ICU Mortality Prediction using Time Series Motifs

Sean McMillan[1], Chih-Chun Chia[1], Alexander Van Esbroeck[1], Ilan Rubinfeld[2], Zeeshan Syed[1]

[1]University of Michigan, Ann Arbor, USA
[2]Henry Ford Hospital, Detroit, USA

Abstract

In this paper, we explore the application of motif discovery (i.e., the discovery of short characteristic patterns in a time series) to the clinical challenge of predicting intensive care unit (ICU) mortality. As part of the Physionet/CinC 2012 challenge, we present an approach that identifies and integrates information in motifs that are statistically over- or under-represented in ICU time series of patients experiencing in-hospital mortality. This is done through a three step process, where ICU time series are first discretized into sequences of symbols (by segmenting and partitioning them into periods of low, medium and high measurements); the resulting sequences of symbols are then searched for short subsequences that are associated with in-hospital mortality; and the information in many such clinically useful subsequences is integrated into models that can assess new patients. When evaluated on data from the Physionet/CinC 2012 challenge, our approach outperformed existing clinical scoring systems such as SAPSII, APACHEII and SOFA, with an event 1 score of 0.46 and an event 2 score of 56.45 on the final test set.

1. Introduction

The Physionet/CinC 2012 Challenge focused on the prediction of in-hospital mortality in intensive care unit (ICU) patients. In addition to baseline clinical variables (e.g., age, gender, height, and weight) as many as 37 in-hospital physiological and laboratory variables were available as time series for each patient over the first 48 hours of admission to the ICU. In our study, we focussed primarily on these time series data, and explored the hypothesis that there is much useful information in motifs (i.e., short characteristic time series patterns) within these data that can aid the prediction of in-hospital ICU mortality. This hypothesis is motivated by the observation that many short patterns in ICU time series have physiological interpretations and significance: for example, drops in blood pressure can correspond to episodes of acute hypotension while short periods of high heart rate may correspond to tachyarrhythmias. Methods that can identify and leverage these short patterns in ICU time series may therefore have value in the assessment of ICU patients.

2. Methods

Our overall approach consists of three steps:

- First, irregularly sampled time series data for each available ICU variable are converted into regularly sampled time series with (potentially missing) values at each two hour interval. We discretize these regularly sampled time series by partitioning the values into low, medium or high classes. In this manner, we transform the original irregularly sampled ICU time series into regularly sampled symbolic sequences.
- Second, the frequencies of short subsequences within these transformed signals are measured and subsequences that occur significantly more often (or less often) in patients experiencing ICU mortality than would be expected by chance alone are identified. This process makes use of statistical measures of discrimination and adjusts for multiple hypotheses.
- Finally, the frequencies of subsequences associated with in-hospital mortality are used are features to train models to evaluate new ICU patients.

Aspects of this work resemble earlier efforts by Chia et al. to predict adverse outcomes following acute coronary syndrome using information in heart rate time series[1]. Notable differences between the approach described here and the work of Chia et al. include extending motif discovery to multivariate time series as well as dealing with data that are irregularly sampled.

In more detail, because of the irregular data sampling inherent in many ICU time series, we organize these data into a temporal grid. The 48 hour recordings in the Physionet/CinC 2012 Challenge training set are divided into 24 non-overlapping 2 hour windows. The time series for each ICU variable is made regularly sampled in the following way. If the variable has a single sample in the 2 hour time window, that value is used in the new time series. If the variable has multiple values in the window, the mean is used. If the variable has no values in the window, the value in the new series is marked as missing. Each

Table 1. 10 features with the highest magnitude coefficients for models using length 1, 2, and 3 patterns respectively. Variables are listed in decreasing order of magnitude. The label (+) indicates a positive association with in-hospital mortality, while (-) indicates a negative association.

