Marked Point Process for Severity of Illness Assessment

Kazi T. Islam KAZI.ISLAM@EMAIL.UCR.EDU
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California 92507, USA

Christian R. Shelton CSHELTON@CS.UCR.EDU
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California 92507, USA

Juan I. Casse JCASSE@CS.UCR.EDU
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California 92507, USA

Randall Wetzel RWETZEL@CHLA.USC.EDU
Laura P. and Leland K. Whittier Virtual Pediatric Intensive Care Unit
Childrens Hospital LA
Los Angeles, CA 90089, USA

Abstract

Electronic Health Records (EHRs) consist of sparse, noisy, incomplete, heterogeneous and unevenly sampled clinical data of patients. They include physiological signals, lab test results, procedural events, clinical notes. Such data can be treated as a temporal stream of events of varied types occurring at irregularly spaced time points. We focus on modeling the temporal dependencies that arise due to the types, timings, and values of different events in such data. We model the event streams, including vital signs, laboratory results contained in two different datasets (MIMIC III — Medical Information Mart for Intensive Care clinical database — and data extracted from EHRs of patients in a tertiary pediatric intensive care unit) using a piecewise-constant conditional intensity model (PCIM), a type of marked point process. Our experiments capture meaningful temporal dependencies and show improvement in hospital mortality prediction over traditional ICU scoring systems.

1. Introduction

Electronic Health Records (EHRs) contain detailed records of patients symptoms, demographics, outcomes, and other encounter information. In this paper, we concentrate on records from stays in intensive care units (ICUs), because they provide an intense, well monitored, and (relatively) short duration episodes. These attributes allow for the collection of large number of patient trajectories with definite outcomes. We expect many of the modeling lessons learned in this setting to be transferable to other areas of patients' medical records.

We view the EHR data of a patient as a timeline of events happening throughout the patient's stay in the ICU, where an event for an individual patient is a new measurement

of a particular physiological variable, a new lab test, a dosage of a particular drug, or a procedure performed, started, or stopped. The measurements associated with such events are indicative of patient states and the subject of constant monitoring from clinicians. But, the temporal dynamics of the stream of events carry extra information for event forecasting and understanding severity of illness. For example, abnormal values for blood pressure could indicate critical states of a patient, but a higher measurement *rate* could also indicate high degree of urgency. Higher values of a certain physiological variable could instigate the clinicians to order a particular lab test. High values of that lab result could result in ordering of a related lab result, or initiate the dosage of a particular drug. These complex temporal dependencies that arise due to time, value, and types of different events throughout the timeline of a patient's stay in an ICU could be vital in understanding the temporal dynamics of the patient's state(e.g. severity of illness). These sort of data are not missing-at-random (or, more accurately, "measured-at-random").

One challenge in modeling medical data is the irregular arrival of different events. Not only are the temporal patterns of measurements indicative of the patients' conditions, but those patterns do not correspond to regular sampling intervals and are dependent on the previous history. Therefore, in this paper we model the timing information in the continuous-time domain. As a result, no fixed sampling rate had to be chosen. Also, the heterogeneous nature of the data requires a model that can capture temporal dependencies between events of varied types. These properties of the complex EHR data stream has led to our choice of using a non-Markovian marked point process model named PCIM (Gunawardana et al., 2011).

Our general goal is to explore how modeling the temporal dynamics of the raw EHR data stream in continuous time domain could facilitate certain critical decision making tasks performed in an ICU. To achieve that goal, we have performed mortality prediction in the ICU using the first 24 hours of ICU data from two different datasets, including the publicly available MIMIC-III database (Johnson et al., 2016).

Severity of Illness (SOI) assessment and mortality modeling is a broad area of research in health informatics. SOI and mortality scores are used to predict patient outcome and often used as the principal measures of quality-of-care comparison among ICUs and stratification for clinical trials. Scoring systems like SAPS-I (Simplified Acute Physiology Score) (Le Gall et al., 1984), SAPS-II (Le Gall et al., 1993), PIM2 (Pediatric Index of Mortality) (Slater et al., 2003), and PRISM3 (Pollack et al., 1996) are the accepted current practices in ICU acuity scoring, but are based on static snapshots of certain clinical variables over a patient's stay in the ICU and ignore the rapidly evolving temporal dynamics of the clinical variables. Several works in the literature have focused on boosting the predictive power of such scoring systems using rich clinical information present in an EHR such as, clinical notes (Ghassemi et al., 2014; Lehman et al., 2012). While the temporal dynamics of the topics derived from the notes using traditional topic modeling techniques are useful (Jo et al., 2015; Ghassemi et al., 2015) for patient risk assessment, the temporal information present in patient trajectories of raw vital signs or lab measurements has not been extensively studied. We conclude that temporal modeling of clinical variables including vital signs and lab tests can complement the standard scoring systems and improve mortality prediction performance.

In this paper we make the following contributions.

