial phase are currently available.
Here are some examples of real-world situations that motivate early stage time-to-event prediction

- When a new treatment option (or medication) becomes available in the healthcare industry, it is important to research how it affects a specific patient population in order to determine the treatment's effectiveness. This patient population is tracked throughout time, and an event in this population is the patient's hospital admission as a result of unsuccessful treatment. When there are only a few patients in the hospital, it is important to measure this treatment's efficacy as soon as feasible [5].
- To increase graduation rates in education, it is crucial to identify students who are in danger of quitting their studies early. In terms of application, being able to develop a reliable prediction model with only preliminary data might be highly beneficial [6].
- Let's look at an example to help illustrate the difficulties and issues around this issue, which is depicted in Figure 1 below. In this example, a longitudinal research with six individuals is done, and data on event occurrence up until time $t_c$is recorded, with the event being experienced by only subjects S2 and S5. Our study tries to predict the occurrence of the event at time t_f (for instance, the study's completion). [15-17]To forecast the event occurrences by the time the study is finished (t_f), only the event occurrences up to the observation time $\boldsymbol{t_c}$. are accessible during the training phase. It should be noted that all subjects at $\boldsymbol{t_c}$. (shown by 'X') except S2 and S5 are suppressed. However, something will occur for subjects S1 and S6 over the course of the time period t_f.

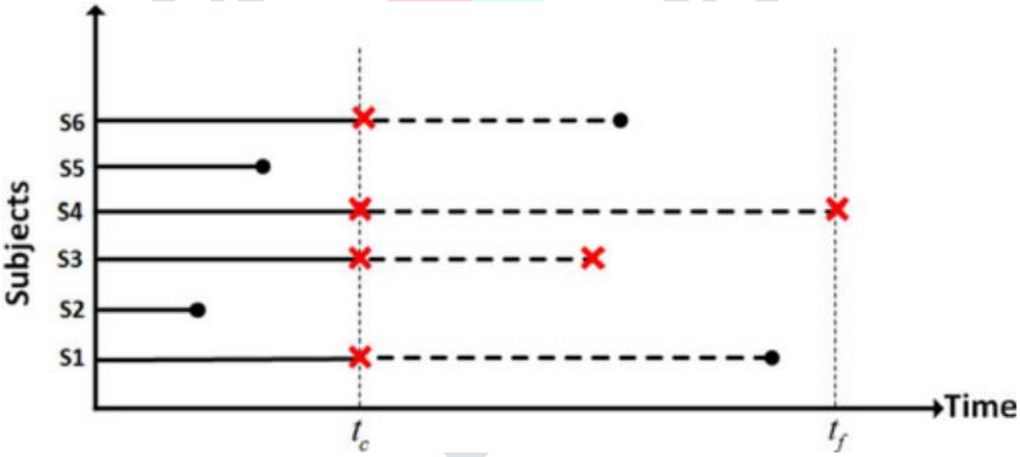


Figure 1. Using information only available up until time $\boldsymbol{t_c}$, an example is used to illustrate the difficulty of event predicting.

- This situation clearly drives the need for algorithms that can accurately forecast events using the training data at time $\boldsymbol{t_c}$ when few events have happened. This is a significant issue in the field of longitudinal research since the only way to gather accurate data in this situation is to wait a long enough time until all available information regarding the occurrence of the event is gathered. [18]

- In this research, we present a novel Kaplan-Meier estimator-based approach to handle censored data. Then, we will develop event prediction models that are trained early in longitudinal studies utilizing a brand-new Early Stage Prediction (ESP) framework. To be more precise, we develop three algorithms—ESP-NB, ESP-TAN, and ESP-BN—using training data obtained at an early stage of the study to reliably estimate the occurrence of events. Based on Naive Bayes, Tree-Augmented Naive Bayes (TAN), and Bayesian Networks, this approach. The proposed system is evaluated using a broad range of benchmark datasets, both synthetic and actual. Our extensive series of tests show that, in comparison to the other alternative techniques, the proposed ESP framework is more effective at forecasting future event occurrences with less training data.

LITERATURE SURVEY

2.1 Bayesian Methods

Now, we'll go over the fundamental principles of three well-known Bayesian prediction techniques, including Naive Bayes, Tree-Augmented Naive Bayes, and Bayesian Network [13]. The use of conditional and prior probabilities is a feature shared by all three approaches. The key difference between them is how they compute the conditional probability terms and model the relationship between the attributes and dependencies.