Rank	L = 1	L = 2	L = 3
1	K (-)	K (-)	K (-)
2	Age (+)	Albumin: Medium Low (+)	Temp: Low Low Low (+)
3	ALP (+)	Age (+)	HR (-)
4	Platelets: Medium (-)	Glucose: Missing High (+)	ALP (+)
5	Temp: Medium (-)	Urine: Low Missing (+)	Albumin: Low Missing Medium (+)
6	Urine: High (-)	ALP: High Missing (+)	PaO2: Missing Missing High (+)
7	Glucose: High (+)	GCS: Medium Missing (+)	Age (+)
8	BUN: High (+)	pH: Missing Low (+)	Temp: Medium High High (-)
9	pH: Missing (+)	Albumin (+)	RR: Medium Medium Medium (-)
10	ALT (+)	Urine: Medium High (-)	ALP: Medium Missing High (+)

variable is then discretized into three equiprobable bins, reflecting low, medium, and high values of the variable. These bins are defined using the full set of values in all patients from the training set, with divisions at the 33rd and 66th percentiles. Missing values are considered as a separate value. This resulted in 4 discrete values for each variable (i.e., low, medium, high, and missing).

As clinically useful patterns, we consider all exact subsequences of a given length in each variable's discrete sequence. We estimate the frequencies with which these patterns occur in the data both within patients who experienced in-hospital mortality and those who did not through direct counting. For length 1 (a single discretized value), when considering the full set of available variables this corresponds to 144 possible patterns. For length 2 (two consecutive discretized values), this corresponds to 576 possible patterns. This number increases exponentially as the length of patterns used increases.

Due to the large number of possible patterns, particularly with longer pattern lengths, we apply the rank sum test on the training data to each pattern's frequencies over patients as a feature selection criterion [2]. The rank sum test (also known as the Mann-Whitney U statistic) is a non-parametric test that assesses whether the medians of two distributions (in this case the frequencies of subsequences in patients who did and did not suffer in-hospital mortality) are significantly different. To account for the testing of many patterns, we use the Bonferroni correction to adjust the p values of the large number of hypotheses [3]. A pre-correction threshold of 0.05 is used to determine significance. Only patterns that are found to differ significantly in occurrence between patients with and without in-hospital mortality are used for training a model to predict in-hospital ICU mortality. In our work, we use the frequencies of selected patterns as features to train a support vector machine (SVM) classifier. Our choice of an SVM classifier is motivated by the ability of this approach to train models in high-dimensional spaces while generalizing well to out of sample data [4].

In addition to the pattern frequencies, we also use a set of baseline variables. Specifically, we use as many available variables comprising the Acute Physiology Score portion of the APACHEII severity score [5], the SAPSII severity score [6], and the SOFA score [7] as possible. If multiple measurements of these variables are available the mean, minimum, and maximum measurements over the course of the 48 hours are used.

Finally, to address the limitation that SVMs do not directly produce probabilistic outputs, we use Platt scaling to assign risk probabilities to patients based on the SVM predictions [8]. Platt scaling fits a sigmoid function to the SVM prediction, converting the SVM's continuous decision values into usable risk probabilities.

3. Experiments

We trained and evaluated our approach on the Physionet/CinC 2012 Challenge A, B and C data sets. In addition to measuring the standard evaluation metrics for the challenge, we also studied the SVM model developed through our approach to identify the most useful features. As the features were normalized before training the SVM model, the magnitude of the coefficients gives an indication of the strength of the relationship between the variable and in-hospital mortality. In addition, to evaluate whether the use of motif frequencies improved performance over only using baseline features, we compared versions of our SVM model with and without baseline features, and with varying lengths of patterns. Evaluation for these experiments was done using 10 random splits of the test set A data into data sets consisting of 60% training data, 20% held out data for parameter selection, and 20% held out

Table 2. AUROC and minimum of positive predictive value and sensitivity for models using only the baseline features, with length 1 pattern features added, and with length 2 pattern features added.

Model	AUROC	Min(PPV, Sens)	Patterns Possible	Patterns Selected
Baseline	0.78	0.42	n/a	n/a
L = 1	0.81	0.45	144	56
L = 2	0.81	0.46	576	104
L = 3	0.82	0.45	2,304	149
L = 1,2,3	0.81	0.43	3,024	260

data for final evaluation. We assessed the AUROC and event 1 score (minimum of PPV and sensitivity) as the measure of improvement.

4. Results

4.1. Official results

Our approach achieved an event 1 score of 0.50, and an event 2 score of 36.63 on test set B and an event 1 score of 0.46 and an event 2 score of 56.45 on test set C.