- We model the temporal dependency in streams of measurement data using piecewise-constant conditional intensity models (PCIMs).
- For modeling the measurement values we extend the PCIM model to allow marks that contain continuous-values, in addition to the discrete-valued labels.
- From the learned models, we assign a risk score to each patient trajectory with vital signs and lab tests and use it as a mortality predictor.

2. Methods

As the data are irregularly sampled, we modeled the timings of the measurements in the continuous-time domain. We discuss the marked point process models used in the next subsection.

Outside of health care, temporal event sequences has been studied in other areas including genetics (Friedman et al., 2000), data center management (Oliner and Stearley, 2007), neuroscience (Truccolo et al., 2005), and web search query logs. Event streams can be modeled in either discrete or continuous time. Discrete time approaches, such as Hidden Markov Models (HMMs) (Rabiner, 1989; Leonard E. Baum, 1966) and Dynamic Bayesian Networks (DBNs) (Dean and Kanazawa, 1988), require discretizing event times to a fixed sampling rate. The irregular sampling property of the EHR data makes the discrete approach less appealing. Recent approaches in modeling continuous time processes include Continuous Time Bayesian Networks (CTBNs) (Nodelman et al., 2002), Continuous Time Noisy-Or (CT-NOR) (Simma et al., 2012), Poisson Cascades (Simma and Jordan, 2012), Poisson Networks (Rajaram et al., 2005; Truccolo et al., 2005), and piecewise-constant conditional intensity models (PCIMs) (Gunawardana et al., 2011). We chose the last, PCIM, which is a class of marked point process, due to its flexible representation of the rate function as a decision tree. This promotes an interpretable and concise representation of the temporal dependencies.

2.1 Piecewise-Constant Conditional Intensity Model

Assume events are drawn from a finite label set L, representing the different event types. An event can then be represented by a pair: a time stamp t and a label $l \in L$. An event sequence x is $\{(t_i, l_i)\}_{i=1}^{n}$, where $0 < t_1 < \ldots < t_n$. We use $h_i = \{(t_j, l_j) | (t_j, l_j) \in x, t_j < t_i)\}$ for the history of event i. Let $t(y)$ for an event sequence y be the time of the last event in y, such that $t(h_i) = t_{i-1}$.

A conditional intensity function $\lambda(t|x)$ associated with a temporal point process is the expected instantaneous rate at which events are expected to occur at time t given the history before t. A conditional intensity model (CIM) is a set of such non-negative conditional intensity functions indexed by the labels $\{\lambda_l(t|x, \theta)\}_{l \in L}$. The likelihood of event sequence x can then be written as

$$p(x|\theta) = \prod_{l \in L} \prod_{i=1}^{n} \lambda_l(t_i|h_i; \theta)^{\mathbf{1}_l(l_i)} e^{-\Lambda_l(t_i|h_i;\theta)}, \tag{1}$$

where $\Lambda_l(t|h;\theta) = \int_{t(h)}^{t} \lambda_l(\tau|h;\theta)d\tau$. If $l' = l$, the indicator function $\mathbf{1}_l(l')$ is one, and it is zero otherwise. $\lambda_l(t|h;\theta)$ is the expected rate of event l at time t given h and model parameters θ. As it is conditional on the entire history, the process is non-Markovian.

A PCIM is a class of CIM where the conditional intensity function is a piecewise-constant function of time for any history. For each label l, a local structure S_l specifies regions in the timeline, where the conditional intensity function is constant and local parameters for each label θ_l represent the values of the intensity function in those regions. Formally, PCIMs are composed of local structures $S_l = (\Sigma_l, \sigma_l(t,x))$ and local parameters $\theta_l = \{\lambda_{ls}\}_{s\in\Sigma_l}$, where Σ_l denotes the set of states where the conditional intensity function is constant, λ_{ls} are non-negative constants representing the intensities in those states, and σ_l is a piecewise constant state function in time that maps a time and a history to Σ_l. In a PCIM, the state function for each label, σ_l, is represented using a decision tree where the states $s \in \Sigma_l$ are the leaves and the internal nodes are binary test functions, formally defined as basis state functions (Gunawardana et al., 2011). They map a time t and a history h to a subtree. If the test functions are picked to be piecewise-constant functions of time for any event history, the intensity function $\lambda_l(t|h) = \lambda_{ls}$, where $s = \sigma_l(t,h)$ becomes piecewise-constant as well. The resulting likelihood of the event sequence x can then be written as

$$p(x|S,\theta) = \prod_{l\in L}\prod_{s\in\Sigma_l} \lambda_{ls}^{c_{ls}(x)} e^{-\lambda_{ls} d_{ls}(x)} \ , \tag{2}$$

where $S = \{S_l\}_{l\in L}$, $\theta = \{\theta_l\}_{l\in L}$. c and d are the sufficient statistics for likelihood calculation. $c_{ls}(x)$ is the total number of events of label l occurring in x that map to state s, and $d_{ls}(x)$ is the total duration when the trajectory for l is mapped to s. An example of a PCIM model is given in Figure 1.

