
Personalized Survival Prediction with Contextual Explanation Networks

Maruan Al-Shedivat
Carnegie Mellon University
alshedivat@cs.cmu.edu

Avinava Dubey
Carnegie Mellon University
akdubey@cs.cmu.edu

Eric P. Xing
Carnegie Mellon University
epxing@cs.cmu.edu

Abstract

Accurate and transparent prediction of cancer survival times on the level of individual patients can inform and improve patient care and treatment practices. In this paper, we design a model that concurrently learns to accurately predict patient-specific survival distributions and to explain its predictions in terms of patient attributes such as clinical tests or assessments. Our model is flexible and based on a recurrent network, can handle various modalities of data including temporal measurements, and yet constructs and uses simple explanations in the form of patient- and time-specific linear regression. For analysis, we use two publicly available datasets and show that our networks outperform a number of baselines in prediction while providing a way to inspect the reasons behind each prediction.

1 Introduction

In survival analysis, the goal is to estimate the occurrence time and the risk of an unfavorable event in the future (e.g. death of a patient) that can inform our decisions at present time (e.g., help to select a treatment). The classical models for this task are the Aalen’s additive model [1] and the Cox’s proportional hazard model [2], which linearly regress attributes of a patient to the hazard function. While suitable for comparing populations of patients, these models were not designed for patient-specific prediction. By reformulating survival analysis as a multi-task classification problem, Yu et al. [3] show that a set of temporally ordered linear classifiers provides much more accurate predictions.

Here, we follow the same classification approach and show that using deep learning methods further improves predictive performance on survival data. While promising, straightforward use of neural networks leads to black-box predictors that lack transparency offered by the linear models. To overcome this issue, we employ *contextual explanation networks* [CEN, 4]—a class of models that learn to predict by generating and leveraging intermediate explanations. Explanations here are defined as instance-specific simple (linear) models that not only help to interpret predictions but are selected by the network to make predictions for each patient at each time interval. CENs can be based on arbitrary deep architectures and can process a variety of input data modalities while interpreting predictions in terms of selected attributes. As we demonstrate in experiments, this approach attains both the best performance as well as interpretability.

2 Background

First, we present the setup used by Yu et al. [3]. The data is represented by patient-specific attributes, \mathbf{X} , and the times of the occurrence of event, \mathbf{T} . These times are converted into m -dimensional binary vectors, $\mathbf{Y} := (y^1, \dots, y^m)$, that indicate the corresponding follow up time. If the death occurred at time $t \in [t_i, t_{i+1})$, then $y^j = 0, \forall j \leq i$ and $y^k = 1, \forall k > i$. If the data point was censored (i.e.,

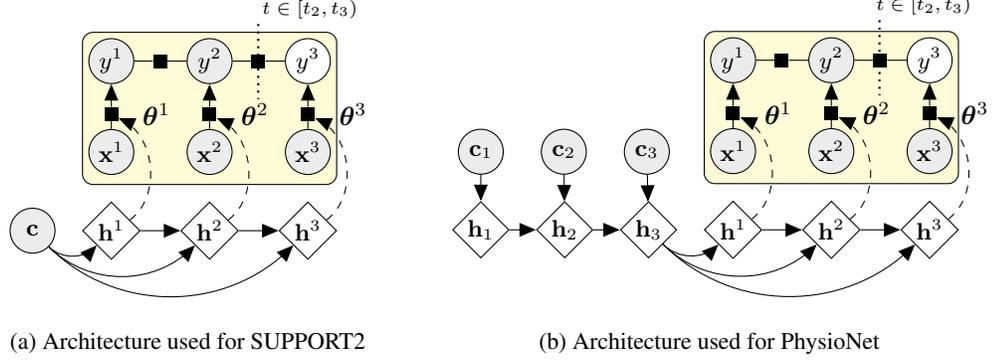


Figure 1: CEN architectures used in our survival analysis experiments. Context encoders were time-distributed single hidden layer MLP (a) and LSTM (b) that produced inputs for another LSTM over the output time intervals (denoted with \mathbf{h}^1 , \mathbf{h}^2 , \mathbf{h}^3 hidden states respectively). Each hidden state of the output LSTM was used to generate the corresponding θ^t that were further used to construct the log-likelihood for CRF.