4.2. Useful features

Table 1 shows the 10 features with the highest magnitude coefficients for models using patterns of lengths 1, 2, and 3.

Of the 10 features with the highest coefficients, many of the measurements are those used to measure systemic or organ dysfunction as in SOFA or the other severity scores. For example, with motifs of length one, moderate measurements of several variables such as platelets and temperature were negatively associated with mortality as would be expected (i.e., extreme values in either direction were found to be unhealthy). Another interesting result is that serum blood urea nitrogen (BUN) was highly associated with in-hospital mortality, despite both APACHEII and SAPSII choosing serum creatinine level [5, 6]. Additionally, in the case of motifs of length one, information about whether pH was missing was negatively associated with mortality. This can most likely be attributed to healthier patients not having the lab work done that would yield that measurement. We also found that age was highly associated with in-hospital mortality for all three motif lengths; this is not surprising given that age is consistently associated with mortality in many application domains and features in both APACHEII and SAPSII.

Our experiments showed that physiological motifs had increased importance in models when motif length was longer. For example, in the experiment with a motif length of three 6 of the top 10 variables in the models were physiological motifs. We believe this reflects the importance of sequential changes in ICU variables rather than snapshot values of these variables as points of time. We also interpret motifs of the kind missing-missing-high as being patient progressions where individuals deteriorated to the level where diagnostic tests or interventions had to be performed. For example, the motif of PaO2 missing-missing-high may likely correspond to patients who deteriorated to the point that they had to be put on a ventilator. There are several other instances of transitions from a value to missing or missing to a value that are positively associated with mortality.

The use of pattern frequencies of any length improved the AUROC and minimum of positive predictive value and sensitivity over the model using only the baseline features. There was little improvement in using longer patterns (L=2 or L=3) over the shorter patterns. Using all patterns of length 3 or less did not improve performance over the use of each pattern length individually. While the number of possible patterns increases exponentially with length, the number of patterns selected grows much slower.

5. Discussion

In this paper, we explored the hypothesis that there is clinically useful information in short patterns within ICU time series to predict in-hospital mortality. The results of our experiments performed on the CinC/Physionet 2012 Challenge data sets support this hypothesis and demonstrate that time series motifs may provide information that is complementary to baseline variables comprising clinical scoring systems such as SAPS, APACHE and SOFA. Importantly, in addition to empirical improvements over these existing scoring systems, a motif-based approach is also readily interpretable and can be used to compactly appreciate changes in patient physiology and interventions with prognostic implications.

While the results of our work are encouraging, we conclude with a brief discussion of some limitations of our study. We observe that our experiments explored a fairly narrow range of motif lengths. Our decision to limit our experiments to a motif length of three helped improve computational efficiency and also restricted the effects of multiple hypotheses when exploring many different motifs. However, it is possible that larger motif lengths (e.g., 4 to 12) may have led to more substantial improvements in predicting in-hospital ICU mortality. Related to this is the observation that our use of a Bonferroni correction may have been unnecessarily conservative and led to important patterns being ignored. We also note that while SVMs are widely used in many application domains, our use of a simple 2-class SVM may have failed to exploit additional improvements (e.g., available through L1-regularization) in discrimination of patients at risk of in-hospital ICU mortality. Finally, we also believe that there may be opportunities to improve performance by modifying other aspects of our study (e.g., using windows of time series features other than 2 hours; using more symbols than just four etc.)

References

[1] Chia C, Syed Z. Computationally generated cardiac biomarkers: Heart rate patterns to predict death following coronary attacks. Proceedings of SIAM International Conference on Data Mining, 2011.

[2] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 1947;18(1):50–60.

[3] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ January 1995;310(6973):170.

[4] Vapnik V. The nature of statistical learning theory. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[5] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical Care Medicine October 1985;13(10):818–29.

[6] Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA The Journal of the American Medical Association 1993;270(24):2957–63.

[7] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine 1996; 22(7):707–710.

[8] Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers 1999;61 – 74.

Address for correspondence:

Sean McMillan
2260 Hayward Street
Ann Arbor, MI 48109
spmcmill at umich.edu