The basis state functions or the test functions need to be carefully chosen to control the capacity of the resulting model. One of the approaches is to index the functions based on predefined time windows and thresholds. Some examples are

- Is the time of day between 6am and 9am?
- Is the number of events with the label B in the past half an hour greater than a threshold?
- Was the most recent event of label A?

The piecewise-constant assumption allow for efficient inference and learning of the model. Gunawardana et al. (2011) showed that the marginal likelihood of an event sequence, x, can be computed in closed form given the structure S using a product of gamma distributions as a conjugate prior for θ. For learning the decision trees greedily, imposing a structural prior allows a closed form Bayesian score to be computed. Given a structure, the rates associated with the states (leaves) can be selected using maximum a posteriori or maximum likelihood estimation.

2.2 Extending PCIM

In the previous sections, we have represented an event sequence as a sequence of pairs, where each pair consists of the time stamp and label representing the time and type of a particular event. For EHR data, we take the label to be the type of measurement (pulse, for example). However, these events also have associated values (the heart rate measurement,

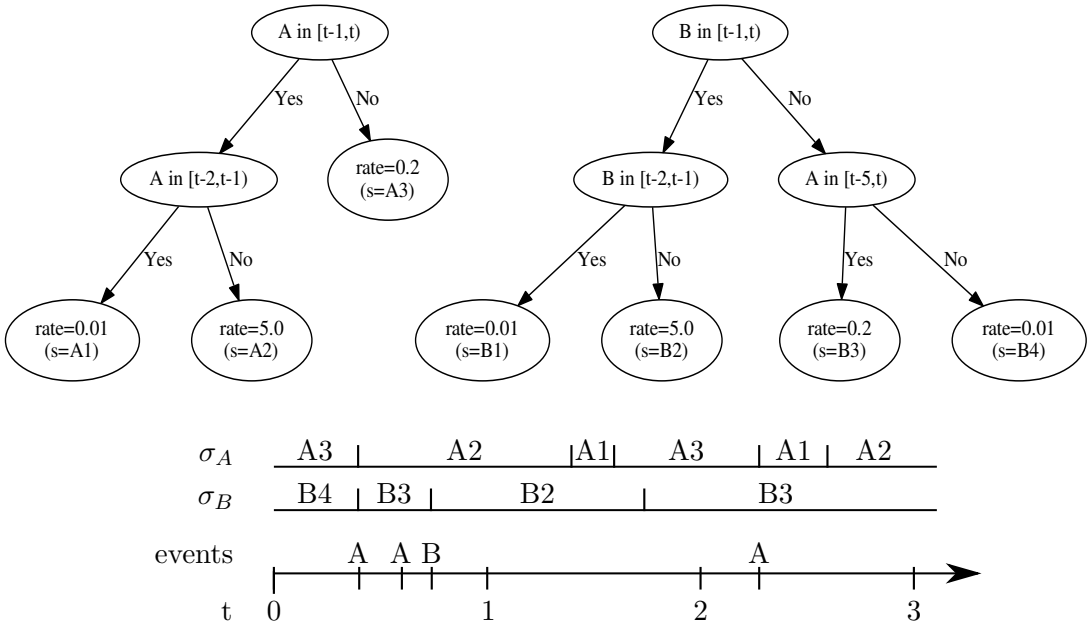


Figure 1: Top: Example decision trees representing a PCIM for two labels A and B. From Gunawardana et al. (2011). Bottom: Example trajectory and the corresponding piecewise-constant assignment of time to states (leaves of the PCIMs).

for example). In this paper, we have focused only on events with numerical values. However, we expect our modeling lessons to be transferable to other forms of event data including clinical notes which can be represented in numerical format, such as topic proportions.

Incorporating the values, an event is now a triple: a time stamp t, a label $l \in L$ and the value of the event $v \in \mathcal{R}$. The notation for an event sequence x then becomes $\{(t_i, l_i, v_i)\}_{i=1}^n$, where $0 < t_1 < \ldots < t_n$. We have extended the original PCIM model to model the values in each state $s \in \Sigma_l$. We impose a Gaussian distribution $\frac{1}{\sigma_{ls}\sqrt{2\pi}} e^{-\frac{1}{2\sigma_{ls}^2}(v-\mu_{ls})^2}$ on the value v associated with an event from a particular state s, represented by a leaf in the decision tree. The parameter set θ_l is now $\{\lambda_{ls}, \mu_{ls}, \sigma_{ls}\}_{s \in \Sigma_l}$, and, after incorporating the product of Gaussians with the product of the exponential distributions in Equation 3, the likelihood is

$$p(x|S_l, \theta_l) = \prod_{s \in \Sigma_l} \Lambda_{ls} \lambda_{ls}^{c_{ls}(x)} e^{-\lambda_{ls} d_{ls}(x)} \ , \tag{3}$$

where

$$\Lambda_{ls} = \left(\frac{1}{\sigma_{ls}\sqrt{2\pi}}\right)^{c_{ls}(x)} \ e^{-\frac{1}{2\sigma_{ls}^2}(u_{ls}(x) - 2\mu_{ls} m_{ls}(x) + \mu_{ls}^2 c_{ls}(x))} \ . \tag{4}$$

