A Critical Review: Deep Survival Analysis

Paper by: Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei
Review by: Naveen Arivazhagan, Jared Dunnmon, Carson Lam, Priyanka Nigam, Darvin Yi

Introduction

Electronic health record (EHR) data includes a variety of information such as patient demographics, medical history, prescriptions issued, test results, and vital signs. The increased availability of EHR data has provided the opportunity to gain insights into disease progression and treatment outcomes, which may help in identifying high-risk patients, designing more individualized treatment plans, and earlier diagnosis. In this paper, the authors focus on the task of survival analysis and propose a new hierarchical generative model, called deep survival analysis, that uses deep exponential families to predict patient risk for some event of interest as well as the time to that event. Their method allows the modelling of observations and covariates jointly conditioned on a rich latent structure, does not need manual selection of an appropriate start time, and can handle heterogeneous or missing data that is common in EHR.

Such forecasts and risk estimates can be used by clinicians to allocate resources appropriately and also to take appropriate interventions. Risk scores are used to aid in making treatment decisions for several diseases including coronary heart disease (CHD), breast cancer, and prostate cancer. In this paper, authors focus on CHD, both because it is a leading cause of death and because there are several effective preventative treatment options, including antiplatelet therapy, statins, and lifestyle changes. Many of these treatment options have been successful in preventing both a first CHD event and subsequent CHD events following a primary one. Thus better CHD risk analysis may aid in treatment planning and CHD event prevention.

Related Work

The authors compare their model to traditional risk score approaches. The traditional approach models the time to the event of interest from a standard starting point and requires data in a specific format – pairs describing a time and whether an observation is either uncensored, in which case the time corresponds to exactly when the event occurred, or censored, in which case the time indicates that the event occurs later than this time. Such time-of-occurrence data is modelled as being drawn from some unknown distribution that is estimated using various methods.

Traditional methods estimate risk scores by regressing over covariates and have several disadvantages. First, regression based models require complete measurement of all the covariates both at training time and during inference; EHR data, which encompasses many more measurements than traditional methods would utilize, suffers from sparse and missing data points. Second, the regression model must be linear over the covariates, as such frameworks are ill-suited to take into account combinatorially many terms that emerge when considering non-linear relationships. Finally, traditional methods require carefully curated data, wherein all patient records are aligned by some synchronization event such as start of clinical trial, intervention date, or condition onset date. Such an event is generally not available in EHR data that is collected at many arbitrary points in a patient's lifetime.

The standard risk score for CHD is the Framingham risk score, which was introduced in 1976 (Kennel et al. 1976) and included age, gender, smoking, systolic blood pressure, total cholesterol, ECG measure of left ventricular hypertrophy, and diabetes. This risk score has been

updated to include systolic and diastolic blood pressure, more advanced measures of diabetes, and the ratio of total cholesterol to HDL cholesterol (Anderson et al. 1991, Wilson et al. 1998).

Methods

The Weibull distribution, commonly used in survival analysis, is utilized to model the time elapsed before a relevant event occurs. The likelihood of an uncensored observation is defined as the probability mass assigned before the event, and the likelihood of a censored observation is defined as the probability mass assigned to the interval after the observation. The Weibull distribution is parametrized by λ and k. k, which determines the shape of the curve, is considered to be a fixed hyper-parameter that is manually tuned, while λ is the softplus of a linear combination of the latent variables sampled from a deep exponential distribution. Note that a and b are drawn from Gaussian distributions.

$$p_{weibull}(t) = (k/\lambda)(t/\lambda)^{k-1}exp(-(t/\lambda)^k)$$
$$\lambda = log(1 + exp(z^T_n\, a + b))$$

x^i, the observed covariates, are assumed to be independent of each other conditioned on the latent variables z_n. x_i are considered to be of two types: 1) real valued for continuous data, 2) binary, zero or nonzero, for count based data. These are approximated by the student-t distribution and bernoulli distribution respectively.

$$x^i_n \sim p(\cdot\,|\beta^i, z_n)$$

The paper proposes a hierarchical generative model based on deep exponential families to model the latent variables, z_n. A Deep Exponential Family (DEF) has L hidden layers that are K dimensional per observation. By virtue of being fully Bayesian, it is able to deal with missing data. Further, DEFs specifically have been shown to provide improvements in modelling missing data in other similar contexts, including text modeling, recommender systems, and images (Ranganath et al., 2015b). The multi-layer model is also able to incorporate non-linear interactions as it stacks the latent variables into multiple layers.

