Prediction using Patient Comparison vs. Modeling: A Case Study for Mortality Prediction

Mark Hoogendoorn[1,2], Ali el Hassouni[1], Kwongyen Mok[1], Marzyeh Ghassemi[2], and Peter Szolovits[2]
[1]VU University Amsterdam, Department of Computer Science,
m.hoogendoorn@vu.nl, {a.el.hassouni, k.y.mok}@student.vu.nl
[2]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Lab,
{mghassem, psz}@mit.edu

***Abstract*— Information in Electronic Medical Records (EMRs) can be used to generate accurate predictions for the occurrence of a variety of health states, which can contribute to more pro-active interventions. The very nature of EMRs does make the application of off-the-shelf machine learning techniques difficult. In this paper, we study two approaches to making predictions that have hardly been compared in the past: (1) extracting high-level (temporal) features from EMRs and building a predictive model, and (2) defining a patient similarity metric and predicting based on the outcome observed for similar patients. We analyze and compare both approaches on the MIMIC-II ICU dataset to predict patient mortality and find that the patient similarity approach does not scale well and results in a less accurate model (AUC of 0.68) compared to the modeling approach (0.84). We also show that mortality can be predicted within a median of 72 hours.**

I. Introduction

Making predictions on future health states can be of great value in the medical domain because such predictions can contribute to disease prevention, early detection, more effective treatment, etc. One way of generating predictions is by applying machine learning techniques to data stored in Electronic Medical Records (EMRs). The EMRs usually cover a variety of aspects of a patient's health state and the type of data typically varies from highly structured (e.g., billing codes) to very unstructured (e.g., notes by the physician). In addition, EMR data is of a highly temporal nature and often contains ample missing values.

In order to cope with the characteristics of EMR data and fully exploit the wealth of information contained in them, a variety of approaches have been developed. Some approaches have focused on extraction of features from EMRs to make them better suited for the generation of *predictive models* (see e.g. [4]). These features typically abstract over the time dimension, combine multiple measurements, and handle sparseness of the data. The features can then be exploited in commonly used classification approaches such as logistic regression. Such methods have also been used to estimate how long in advance certain predictions can be made (cf. [3]). Another category of approaches has focused on defining *patient similarity*, driving more instance-based learning (e.g., k-nearest neighbor) approaches. Here, comparison of patient measurements that cover differently sized time windows or shifted data are examples of challenges for which solutions have been developed.

While both approaches described above are appealing, very little work has been done that compares the two. In this paper, we aim to make such a comparison using the MIMIC-II dataset [8] to predict mortality among ICU patients by using their EMR data between ICU intake and discharge. We re-implement the approach developed in [4] for the *predictive modeling* category and make it available as a benchmark. This is similar to other work where windowing, aggregation or modeling of structured numerical data creates a single feature matrix that can be fed into a structured deterministic classifier [1], [2], [5], [7]. We additionally study the approach in combination with a Cox model and investigate how long in advance predictions of mortality can be made. For the *patient similarity* case we use dynamic time warping for numerical data and we combine this with a tailored variant of a k-nearest neighbor approach that we have developed ourselves. We characterize the algorithms' accuracy and speed.

This paper is organized as follows. First, we will provide a description of the dataset in Section II. The methods used to make predictions are presented in Section III. Next, Section IV presents the experimental setup and the accompanying results. Finally, Section V concludes the paper.

II. Dataset Description & Preparation

The MIMIC-II V2.6 database [8] contains 26,647 patients of which 8,7% died at the ICU. For each patient, it holds a huge variety of measurements, obtained from ICU systems and hospital archives. The ICU data covers bedside monitoring (including vital signs, waveforms, trends, and alarms) as well as chart data (covering fluids, medications and progress notes). The hospital archives cover an even greater variety of information, but we only use the demographics data and thus limit the scope of our features to the acute ICU data combined with some background information on the patient.

For the selection and preliminary pre-processing of the data, we draw inspiration from [4]. Essentially, the following types of measurements are distinguished:

1) **continuous and ordinal measurements:** a number of continuous and ordinal measurements are present in the dataset, see [4], pp. 32-33 for a full overview. Categories of measurements are cardiovascular, chemistries, hematology, arterial blood gases, and

ventilation measurements, covering 64 measurements in total. For each of these measurements, an acceptable range is specified. Values outside of this range are not used. Furthermore, a so-called hold time is expressed per measurement, indicating how long the measurement is assumed to remain valid after it has been observed.