Here, $u_{ls}(x) = \sum_{v \leftarrow s} v^2$, $m_{ls}(x) = \sum_{v \leftarrow s} v$ are the sufficient statistics where $v \leftarrow s$ indicates that v is the value of an event of the portion of x that has been mapped to state s. We add a normal-gamma distribution as the (independent) prior over each μ_{ls}-σ_{ls} pair of parameters

to allow for closed-form Bayesian scores for structure learning via the greedy tree-growing algorithm.

3. Cohorts

Mortality modeling is often performed among heterogeneous population of patients to compare quality of care between different ICUs. To demonstrate the generality of our proposed modeling approach, we focus on two different cohorts, a cohort of pediatric patients in a tertiary PICU and a cohort of adult patients.

3.1 Cohort 1

We used data from the EHR archive of PICU at Children's Hospital Los Angeles, which consists of 11684 patient episodes, collected over a period of 10 years. It includes demographics, outcomes, PIM2 and PRISM3 scores, and times and value for measurements of vital signs, interventions, drugs, and lab tests.

3.2 Cohort 2

We also conducted our experiments on the publicly available MIMIC-III database which integrates de-identified clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. We removed patients younger than 18 years old to focus on adult patients in an ICU. The dataset includes vital signs, laboratory results, medications, charted observations, clinical notes.

4. Experiments

To show the effectiveness of our proposed temporal dependency modeling approach, we conducted experiments to predict in-hospital mortality based on the first 24 hours of clinical data for both the cohorts. The pre-processing step involves aggregating multiple related variables and normalization. Next, using a hold-out validation set, we select the final set of variables to be included in the PCIM models, based on the univarite performance of each variable. Critical components of the models such as the basis state functions were also chosen using the validation set. Next, we learned two separate sets of PCIM models on the training set for the two classes of patients: patients who died in hospital and those who survived. Finally, we obtain a severity of illness score from the two sets of models using the log-odds ratio and use this score as a mortality predictor. Details for each of the steps in our experimental process are explained in the following subsections.

4.1 Data Pre-Processing and Experimental Setup

For both datasets, some of the most commonly used clinical variables used in ICU scoring systems (PRISM3, SAPS-I, SAPS-II) were used to compose the primary sets of variables to perform our experiments. For the MIMIC-III dataset, we aggregated multiple related variables into one using a publicly available codebase[1]. The same codebase was also used

1. https://github.com/MIT-LCP/mimic-code

number of measurements of variable X in past t hours is $\geq \tau$
number of measurements of variable X in past t hours with value $\geq \tau'$ is $\geq \tau$
current time t is within a particular interval within the day

Table 1: Binary Tests for learning PCIM. X denotes any measurement variable and τ and τ' are each one of a set of thresholds.

to extract patient mortality outcomes, SAPS-I, SAPS-II score, and measurements for the first 24 hours. The CHLA data were similarly mapped.

For the cohort of pediatric patients, we normalized the values of age-dependent variables, such as heart rate, dividing by the median value among healthy children of the same age range and sex, according to published tables (Fleming et al., 2011). Then, we performed z-score normalization for all variables and removed data falling outside of ± 4 standard deviations from the mean of each variable. For adult patients there is relatively low variability on variables across different age groups and therefore we performed only z-score normalization.

The two class labels are "death" and "survival." We take the former to be the positive class. For both datasets, we randomly perform a (65-15-20) training-validation-test split. The validation set was used for hyper-parameter tuning and selection of the set of binary tests (both type and parameters). In order to make comparisons with PIM2 and PRISM3, we ran our final experiments for Cohort 1 on patient episodes with both PIM2 and PRISM3 scores reported in the dataset (PRISM3 scores of 0 were dropped). This resulted in a final dataset containing 4601 patient episodes (mortality rate 6.2%). Constructing only the test set with patient episodes having both scores reported would have introduced bias. For MIMIC-III, the final dataset included 34971 patient episodes corresponding to the first ICU and hospital stay of the patient (in hospital mortality rate 11.67%).

4.2 Choice of Binary Tests

The base binary tests available for internal nodes for the PCIM are listed in Table 1. The tests for checking whether the total number of measurements within some time interval is greater than a particular threshold (τ) is to allow the model to capture the burstiness in measurements of certain vital signs and lab tests. These tests allow modeling of self-excitation (burstiness) and self-suppression. We also introduce a test to check whether the number of measurements with value $\geq \tau'$ is greater than a particular threshold (τ) within a particular time interval. This allows the rate of subsequent events to depend on the values of previous events (e.g. measurements with extreme values of a particular variable causing more events of the same or a different variable in near future).