2.1.1 Naïve Bayes classifier

Naive Bayes is a popular probabilistic model with many applications. Assume we have a training set that looks like Figure 2.1.1. Where details concerning the occurrence of the event are available till time $\boldsymbol{t_c}$. For subject I, the event probability can be calculated using the Naive Bayes method as follows:

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

The prior probability of the event occurring at time tc is the first element of the numerator. A conditional probability distribution, which makes up the second component, can be calculated as follows:

$$P(x = x_j | y(t_c = 1) = \frac{\sum_{j=1}^{n} \quad (y(t_c = 1, x_{ij} = x_j)}{\sum_{i=1}^{n} \quad (y(t_c) = 1)}$$

As a result, in Naive Bayes, it is a natural estimate for the likelihood function. The ratio of the total number of observations to the number of times a certain value was observed is the estimated likelihood that a random variable would take that value. This formula works for discrete qualities, but it can also work well for continuous variables.

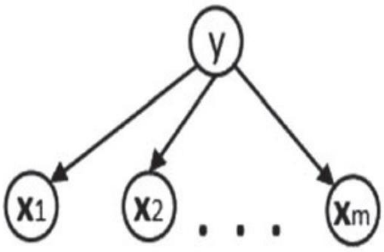


Figure 2.1.1 Naive Bayes Classifier

2.1.2 Tree-Augmented Naïve Bayes classifier

The assumption of attribute independence is relaxed in the Tree-Augmented Naive Bayes, a version of the Naive Bayes. The Naive Bayes model is forced into a tree structure using the TAN method, which restricts the interaction between the variables to one level. This method permits each attribute x_j to depend on the class as well as a maximum of one additional attribute $x_p(j)$, known as $x_j's$ parent. Figure 2.1.2 provides an illustration of the fundamental makeup of the dependency in naive bayes and TAN. The tree for the TAN model should initially be built based on the conditional mutual information [11] between two characteristics as given the training set (x, $y(t_c)$).

$$I(y(t_c)) = \sum_{x_j, x_k, y(t_c)} \quad P(x_j, x_k, y(t_c) log \frac{P(x_j, x_k | y(t_c))}{P(y(t_c))p(x_k | y(t_c))}$$

When the value of y (t_c) is known, the information that x_k offers about x_j is measured by this function. Then, a full undirected graph is created, in which the attributes are represented by the vertices, and the edge weights are determined by Eq. A directed tree is produced by choosing a random root variable and setting the direction of all the edges that branch out from the root after building a maximum weighted spanning tree. After the tree has been built, the conditional probability of each attribute with respect to its parent and class label is calculated and saved. As a result, the following formula can be used to quantify the probability of an event occurring at time t_c:

$$P(x, t \leq t_c) = \frac{P(y(t_c) = 1, t \leq t_c) \prod_{j=1}^{m} \quad P(x_j | y(t_c) = 1, x_p(j))}{P(x, t \leq t_c)}$$

The numerator consists of the prior probability of the event occurring at time t_c and the conditional probability distributions, which can be calculated using maximum likelihood estimation [14].

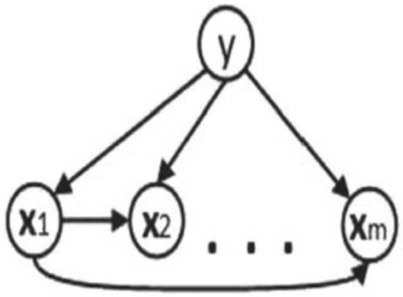


Figure 2.1.2 Tree-Augmented Naïve Bayes classifier

Kaplan-Meier estimator [10]

The product-limit formula predicts the pro even in cases where part of the items are not visible to fail or die. a portion of living things or objects that are physical. The actuarial and reduced-sample techniques are also being studied.