2) **categorical measurements and observations:** a list of categorical variables, typically covering a status or diagnosis assigned by the medical staff (e.g. heart rythm, risk for falls, Riker Sedation-Agitation Scale (SAS)). Each of these is mapped to binary or ordinal values, as in [4, pp. 34-35]. We use 15 such categorical values, mapped to 28 binary variables and 7 ordinal ones. Again, a hold time is indicated.
3) **medications:** a list of relevant intravenous medications (51 in total) with the accompanying dosage [4, p. 36], normalized to the weight of the patient.
4) **input/output measurements:** the input variables in this category are related to blood only (red blood cells, other blood inputs) as the other fluid inputs (e.g. glucose) are considered under the chemistries of the continuous and ordinal measurements listed before. The output is related to urine production [4, pp. 37-38].
5) **demographics:** only sex, age, and the ethnicity are considered.

In order to preprocess the dataset, we obtain all the time stamped data related to the variables described above from the database. We extrapolate data for measurements where no value is recorded for a time point that is within the hold range of a previous observation. The granularity of the dataset is one set of measurements for every 15 minutes of ICU time.

III. Methodology

In this Section, we explain the configuration of the two learning algorithms as well as the feature extraction in more detail. First we explain the generic steps for both algorithms, followed by the experimental setting.

A. Generic Steps

Figure 1 shows an overview of the pipeline for both setups of the algorithms. Essentially, two generic steps are performed. First, all missing time points between ICU intake and discharge are inserted with unknown values. Thereafter, missing data is imputed following the aforementioned hold principle: for each measurement a so-called hold time is indicated expressing how long a previous measurement is considered valid. In case a missing value is encountered that falls within the hold range of the last preceding known measurement, that value is inserted.

B. Predictive Modeling

There are three main steps within the predictive modeling component (Figure 1).

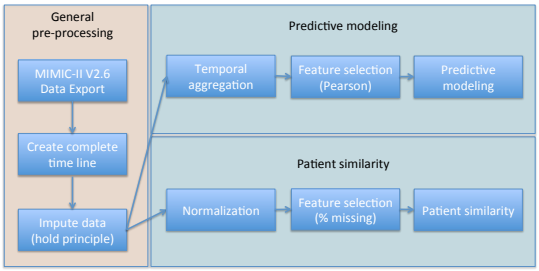

Fig. 1. Overview learning approaches

1) Temporal Aggregation: In order to generate a high-quality predictive model, some form of feature engineering and selection is needed. Essentially, features can abstract along two dimensions: the temporal dimension and the measurement dimension. In this paper, we pursue temporal abstraction (cf. [4]), varying by the type of data. For the *continuous and ordinal measurements* we derive the minimum, maximum and mean values over the selected time period, as well as the standard deviation. In addition, the slope of a linear regression model fit to the observed values is derived over fixed time windows (4 or 28 hours, or both, depending on the type of measurement, see [4]). In the case of the *categorical measurements* the average value over the total time period is determined. Note that all these attributes have been transformed to binary or ordinal already as indicated in the previous Section. In the *medications* we take the average dose of medications provided during the time period and the same holds for *input/output measurements* where we take the average input and output values. In addition, we consider a number of specific *derived variables* that have been found relevant in the literature (see [4, pp. 41-42] for an overview).

For aggregation of the data we can use different settings, for instance by taking only the first day of data, the first 40 hours, etc. This allows us to study the impact of more data becoming available upon the predictive performance. To align the time lines of patients we start our aggregation at a fixed time (11 pm) to make sure day and night rhythms of patients are consistent. The approach taken is inspired by the SDAS*n* approach (for Stationary Daily Acuity Score) in [4].

2) Feature selection and Predictive Modeling: After having engineered the features, we perform a selection of the most promising features. To determine what features we include in the model we use the Pearson correlation coefficient between each feature and the outcome we seek to predict. We perform this calculation on the whole dataset. We then create a model based on the most correlated attributes. In order to avoid having too much dependence in the variables, we use an iterative process where features with the highest correlation with the target (i.e. mortality) are selected if they are not highly correlated (≥ 0.2 for the logistic regression model and ≥ 0.7 for the Cox model) to a feature already part of the set. The number of features to include is determined by means of a 5-fold cross-validation approach. Note that this feature selection process differs from [4]. We decided to use this different approach as we did not have access to

domain experts to select features.

C. Patient Similarity with K-Nearest Neighbor

The steps for patient similarity computation are below.

1) Normalization: We apply a simple normalization approach by just scaling the values to a $[0, 1]$ range.