The full bank of binary tests are chosen by selecting the parameters from pre-determined sets of values. In particular, "past t hours" uses $t \in \{1/60, 1/12, 1/4, 1/2, 1, 24\}$; "value $\geq \tau'$" uses $\tau' \in \{-3.0, -2.5, -2, -1.5, 1.5, 2, 2.5, 3.0\} \times$std. dev.$+$mean; "number of measurements $\geq \tau$" uses $\tau \in \{0, 1, 5, 10, 20, 30, 50\}$; and "current time is within a particular interval of the day" uses range of one of "within δt minutes of the top of the hour," where $\delta t \in \{1, 2, 5, 15\}$ or within the first t hours of the day where $t \in \{1, 2, 3, \ldots, 24\}$.

4.3 Severity of Illness Score computation

Denote the set of models learned for the "died" class as M_{death} and the set for the "survive" class as M_{survive}. Both sets include PCIM models learned on the measurement timings and values of the selected variables. We score patient with event sequence x with the log-odds: $\log \frac{p(M_{\text{death}}|x)}{p(M_{\text{survive}}|x)}$ and directly use it as a mortality predictor.

4.4 Full Severity Score

We use a support vector machine (SVM) as the final classifier using the computed severity of illness score (as above) and other features (Pedregosa et al., 2011). For MIMIC-III, the static features include age, minimum GCS, minimum PaO2/FiO2 ratio for the ventilated patients, comorbidity, and urine output in the first 24 hours. For Cohort 1, we used PIM2 score as a feature. For experiments with both vital and lab measurements, we use the log-odds as a feature separately for each set of measurements. While each PCIM tree models the rate for one particular variable, the binary tests in the tree can look at all available variables (both vitals and labs) to capture the temporal dependencies with other variables.

To solve the problem of large class imbalance, we randomly re-sample instances from the minority class ('death') to preserve a 70-30 ratio between the two classes. However, test set distribution was not modified. A 5-fold cross validation with grid search on the training set is used to select the kernel function (linear/RBF) and tune C and the kernel parameters.

4.5 Baselines

For evaluating the predictive performance of our approach, we compare with PIM2 and PRISM3 scores for cohort 1, and SAPS-I and SAPS-II scores for cohort 2.

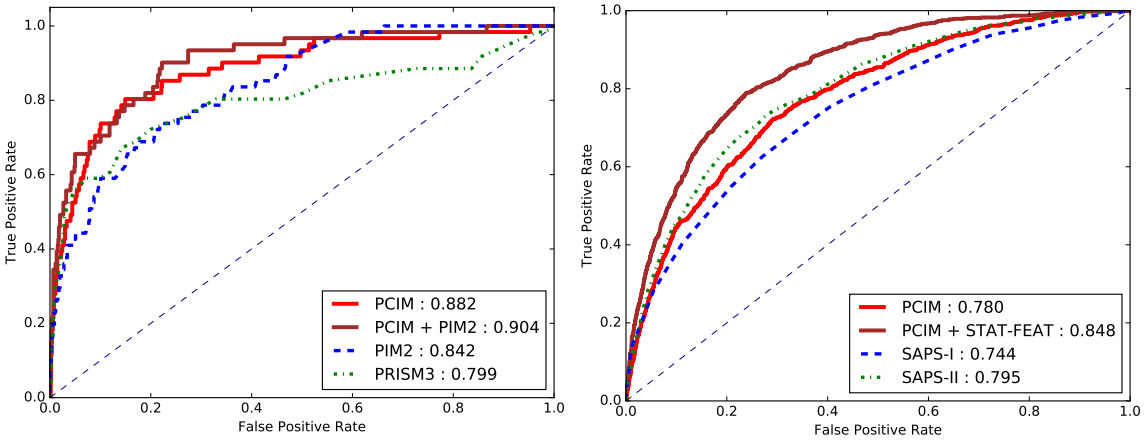
5. Results and Discussion

In this section, we present a comparative evaluation of our method with the standard ICU scoring systems. Statistical significance of the difference between the ROC curves presented was measured using MedCalc statistical software[2] (DeLong et al., 1988).

Figure 2(a) shows the comparison of ROC curves for PIM2, PRISM3, and SOI scores computed from PCIM models. SOI scores computed from the learned PCIM models outperform standard scoring system baselines, PIM2 and PRISM3 ($p(\text{PCIM} \sim \text{PRISM3}) < 0.024$ and $p(\text{PCIM} \sim \text{PIM2}) = 0.146$). The reason for relatively higher p-values is highly likely to be caused by the small test set size of Cohort 1. PCIM models do not directly incorporate information, such as admission baseline features which capture how ill a child was at the time of admission. PIM2 is largely based on such features calculated within two hours of admission and outperforms PRISM3 in our experiments on the holdout testset. Best performance is achieved when PIM2 is combined with the SOI score obtained from the PCIM models by adding PIM2 as an extra feature to the SVM.