"The origins of this study can be traced back to Paul Meier's interaction with Greenwood's paper1 on the protracted nature of cancer in 1952 at Johns Hopkins University (now the University of Chicago). I developed an interest in the repeaters' vacuum tube lifespan in the telephone cables buried in the ocean a year later while working at Bell Telephone Laboratories. When I gave John W. Tukey my manuscript, he told me about Meier's work, which was already well-known among some of our coworkers. The Journal of the American Statistical Association accepted both articles and suggested a joint paper. We had to communicate extensively over the course of four years to resolve our divergent viewpoints, and we worried that someone else may publish the concept during that time. "The nonparametric estimate defines a discrete distribution, in which all the probability is concentrated at a finite number of points, or (for a large sample) an actuarial approximation thereto, giving the probability in each of a number of succeeding intervals.

METHODOLOGY

The ESP algorithm

We will now go over the two steps of the ESP algorithm. Utilizing training data up until time t_c, the conditional probability distribution is computed in the first phase Since we are already extrapolating (in a sense, approximating) “In the prior probability component, it is not recommended to perform a similar approximation on the likelihood component. Additionally, due to the numerous complications involved in estimating the likelihood component, extrapolating that component is not practical. Since we only have data up until t_c, we presume that the combined Bayesian probability estimation remains constant over time. There is no logical way to determine the likelihood from the data beyond t_c. This is a reasonable assumption in survival data when the covariates do not rely on the time and is particularly successful in practice in the face of limited data because the association between the features at time t_c do not significantly change until the end of the trial [60]. However, as time goes on, it becomes necessary to update the prior probability of an event occurring since we lack the data to calculate the joint probability with certainty at the specified future time t_f. Using various extrapolation approaches, we extrapolate the prior probability of event occurrence at time t_f that is past the observed time in the second phase.

5.1.1 ESP Naive Bayes (ESP-NB)

The ESP-NB can be stated as follows using the Naive Bayes approach using Eq. (1) and the extrapolation method described in the preceding section:

$$P\left(x, t \le t_f\right) = \frac{F\left(t_f\right)\prod_{j=1}^{m} \quad P(x = x_j | y(t_c) = 1)}{P(x, t \le t_f)}$$

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

According to Eq. (4), the probability of an event occurring based on the TAN approach for time tf can be calculated as follows:

$$P\left(x, t \le t_f\right) = \frac{F\left(t_f\right)\prod_{j=1}^{m} \quad P(x_j | y(t_c) = 1, x_p(j))}{P(x, t \le t_f)}$$

Algorithm 1. Early-Stage Prediction (ESP) Framework: -

Require: Training data $D_n(t_c) = (x, y(t_c), T), t_f$

Output: Probability of event at time t_f

```
Phase 1: Conditional probability estimation at $t_c$
For j= 1,…….., m
    P($x_j|y(t_c) = 1$)
End
Phase 2: Predict probability of event occurrence at $t_f$
Fit AFT model to $D_n(t_c)$
P($y\left(t_f\right) = 1, t \le t_f) = F(t)$
For i= 1, ………., n
    Estimate P(y,( $t_f$)=1| $x_i$ , $t \le t_f$)
End
Return P(y($t_f$) = 1| $x, t \le t_f$)
```

5.1.3 ESP Bayesian Network (ESP-BN)

Using the data t_c up until now, we must first construct a network for the Bayesian network. A Bayesian network classifier will be trained using the Hill-climbing structure learning method. The next stage is to predict the likelihood that an event will occur at the conclusion of the research t_f, once we have learned the topology of the Bayesian network." We can use the various extrapolation strategies previously discussed for this purpose. As a result, the posterior probability estimation for the event's occurrence at time t_f is given by

$$P\left(x, t \le t_f\right) = \frac{F\left(t_f\right)\prod_{j=1}^{m} \quad P(x_j | y\left(t_f\right) = 1, \ Pa(x_j))}{P(x, t \le t_f)}$$

This indicates that, when compared to its underlying models, ESP enhances prediction performance without adding complexity.