we lack information for times after $t \in [t_i, t_{i+1})$, the targets (y^{i+1}, \dots, y^m) are regarded as latent variables. Note that only $m + 1$ sequences are valid, i.e., assigned non-zero probability by the model, which allows to write the following linear model:

$$p(\mathbf{Y} = (y^1, \dots, y^m) \mid \mathbf{x}, \Theta) = \frac{\exp(\sum_{t=1}^m y^t \mathbf{x}^\top \theta^t)}{\sum_{k=0}^m \exp(\sum_{t=k+1}^m \mathbf{x}^\top \theta^t)} \quad (1)$$

The model is trained by optimizing a regularized log likelihood w.r.t. $\Theta := \{\theta^t\}_{t=1}^m$. After training, we get a set of linear models, one for each time interval, used for predicting the survival probability.

3 Contextual Explanation Networks for Survival Analysis

Here, we take the same structured prediction approach but consider a slightly different setup. In particular, we assume that each data instance (patient record) is represented by three variables: the *context*, \mathbf{C} , the *attributes*, \mathbf{X} , and the *targets*, \mathbf{Y} . Our goal is to learn a model, $p_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{X}, \mathbf{C})$, parametrized by \mathbf{w} that can predict \mathbf{Y} from \mathbf{X} and \mathbf{C} . Note that inputs have two representations, \mathbf{X} and \mathbf{C} , where \mathbf{X} is a set of attributes that will be used to interpret predictions¹. Contextual explanation networks (CENs) are defined as models that assume the following form:

$$\mathbf{Y} \sim p(\mathbf{Y} \mid \mathbf{X}, \theta), \quad \theta \sim p_{\mathbf{w}}(\theta \mid \mathbf{C}), \quad p_{\mathbf{w}}(\mathbf{Y} \mid \mathbf{X}, \mathbf{C}) = \int p(\mathbf{Y} \mid \mathbf{X}, \theta) p_{\mathbf{w}}(\theta \mid \mathbf{C}) d\theta \quad (2)$$

where $p(\mathbf{Y} \mid \mathbf{X}, \theta)$ is a predictor parametrized by θ . Such predictors are called *explanations*, since they explicitly relate interpretable variables, \mathbf{X} , to the targets, \mathbf{Y} . The conditional distribution $p_{\mathbf{w}}(\theta \mid \mathbf{C})$ is called the *context encoder* processes the context representation, \mathbf{C} , and generates parameters for the explanation, θ .

For survival analysis, we want explanations to be in the form of linear CRFs as given in (1). Hence, our contextual networks with CRF-based explanations are defined as follows:

$$\begin{aligned} \theta^t &\sim p_{\mathbf{w}}(\theta^t \mid \mathbf{C}), \quad t \in \{1, \dots, m\}, \quad \mathbf{Y} \sim p(\mathbf{Y} \mid \mathbf{X}, \theta^{1:m}), \\ p(\mathbf{Y} = (y^1, y^2, \dots, y^m) \mid \mathbf{x}, \theta^{1:m}) &\propto \exp \left\{ \sum_{t=1}^m y^t (\mathbf{x}^\top \theta^t) + \omega(y^t, y^{t+1}) \right\} \quad (3) \\ p_{\mathbf{w}}(\theta^t \mid \mathbf{C}) &:= \delta(\theta^t, \phi_{\mathbf{w}, \mathbf{D}}^t(\mathbf{c})), \quad \phi_{\mathbf{w}, \mathbf{D}}^t(\mathbf{c}) := \alpha(\mathbf{h}^t)^\top \mathbf{D}, \quad \mathbf{h}^t := \text{RNN}(\mathbf{h}^{t-1}, \mathbf{c}) \end{aligned}$$

A few things to note here. First, the model generates explanations for each patient and for each time interval. Second, depending on the nature of the context representation, \mathbf{C} , CENs process it

¹It is common to have the data to be of multiple representations some of which are low-level or unstructured (e.g., image pixels, sensory inputs), and other are high-level or human-interpretable (e.g., categorical variables). To ensure interpretability, we would like to use deep networks to process the low-level representation (the *context*) and construct explanations as *context-specific probabilistic models* on the high-level features.