ROC curves for SAPS-I, SAPS-II, and SOI scores computed from PCIM models on the MIMIC-III dataset are presented in Figure 2(b). Similar to Cohort 1, pre-admission and in-ICU information, such as ventilation and comorbidity were not directly incorporated in the

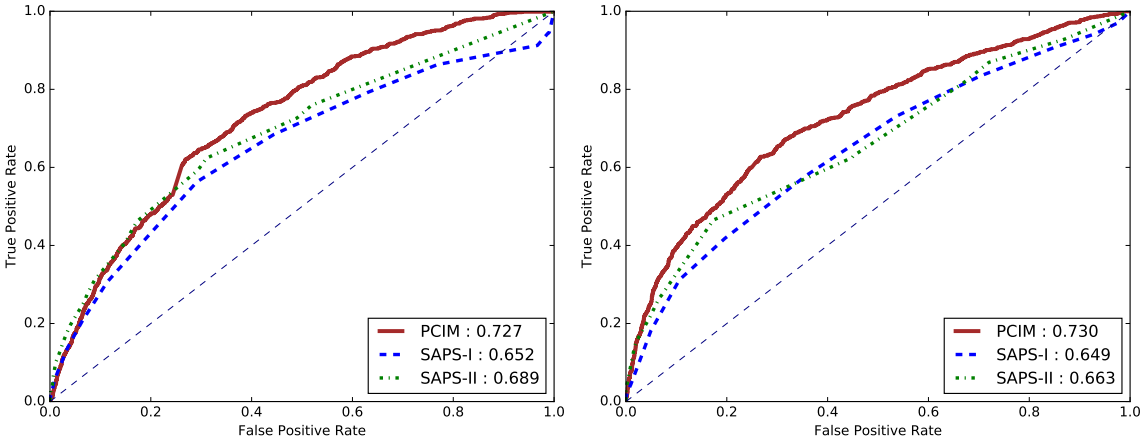
2. https://www.medcalc.org



(a) Cohort 1 (CHLA)

(b) Cohort 2 (MIMIC-III)

Figure 2: Comparison of all methods for in hospital mortality prediction



(a) Scores from lab variables only (MIMIC-III)

(b) Scores from vital signs only (MIMIC-III)

Figure 3: Comparison of all methods for the scoring of labs and vitals variables in MIMIC-III

PCIM models. When added to our method (again, as features to the SVM), we outperform both SAPS-I and SAPS-II with statistical significance (AUROC difference has $p < 0.0001$).

For MIMIC-III, Figure 3 shows the ROC curves using only labs or only vital signs. Scores computed from the four vital variables significantly outperform (AUROC, $p < 0.0001$) the isolated vitals scores of both SAPS-I and SAPS-II, computed using certain ranges of values, Figure 3(b). The area under the ROC curve difference computed only using the lab variables is similarly statistically significant ($p(\text{PCIM} \sim \text{SAPS-I}) < 0.0001$ and $p(\text{PCIM} \sim \text{SAPS-II}) = 0.0001$), Figure 3(a).

6. Differences with existing Recurrent Neural Network based Methods

While existing RNN based approaches (Lipton et al., 2015; Aczon et al., 2017; Harutyunyan et al., 2017; Lipton et al., 2016; Che et al., 2016) can strongly capture the temporal trend of the values of the variables, our proposed methodology is also taking into account the intensity of different events by directly modeling the generation of a new event of a particular type with respect to previous events' values, timings, and types. One particular advantage to directly model the event generation process in continuous time domain is to avoid discrete window based aggregation and any form of data imputation. Discrete time approaches usually assume the data are "missing at random," which does not always hold for irregularly sampled data. Also, this leads to reduced variability of more frequently measured signals. Imputing normal values for missing data results in loss of distinction between a truly normal and missing measurement. Also, forward and back filling imputation strategies discard information about the timings of the measurements. Another advantage of our method is the model interpretability, added flexibility of model selection through the choice of basis state functions, and applicability in small to intermediate datasets, where neural network based methods might be an overkill and prone to overfitting. We would like to explore connections between Recurrent Neural Networks and conditional intensity functions of marked point process (Xiao et al., 2017; Du et al., 2016) in our future works to extract the advantages of both methods simultaneously.

7. Related Works

Standard ICU scoring systems developed for pediatric (e.g. PIM2, PRISM3) and adult (e.g. SAPS-I, SAPS-II) ICU units rely on scoring a patient based on a range of values of particular physiological variables recorded within the first 24 hours, in addition to information such as pre-ICU procedures, admission category and in-ICU ventilation data. However, they largely ignore the temporal nature of the data and are often susceptible to missing values. These scores are good for early detection of severe illness and are excellent benchmarking tools for determining ICU performance.