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 $t_c$
$E_G \longleftarrow \emptyset, estimate\ P(G|\ D_n(t_c))$
$score_{final} \longleftarrow \infty, score = MDL(BN, D_n(t_c))$ (Eq. (5))
While $score_{final} > score$
    $score_{final} \longleftarrow\ score$
    For every add/remove/reverse $E_G$ on G
        Estimate P($G_{new}|D_n(t_c)$)
        $score_{new} = MDL(BN_{new}, D_n(t_c))$
    Select neetwork structure with minimium $score_{new}$
    If $score > score_{new}$
    $score \longleftarrow score_{new}, G \longleftarrow G_{new}$
Phase 2: Forcasting event occurrence at $t_f$
    Fit AFT model to $D_n(t_c)$
    P(y($t_f$) $= 1, t \leq t_f) = F(t)$
    For all I in $D_n(t)$
        Estimate p($y_i(t)|x_i$)
        Weibull using Eqs. (7),(16) and (18)
        Log logistic using Eqs. (7),(17) and (18)
    End for
    Return P(y($t_f$)=1|x, $t \leq t_f$)
```

19. ACCURACY CURVE

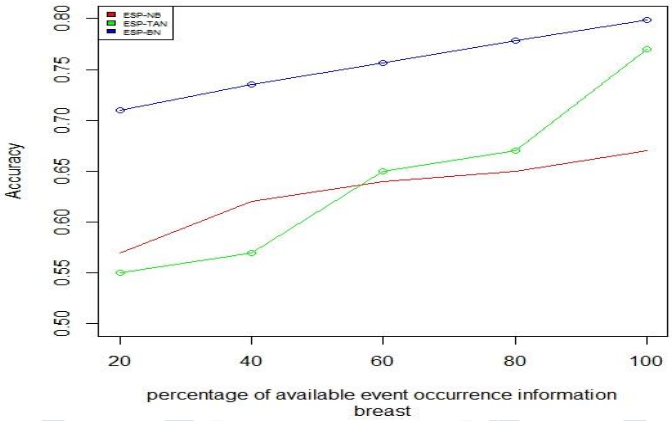
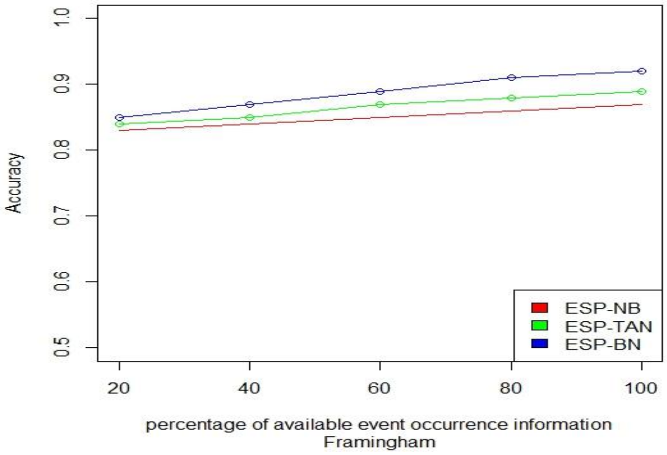


Fig 1: Colon Accuracy Curve



20. Fig 2: Framingham Accuracy curve

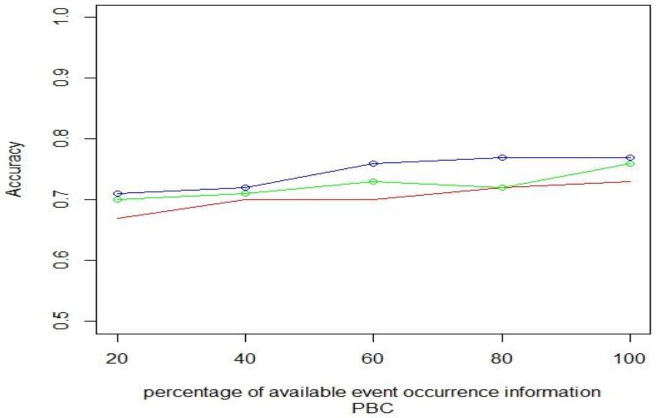


Fig 3: PBC Accuracy curve

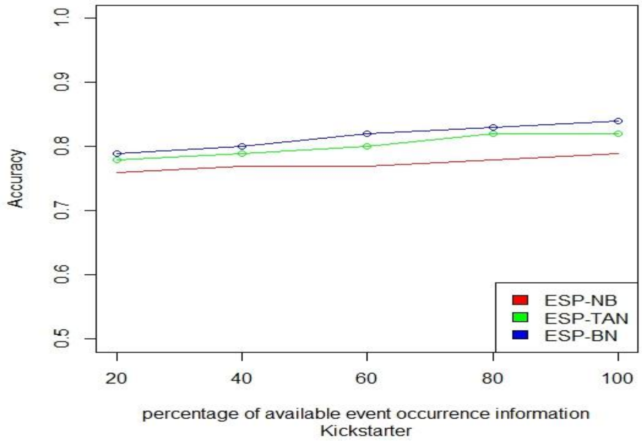


Fig 4: Kick starter Accuracy curve