$$z_n \sim DEF(W)$$

Given the covariates, predictions are made from the posterior. However, since solving this is intractable, it is instead approximated by the mean-field family and variational inference methods are used.

$$p(t|x) = \int_z p(t|z)p(z|x)dz$$

Dataset

The authors compare the results of survival analysis using the baseline CHD risk score, the Framingham CHD risk score, to using the deep analysis model on a dataset of 313,000 adult patients with at least one data point for vitals, laboratory tests, medications, and diagnosis codes in at least 5 months. The EHR data contains is collected from varied settings, including inpatient, outpatient, and emergency room visits. The data includes 9 vital signs, 79 laboratory test measurements, 5,262 medication orders, and 13,153 diagnosis codes.

This dataset was split 84-8-8 into training, validation, and test sets. This data was preprocessed to bin all observations by the month in which they occur. Continuous data was binned by computing the average of this data over the entire month, and discrete data was binned using an indicator for whether it was present during that month or not. This dataset has

a significant amount of missing information, as only 11.8% of patients have data for an entire month and only 1.4% of the months have all of the data required in the baseline CHD risk score.

Results

The authors use a metric known as concordance index to evaluate model performance. It is the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered, i.e. all uncensored pairs. Their results show that the deep survival model outperforms the baseline CHD risk score, which obtains a concordance score of 65.57%, for all dimensionalities of the hidden layers in the deep exponential family model. The best tuned Deep Survival Analysis model using a deep exponential family has a hidden layer dimensionality of 50, and obtains a score of 73.11%.

Critique

The authors evaluate their model based only on concordance index. Concordance index only measures how accurate the ordering of survival times is and does not factor in whether the individual times were predicted accurately. Due to this, the evaluation and comparison to the baselines is still incomplete.

The authors analyze the best deep survival model to determine which features provide the most predictive power by computing the likelihood of the data with all other features hidden. They found that the model using only diagnosis codes yielded the highest predictive likelihood. However it seems like this is to be expected as the diagnosis is based on other types of data including vitals and laboratory tests, and medications prescribed are based on the diagnosis (Harrell et al. 1982). Thus diagnosis information likely provides a good summary of the other data.

The DEF model is trained using a population dataset vastly differently from the population studied in the Framingham Heart study, the study from which the Framingham CHD risk score is derived (Washington Heights NY., Framingham MA.), yet tested against a subset of the model's original dataset (Dawber et al. 1963). It is unlikely that the differences are captured by diagnosis codes alone. A more objective test would be to compare against a third dataset or at least try to match the dataset with respect to demographics.

The model also makes forecasts based only on the observations made at a single point of time and does not factor in the full patient history. This full-history data that contains the evolution of various parameters through time could help in making more accurate predictions.

Lastly, the proposed model still does not handle the rich textual and image data collected over a patient's history. The task may however benefit if this multi-modal information could be incorporated through vector-based distributed representations characteristic of deep learning.

Conclusion

The authors develop a method that leverages the easily available EHR data to develop a method to forecast risk of coronary heart disease. Earlier methods relied on carefully curated task specific datasets, since EHR data suffered from problems such as missing data, data sparsity, heterogeneous data, and no common start point across which different cases could be compared. The fully Bayesian, hierarchical generative model based approach developed here is able to deal with these problems and achieves a new state of the art result in survival analysis.

References

1. Kannel WB, McGee D, and Gordon T (1976) A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 38: 46-51.
2. Anderson KM, Wilson PWF, Odell PM, Kannel WB (1991) An updated coronary risk profile. A statement for health professionals. *Circulation* 83: 356-362.
3. Wilson, PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* 97: 1837-1847.
4. R Ranganath, L Tang, L Charlin, and DM Blei. Deep exponential families. In Proceedings AISTATS (International Conference on Artificial Intelligence and Statistics).
5. FE Harrell, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. JAMA, 247(18):2543–2546, 1982.
6. TR Dawber, WB Kannel, LP Lyell. An approach to longitudinal studies in a community: the Framingham Study. *Ann N Y Acad Sci.* 1963;107:539–556.