2) Feature selection: For feature selection, we use the percentage of missing values as a selection criterion and try different settings in a 5-fold cross-validation setting. Experiments showed that a similar feature selection approach as used in the predictive modeling approach performed worse.

3) Patient similarity: In order to define the patient similarity, we use dynamic time warping for highly time varying features. In this case we have selected the heart rate, respiration, the nocturnal, systolic, and diastolic blood pressure, and the oxygen saturation as such features. To ease the computational demands we use the Keogh lower bounds (cf. [6]) instead of computing the full mapping, and perform this for a fixed window size of an hour (i.e. four time points). For all other measurements we average the values observed for the patients and compute the Euclidean distance. Of course, not all measurements might have a value during the investigated time frame, therefore we only compare values of features that have at least one measurement for both patients and average their distances. We add a penalty for each feature that cannot be matched (i.e. where at least one of the patients does not have a single measurement related to the feature). The parameter C determines the height of the penalty.

Assuming we have a set DTW of highly time varying features and REG of other features where the vector $T_{p,i}$ represents the recorded values of a variable i for patient p during the period under consideration, i.e. $t_{p,i,1}$, ..., $t_{p,i,n}$ in case of n time points. Then the distance between a patient a and b is defined according to equation 1.

$$dist(a,b) = \frac{\sqrt{\sum_{i \in DTW} dtw(T_{a,i}, T_{b,i})^2 + \sum_{j \in REG} mdist(T_{a,j}, T_{b,j})^2 + penalty(a,b)}}{feat\_matched(a,b)} \quad (1)$$

where

$$penalty(a,b) = C \cdot (1 + |DTW| + |REG| - feat\_matched(a,b)) \quad (2)$$

$$dtw(T_1, T_2) = \begin{cases} keogh\_bound(T_1, T_2) & \text{if } (|T_1| > 0) \wedge (|T_2| > 0) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$mdist(T_1, T_2) = \begin{cases} (\langle T_1 \rangle - \langle T_2 \rangle) & \text{if } (|T_1| > 0) \wedge (|T_2| > 0) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$feat\_matched(a,b) = \sum_{i \in DTW \cup REG} \begin{cases} 1 & \text{if } (|T_{a,i}| > 0) \wedge (|T_{b,i}| > 0) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

After obtaining the distances, we select the k closest patients and assign the average class score (i.e. the sum of all positive cases among the k nearest neighbors divided by k) as the risk for that specific patient.

D. Experimental Setting

In order to evaluate our approach we test our approach on the aforementioned MIMIC-II V2.6 dataset. We apply certain criteria for selection of the patients in the dataset: at least one BUN (Blood Urea Nitrogen) observation, one GCS (Glasgow Coma Scale) observation, one hematocrit observation, one heart rate observation, one IV medication recorded, and the patients should receive adult care. In addition, we remove patients who left the ICU within 24 hours. This is done to guarantee that each patient has some relevant data, although it reduces the dataset from 26,647 to 13,923 patients while resulting in an almost identical class distribution.

IV. Results

In this section we describe the results obtained using the different approaches and compare them. In addition, we perform an in-depth earliest mortality prediction analysis.

A. Predictive Modeling versus Patient Similarity

We run experiments to compare the *predictive modeling* method with the *patient similarity* method. We randomly sample varying numbers of patients (ranging between 150 and 2500) to study the influence of the number of patients upon the accuracy of the predictions and the computation time. We only consider the data of the first day at the ICU. As a means of evaluation we use a stratified 5-fold cross-validation approach and use the average Area Under the Curve (AUC) over the 5 folds as a performance metric.

For the *predictive modeling* case we use the logistic regression approach with 50 features (this number has been determined based on experiments using a 5-fold cross-validation setting). We use L2-regularization with a cost of 150 and a tolerance of $1e^{-6}$. These results were obtained using a grid search by cross-validation over the training set. The resulting AUC's for varying numbers of patients are shown in Figure 2. Note that the performance seems to decrease slightly when moving to 2500 patients, this is most likely due to the fixed set of features we use that has been based on a smaller set of patients.