AUC CURVE

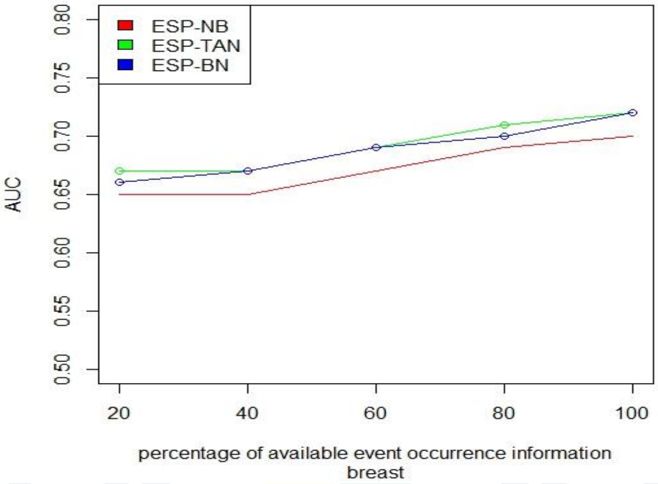
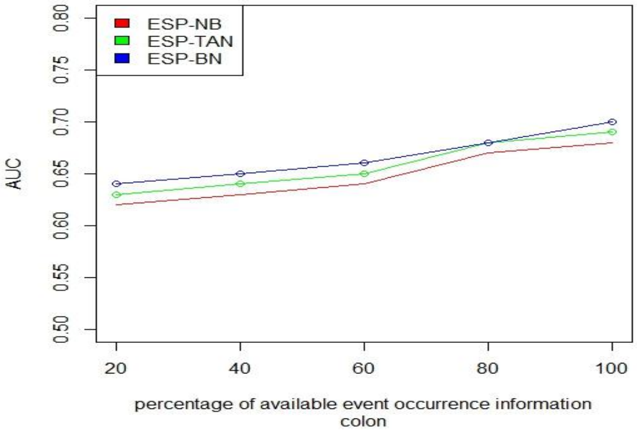


Fig 5 : Breast AUC curve

21.



22. **Fig 6: Colon AUC curve**

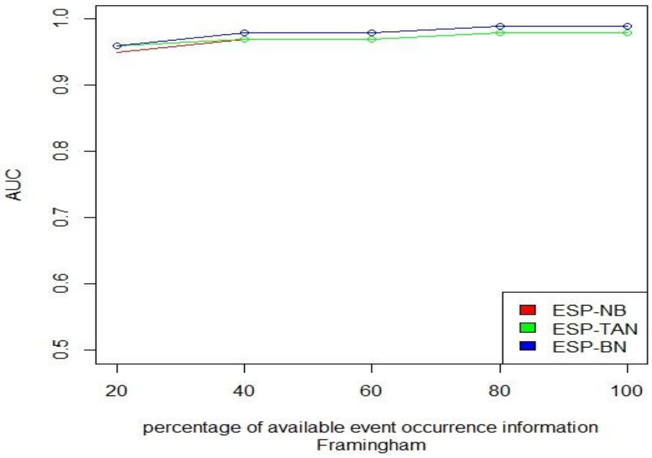


Fig 7: Framingham AUC curve

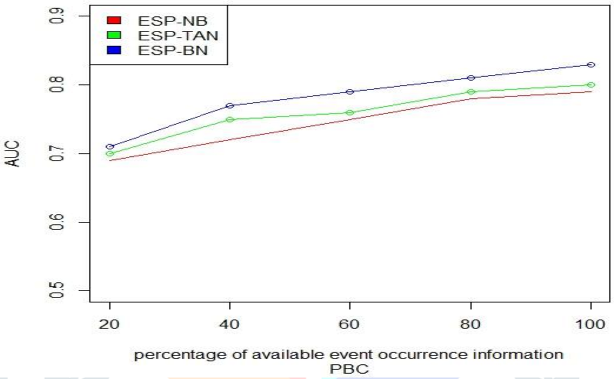


Fig 8: PBC AUC curve

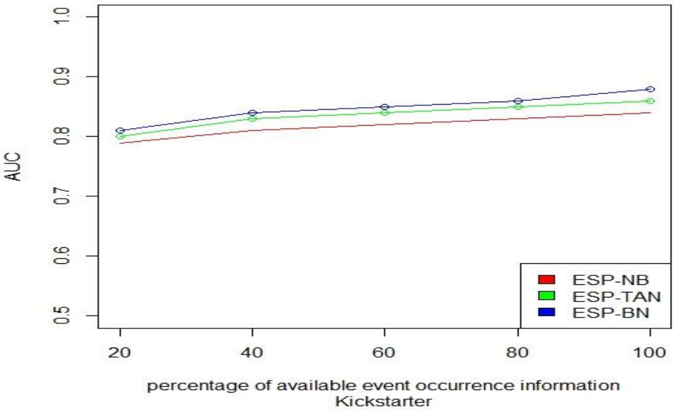


Fig 9: Kick starter AUC

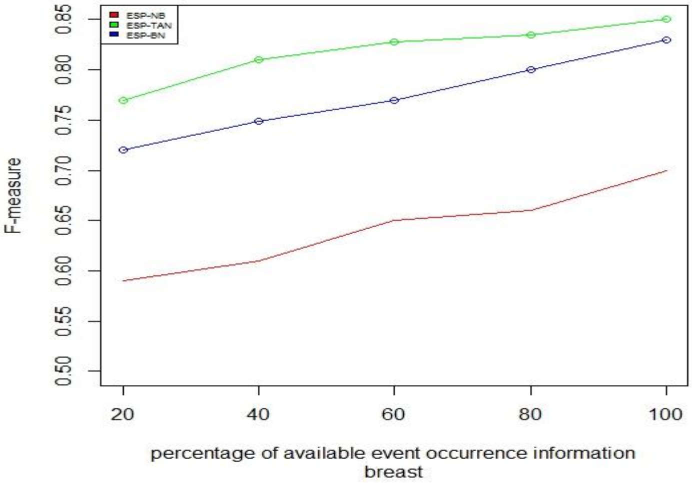
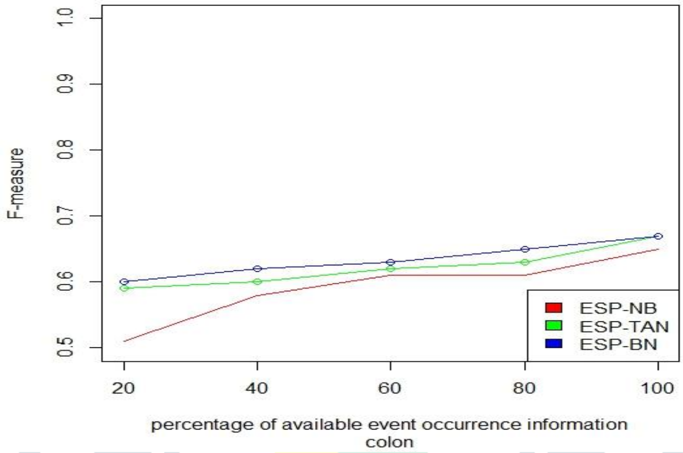


Fig 10 F-Measure Curve Figure



23. Fig 11: Colon F-measure curve

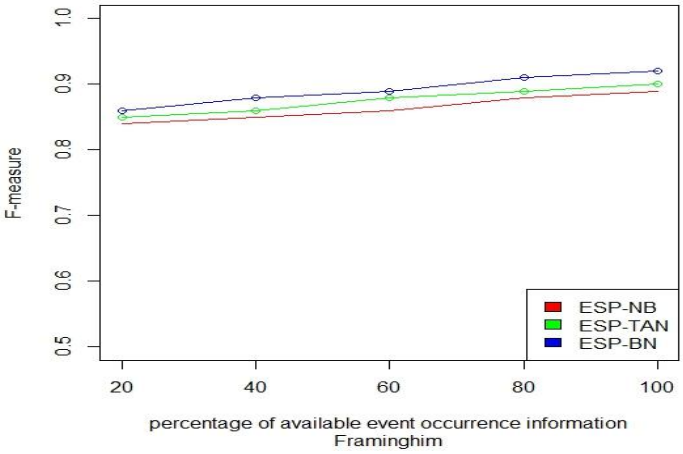
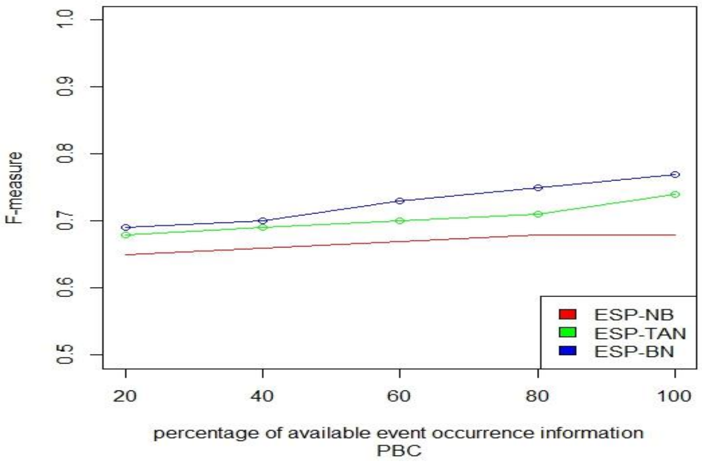
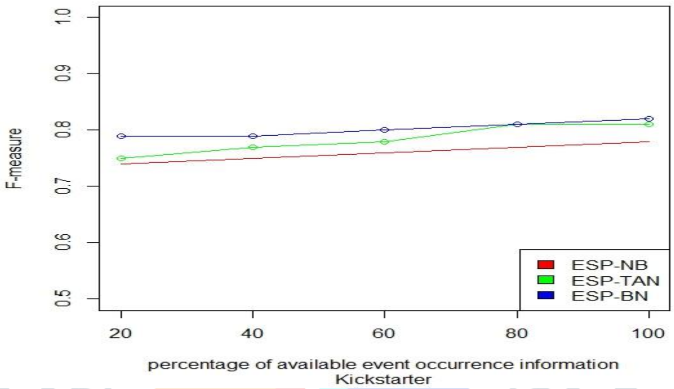


Fig 12: Framingham F- measure curve



24. **Fig 13: PBC F-measure curve**

25.



\

Fig 14: Kick starter F-measure curve

CONCLUSION

In many real-world application fields, it is essential to make predictions about the future based solely on the information acquired at the start of longitudinal studies. By fitting an event data set with sparse early occurrences to a statistical distribution of time, we developed a new framework for early stage event prediction in this study. One of the common aspects of longitudinal data is the existence