Recent work has focused on extracting more meaningful features from sources like clinical notes, which is ignored in traditional ICU scoring systems. Both Lehman et al. (2012) and Ghassemi et al. (2014) used non-parametric topic modeling techniques, hierarchical Dirichlet process and latent Dirichlet allocation (LDA) respectively, for learning the latent topic structure of nursing notes, which improves mortality prediction performance over SAPS-I and SAPS-II. Ghassemi et al. (2015) used multi-task Gaussian process (MTGP) models to

model irregularly sampled clinical data and new clinical notes for acuity forecasting. Their approach shows that hyperparameters learned from MTGP models capturing the temporal change in topic membership of notes recorded for a particular patient can improve mortality prediction performance, if combined with SAPS-I and latent topic features. Jo et al. (2015) proposed a model combining LDA and Hidden Markov Models (HMMs) to capture the temporal dynamics of underlying patient states from the clinical notes and improve long term mortality prediction.

Lipton et al. (2015) used Long Short Term Memory Recurrent Neural Networks (LSTM-RNNs) in disease phenotyping. However, they employed a discrete (hourly) time window based approach and imputed values in missing windows using either forward or back-filling strategy. In a subsequent work (Lipton et al., 2016), better performance was achieved using binary indicator variables for the time windows with missing values, however it still relied upon a window based scheme. Aczon et al. (2017) developed a dynamic mortality risk prediction model using LSTM-RNNs in pediatric ICU data. They didn't resample the data at any fixed sampling rate, however depended upon zero or forward imputation for the values of variables, which had no recorded value at a particular time-point where at least one variable was recorded for a particular patient. Che et al. (2016) proposed a model based on Gated Recurrent Units (GRUs) with trainable decays to capture the temporal structure of the missing values. The AUROC achieved on the MIMIC-III dataset (19714 admission records) with 24 hours of data is 0.78821. The feature set however, consisted of 91 extra variables in addition to the 8 lab variables used separately in one of our experiments, achieving an AUROC of 0.727. None of these works however, focused on capturing the intensities of events of varied types and their dependence on previous history across the timeline of a patient stay.

Marlin et al. (2012) focused on unsupervised learning from time series data collected from a pediatric ICU. While their approach shows non-uniformities among patients across different clusters extracted, their approach also employs a discrete time window and ignores measurement timings. Joshi and Szolovits (2012) modeled patient acuity using an unsupervised learning approach, radial domain folding, which in an organ specific manner, learns lower dimensional abstractions from routinely generated physiological data. Logistic regression models trained on the learned RDF layers outperform SAPS-II scoring system in mortality prediction. Weiss and Page (2013) used a forest-based point process model (an extension of the PCIM) to predict future onset of myocardial infractions. Based on the rates learned from event history, rates of future events are forecast. Their approach shows the strength of continuous time models in medical data, however doesn't address the temporal modeling of routinely measured physiological signals. Saria et al. (2010) focused on a non-parametric Bayesian method for data analysis in continuous time series including health care data, based on topic models.

8. Conclusion

Our contribution has been to demonstrate the efficacy of directly modeling the continuous-time temporal dependencies of discrete events recorded in an EHR. Our results indicate that a severity of illness score computed using our proposed modeling technique improves the performance of hospital mortality prediction.

One of the limitations of our approach is the use of a generative model for the task of classification, using the log-odds calculated from the representative models for the two class labels, learned separately. However, careful choice of the basis state functions from a validation set resulted in significant discrimination between learned models of the two classes and improved prediction performance. Discriminative training might better accentuate the temporal differences between the two classes. This is left as a future work.

Our method provided one prominent feature for an SVM classifier. We added other static measurements (age, minimum Glasgow Coma Scale, minimum ratio of partial pressure arterial oxygen and fraction of inspired oxygen, total urine output, Pa02, and co-morbidity) as features to the SVM. These could have also been added as possible values for binary tests in the PCIM model. This, particularly in conjunction with disciminative PCIM training, might work better as a classifier. We also focused only on nine variables for MIMIC-III because those are used in SAPS-II. The model might also be improved with an more expansive variable set. For the Cohort 1 data, we used more variables and saw a larger improvement. In one case, this was not statistically significant; we expect this is because of the relatively smaller sample size.

Acknowledgement

This work was supported by DARPA (FA8750-12-2-0010).

References

Melissa Aczon, David Ledbetter, L Ho, Alec Gunny, Alysia Flynn, Jon Williams, and Randall Wetzel. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. *arXiv preprint arXiv:1701.06675*, 2017.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

Thomas L Dean and Keiji Kanazawa. Probabilistic temporal reasoning. In *AAAI*, pages 524–529, 1988.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564. ACM, 2016.

Susannah Fleming, Matthew Thompson, Richard Stevens, Carl Heneghan, Annette Plüddemann, Ian Maconochie, Lionel Tarassenko, and David Mant. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *The Lancet*, 377(9770):1011–1018, 2011.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB '00, pages 127–135, New York, NY, USA, 2000. ACM. ISBN 1-58113-186-0. doi: 10.1145/332306.332355. URL `http://doi.acm.org/10.1145/332306.332355`.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI*, pages 446–453, 2015.

Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems*, pages 1962–1970, 2011.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

Yohan Jo, Natasha Loghmanpour, and Carolyn Penstein Rosé. Time series analysis of nursing notes for mortality prediction via a state transition topic model. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1171–1180, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6. doi: 10.1145/2806416.2806541. URL `http://doi.acm.org/10.1145/2806416.2806541`.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1276. American Medical Informatics Association, 2012.

Jean-Roger Le Gall, Philippe Loirat, Annick Alperovitch, Paul Glaser, Claude Granthil, Daniel Mathieu, Philippe Mercier, Remi Thomas, and Daniel Villers. A simplified acute physiology score for ICU patients. *Critical care medicine*, 12(11):975–977, 1984.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

Li-Wei H Lehman, Mohammed Saeed, William J Long, Joon Lee, and Roger G Mark. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In *AMIA*. Citeseer, 2012.

Ted Petrie Leonard E. Baum. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966. ISSN 00034851. URL http://www.jstor.org/stable/2238772.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 2016.

Benjamin M. Marlin, David C. Kale, Robinder G. Khemani, and Randall C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 389–398, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0781-9. doi: 10.1145/2110363.2110408. URL http://doi.acm.org/10.1145/2110363.2110408.

Uri Nodelman, Christian R. Shelton, and Daphne Koller. Continuous time Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 378–387, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-897-4. URL http://dl.acm.org/citation.cfm?id=2073876.2073921.

Adam Oliner and Jon Stearley. What supercomputers say-an analysis of five system logs. In *IEEE/IFIP Conf. Dep. Sys. Net*, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Murray M Pollack, Kantilal M Patel, and Urs E Ruttimann. PRISM III: an updated pediatric risk of mortality score. *Critical care medicine*, 24(5):743–752, 1996.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989. ISSN 0018-9219. doi: 10.1109/5.18626.

Shyamsundar Rajaram, Thore Graepel, and Ralf Herbrich. Poisson-networks: A model for structured point processes. In *Proceedings of the 10th international workshop on artificial intelligence and statistics*, pages 277–284. Citeseer, 2005.

Suchi Saria, Daphne Koller, and Anna Penn. Learning individual and population level traits from clinical temporal data. In *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop*. Citeseer, 2010.

Aleksandr Simma and Michael I. Jordan. Modeling events with cascades of poisson processes. *CoRR*, abs/1203.3516, 2012. URL http://arxiv.org/abs/1203.3516.

Aleksandr Simma, Moisés Goldszmidt, John MacCormick, Paul Barham, Richard Black, Rebecca Isaacs, and Richard Mortier. CT-NOR: representing and reasoning about events in continuous time. *CoRR*, abs/1206.3280, 2012. URL `http://arxiv.org/abs/1206.3280`.

Anthony Slater, Frank Shann, and Gale Pearson. PIM2: a revised version of the paediatric index of mortality. *Intensive Care Medicine*, 29(2):278–285, 2003. ISSN 1432-1238. doi: 10.1007/s00134-002-1601-2. URL `http://dx.doi.org/10.1007/s00134-002-1601-2`.

Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005.

Jeremy C Weiss and David Page. Forest-based point process for event prediction from electronic health records. In *Machine learning and knowledge discovery in databases*, pages 547–562. Springer, 2013.

Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M Chu. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*, pages 1597–1603, 2017.

Appendix: Final Variables

Best motor response (GCS)	Best verbal response (GCS)
Glasgow coma scale total	Eye opening response(GCS)
Capillary refill rate (sec)	Diastolic blood pressure non-invasive (mmHg)
Systolic blood pressure invasive (mmHg)	Systolic blood pressure non-invasive (mmHg)
Pulse oximetry percentage	Heart rate (bpm)
Respiratory rate (bpm)	EtCO2 (mmHg)
Left pupillary response	Right pupillary response
Temparature (C)	

Table 2: Vital Signs (Cohort 1: CHLA)

ABG Base excess (mEq/L)	ABG PCO2 (mmHg)
ABG pH	Albumin level (g/dL)
BUN (mg/dL)	Bicarbonate serum (mEq/L)
Bilirubin total (mg/dL)	CBG PCO2 (mmHg)
CBG pH	Calcium ionized (mg/dL)
Calcium total (mg/dL)	Chloride (mEq/L)
Glucose (mg/dL)	P/F ratio
PT	PTT
Platelet count (K/uL)	Potassium serum (mEq/L)
Sodium serum (mEq/L)	VBG Base excess (mEq/L)
VBG PCO2 (mmHg)	VBG pH
White blood cell count (K/uL)	

Table 3: Labs (Cohort 1: CHLA)

Bicarbonate	BUN
Sodium	Potassium
WBC	Heart Rate
Respiratory Rate	Systolic Blood Pressure
Temperature(C)	

Table 4: Labs and Vital Signs (Cohort 2: MIMIC-III)