and generate θ^t for each time step using a recurrent encoder (Figure 1). We use a deterministic RNN-based encoder, ϕ^t , that selects parameters for explanations from a global dictionary, \mathbf{D} , using soft attention (for details on dictionary-based context encoding, see [4]). Finally, the potentials between attributes, \mathbf{x} , and targets, $y^{1:m}$, are linear functions parameterized by $\theta^{1:m}$; the pairwise potentials between targets, $\omega(y_i, y_{i+1})$, ensure that configurations ($y_i = 1, y_{i+1} = 0$) are improbable (i.e., $\omega(1, 0) = -\infty$ and $\omega(0, 0) = \omega_{00}, \omega(0, 1) = \omega_{01}, \omega(1, 1) = \omega_{10}$ are learnable parameters). Given these constraints, the likelihood of an uncensored event at time $t \in [t_j, t_{j+1})$ is

$$p(T = t | \mathbf{x}, \Theta) = \exp \left\{ \sum_{i=j}^m \mathbf{x}^\top \theta^i \right\} / \sum_{k=0}^m \exp \left\{ \sum_{i=k+1}^m \mathbf{x}^\top \theta^i \right\} \quad (4)$$

and the likelihood of an event censored at time $t \in [t_j, t_{j+1})$ is

$$p(T \geq t | \mathbf{x}, \Theta) = \sum_{k=j+1}^m \exp \left\{ \sum_{i=k+1}^m \mathbf{x}^\top \theta^i \right\} / \sum_{k=0}^m \exp \left\{ \sum_{i=k+1}^m \mathbf{x}^\top \theta^i \right\} \quad (5)$$

The joint log-likelihood of the data consists of two parts:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \Theta) = \sum_{i \in \text{NC}} p(T = t_i | \mathbf{x}_i, \Theta) + \sum_{j \in \text{C}} p(T > t_j | \mathbf{x}_j, \Theta) \quad (6)$$

where NC is the set of non-censored instances (for which we know the outcome times, t_i) and C is the set of censored instances (for which only know the censorship times, t_j). The objective is optimized using stochastic gradient descent. See [4] for more details.

4 Experiments

In our experiments, we consider the datasets, models, and metrics as described below. We compare CENs with a number of baselines quantitatively as well as visualize the learned explanations.

Datasets. We use two publicly available datasets for survival analysis of of the intense care unit (ICU) patients: (a) SUPPORT2², and (b) data from the PhysioNet 2012 challenge³. The data was preprocessed and used as follows:

- **SUPPORT2:** The data had 9105 patient records and 73 variables. We selected 50 variables for both \mathbf{C} and \mathbf{X} features. Categorical features (such as `race` or `sex`) were one-hot encoded. The values of all features were non-negative, and we filled the missing values with -1. For CRF-based predictors, the survival timeline was capped at 3 years and converted into 156 discrete intervals of 7 days each. We used 7105 patient records for training, 1000 for validation, and 1000 for testing.
- **PhysioNet:** The data had 4000 patient records, each represented by a 48-hour irregularly sampled 37-dimensional time-series of different measurements taken during the patient’s stay at the ICU. We resampled and mean-aggregated the time-series at 30 min frequency. This resulted in a large number of missing values that we filled with 0. The resampled time-series were used as the context, \mathbf{C} , while for the attributes, \mathbf{X} , we took the values of the last available measurement for each variable in the series. For CRF-based predictors, the survival timeline was capped at 60 days and converted into 60 discrete intervals.

Models. For baselines, we use the classical Aalen and Cox models and the CRF from [3], where all used \mathbf{X} as inputs. Next, we combine CRFs with neural encoders in two ways:

- (i) We apply CRFs to the outputs from the neural encoders (denoted MLP-CRF and LSTM-CRF, all trainable end-to-end). Similar models have been show very successful in the natural language applications [5]. Note that parameters of the CRF layer assign weights to the latent features and are no longer interpretable in terms of the attributes of interest.
- (ii) We use CENs with CRF-based explanations, that process the context variables, \mathbf{C} , using the same neural networks as in (i) and output parameters for CRFs that act on the attributes, \mathbf{X} .

²<http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets>.

³<https://physionet.org/challenge/2012/>.

Table 1: Performance of the classical Cox and Aalen models, CRF-based models, and CENs that use LSTM or MLP for context embedding and CRF for explanations. The numbers are averages from 5-fold cross-validation; the std. are on the order of the least significant digit. @K denotes the temporal quantile, i.e., the time point such that K% of the patients in the data have died or were censored before that point.