For the nearest neighbors approach, a value of $k = 1$ was selected and 132 features were selected as this showed superior performance in a 5-fold cross-validation setting (0.68 compared to 0.65 for $k = 2$ and 0.54 for $k = 10$) in an initial experiment where we experimented with 150 to 1000 patients. The fact that such a small k is best is surprising but might be caused by the unbalance in the dataset or the huge number of missing values. We set $C = 1$ and select the features with values present in more than 85% of the cases. Figure 2 shows the result obtained with different numbers of patients. From this Figure we can clearly see that the accuracy of the model decreases as we increase the number of patients. Reasons for this unexpected behavior could be the fact that many features will not have a match due to missing data, and that the distance function is not robust in approximating the true distance given the number of features we use and the missing data. The best performance is much lower than that of the predictive modeling approach (0.68

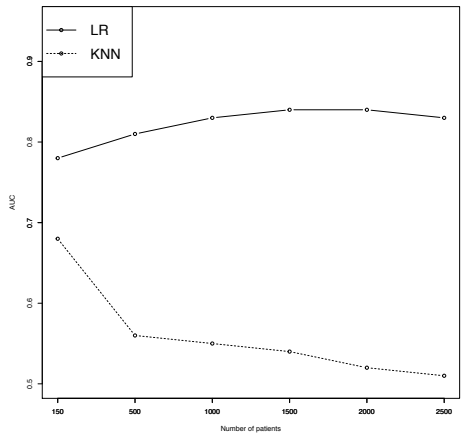

Fig. 2. Sensitivity of LR and KNN to the number of patients used for training.

versus 0.84). The difference is significant (using a paired t-test, p=0.00275). We do see that dynamic time warping helps: it increases the best AUC from 0.66 to 0.68 (150 patients). This difference is also significant (paired t-test, $p < 0.05$).

When comparing the runtimes of the algorithms the KNN approach obviously scales a lot worse compared to the logistic regression approach. For 2500 patients KNN takes 9,980 minutes to run while logistic regression is finished in just under one minute.

B. Earliest Prediction Time Analysis

Furthermore, we want to explore how early in the process we can detect impending patient mortality. We select the Cox model due to its ability to handle the time dimension, which is required for this exploration. We train the Cox model on the complete history of a thousand patients. Furthermore, we select 100 features with a regularization parameter of 0.01 and a tolerance of 1e-07 similar to the approach applied for logistics regression. We apply the model to a set-aside test set (of 967 patients) and select the point on the AUC curve where the curve starts to flatten and set the threshold of the model accordingly. We then apply the model to the test set and select the true positives. For this group of patients we explore how long before the actual moment of death the model starts to forecast that the patient will die. To accomplish this, we iteratively remove time frames of 4 hours starting from the moment of death and work backwards. This is in line with [3]. We obtain an AUC on the set aside test set of 0.78 from the Cox model, substantially lower than the performance of the logistic regression model. We select a point with a true positive rate of 0.87 and a false negative rate of 0.44 (where the curve flattens). When we explore the true positives (67 in total), the mean time to death when we actually predict death is 153 hours. This is promising, but heavily influenced by outliers. The median is 72 hours. Figure 3 shows results on a per patient basis when the model is able to predict mortality correctly. Note that we only consider patients that were predicted correct using their full history. We did not observe inconsistent predictions in this set (i.e. the model alternating between different predictions), only single changes in the prediction were observed.

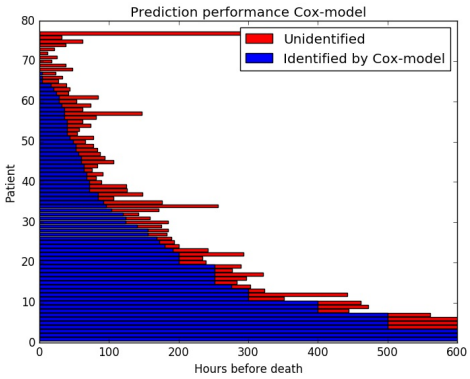

Fig. 3. Earliest prediction times

V. Discussion

While ample research has been reported on using EMRs for predictive modeling of certain health states, very few have focused on comparing approaches that are based on similarity with those that construct a model. This paper makes such a comparison for mortality prediction. The results, at least for our data, showed that performance of a modeling approach is superior to that of a patient similarity approach, both in terms of predictive performance as well as scalability. When studying the predictive models that result from the most accurate approach in more detail, we see that we can predict mortality relatively early on in the process for those we are able to identify correctly. For future work, we aim to study how generalizable these results are.

References

[1] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of KDD 2014*, pages 75–84. ACM, 2014.

[2] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proc. Twenty-Ninth AAAI Conf. on Artificial Intelligence*, 2015.

[3] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

[4] C. Hug. *Detecting hazardous intensive care patient episodes using real-time mortality models*. PhD thesis, 2009.

[5] R. Joshi and P. 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.

[6] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[7] L.-w. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA Annual Symposium Proceedings*, volume 2012, page 505. American Medical Informatics Association, 2012.

[8] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.