of censored instances, which are situations where the outcome is unclear after a certain period of time has passed throughout the inquiry. By using the Kaplan-Meier estimator to determine the likelihood of an occurrence and the likelihood of being censored, we developed a new method to manage such censored data instead of deleting it. The main objective of this work is to demonstrate that, using the available (limited) data on event occurrence, it is possible to create predictions that will be more accurate by the conclusion of the research period. Since it takes time to collect enough training data about event occurrence, this is essential in longitudinal survival research. The suggested ESP-based model changes prior probabilities of event occurrence by using Weibull and Log-logistic distributions to suit time-to-event information. We developed three brand-new Bayesian algorithms using this technique, using the training data obtained at the start of the investigation, to precisely predict the event occurrence for upcoming time points. Our exhaustive evaluations on both synthetic and real datasets demonstrate that the proposed ESP-based algorithms outperform the widely used Cox model and other well-liked classification strategies in estimating occurrences at future time points.

REFERENCES

1. C. Chatfield, Time-Series Forecasting. CRC Press, 2000.
2. 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.
3. G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," Neurocomputing, vol. 50, pp. 159–175, 2003.
4. 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.
5. 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.
6. O. Chapelle, B. Scholkopf, and A. Zien, ¨ Semi-supervised learning. MIT press Cambridge, 2006, vol. 2.
7. Z. Zhou and M. Li, "Semi-supervised regression with co-training." in IJCAI, 2005, pp. 908–916.
8. L. Gordon and R. Plshen, "Tree-structured survival analysis," Cancer Treat Reports, vol. 69, no. 10, pp. 1065–1074, 1985.
9. M. R. Segal, "Regression Trees for Censored Data," Biometrics, vol. 44, no. 1, pp. 35–47, 1988.
10. 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.
11. 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.
12. K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in Advances in Neural Information Processing Systems. MIT Press, 1998, pp. 368–374.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.