SUPPORT2					PhysioNet Challenge 2012				
Model	Acc@25	Acc@50	Acc@75	RAE	Model	Acc@25	Acc@50	Acc@75	RAE
Cox	84.1	73.7	47.6	0.90	Cox	93.0	69.6	49.1	0.24
Aalen	87.1	66.2	45.8	0.98	Aalen	93.3	78.7	57.1	0.31
CRF	84.4	89.3	79.2	0.59	CRF	93.2	85.1	65.6	0.14
MLP-CRF	87.7	89.6	80.1	0.62	LSTM-CRF	93.9	86.3	68.1	0.11
MLP-CEN	85.5	90.8	81.9	0.56	LSTM-CEN	94.8	87.5	70.1	0.09

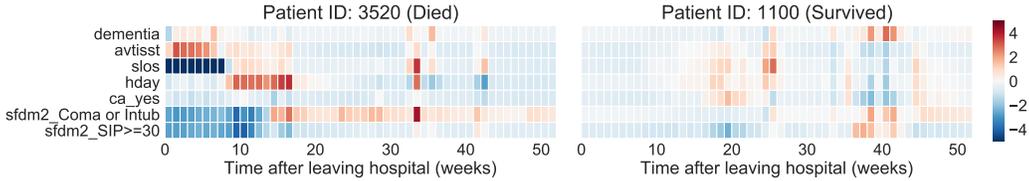


Figure 2: Weights of the CEN-generated CRF explanations for two patients from SUPPORT2 dataset for a set of the most influential features: *dementia* (comorbidity), *avtisst* (avg. TISS, days 3-25), *slos* (days from study entry to discharge), *hday* (day in hospital at study admit), *ca_yes* (the patient had cancer), *sfdm2_Coma or Intub* (intubated or in coma at month 2), *sfdm2_SIP* (sickness impact profile score at month 2). Higher weight values correspond to higher feature contributions to the risk of death after a given time point.

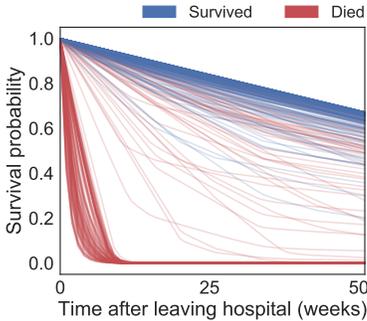


Figure 3: CEN-predicted survival curves for 500 random patients from SUPPORT2 test set. Color indicates death within 1 year after leaving the hospital.

Metrics. Following Yu et al. [3], we use two metrics specific to survival analysis: (a) accuracy of correctly predicting survival of a patient at times that correspond to 25%, 50%, and 75% population-level temporal quantiles (i.e., time points such that the corresponding percentage of the patients in the data had their time of the last follow up prior to that due to censorship or death) and (b) the relative absolute error (RAE) between the predicted and actual time of death for non-censored patients.

Quantitative results. The results for all models are given in Table 1. Our implementation of the CRF baseline reproduces (and even slightly improves) the performance reported by Yu et al. [3]. CRFs built on representations learned by deep networks (MLP-CRF and LSTM-CRF models) improve upon the plain CRFs but, as we noted, can no longer be interpreted in terms of the original variables. On the other hand, CENs outperform neural CRF models on certain metrics (and closely match on the others) while providing explanations for the survival probability predictions for each patient at each point in time.

Qualitative results. To inspect predictions of CENs qualitatively, for any given patient, we can visualize the weights assigned by the corresponding explanation to the respective attributes at each time interval. Figure 2 shows explanation weights for a subset of the most influential features for two patients from SUPPORT2 dataset who were predicted as survivor and non-survivor. These explanations allow us to better understand patient-specific temporal dynamics of the contributing factors to the survival rates predicted by the model (Figure 3). This information can be used for model diagnostics (i.e., help us understand whether we can trust a particular prediction) and as more fine-grained information useful for decision support.

References

- [1] O.O. Aalen. “A linear regression model for the analysis of life time”. In: *Statistics in Medicine*, 8(8):907–925 (1989).
- [2] DR Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1972), pp. 187–220.
- [3] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin & Vickie Baracos. “Learning patient-specific cancer survival distributions as a sequence of dependent regressors”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 1845–1853.
- [4] Maruan Al-Shedivat, Avinava Dubey & Eric P Xing. “Contextual Explanation Networks”. In: *arXiv preprint arXiv:1705.10301* (2017).
- [5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu & Pavel Kuksa. “Natural language processing (almost) from scratch”. In: *Journal of Machine Learning Research* 12.Aug (2011).