Imputation-Enhanced Prediction of Septic Shock in ICU Patients

Joyce C. Ho
Cheng H. Lee
Joydeep Ghosh

ECE Department, University of Texas at Austin

HI-KDD '12, August 12, 2012, Beijing, China

Outline

Introduction

What is sepsis?

- Sepsis is a systemic inflammatory response to infection
- Septic shock is sepsis-induced hypotension despite adequate IV fluid

Motivation

- 11^{th} leading cause of death in 2010[1]
- Estimated $14.6 billion spent on sepsis hospitalizations in 2008[2]
- Intensive Care Unit (ICU) bed utilization as high as 25%
- Mortality rate of 12.8% for sepsis and 45.7% for septic shock

---

[1] http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_04.pdf

[2] http://www.cdc.gov/nchs/data/databriefs/db62.pdf

Predicting Sepsis and Septic Shock

Two-stage approach

1. Can we predict the development of sepsis with only measurements available from the first 2 hours of ICU admittance?
2. Given a patient has sepsis, can we predict potential complications at least an hour prior to the onset of septic shock?

Model Objectives

- Generalizes to all ICU patients
- Uses simple and accessible approaches
- Focuses on commonly observed, non-invasive clinical measurements
- Relies on intermittent and noisy clinical data
- Addresses the problem of patients with partially missing observations

Patients

Patients gathered from Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) database, one of the most extensive publicly available ICU data resource

Patient Study Size

	Positive Cases	Total
Sepsis	1,310 ($\sim$10.8%)	12,179
Septic Shock	328 ($\sim$31.2%)	1,052

Missing Value Imputation

Conditional mean approach using patient's gender and age group

Matrix factorization-based approach[1]

- Singular Value Decomposition (SVD) - uses linear combination of k-eigenvalues to estimate missing values
- Probabilistic Principal Component Analysis (PPCA) - uses an Expectation-Maximization (EM) approach for principal component analysis (PCA) assuming normal probabilistic distribution
- Bayesian Principal Component Analysis (BPCA) - similar to PPCA, except likelihood of estimated value is calculated using a Bayesian estimation method

[1] BioConductor `pcaMethods` package in R

Sepsis Prediction Features

Clinical history (H)
Demographics: Gender, age, height, weight
ICU unit: Care, service
Clinical scores: SAPS-I, SOFA

Initial physiological state (P)
Cardiac: Heart rate, systolic bp, diastolic bp
Respiratory: Respiratory rate, SpO_2
Other: Temperature

Sepsis Prediction with Logistic Regression Models

AUC comparison using non-imputed and imputed feature sets

	None	Imputed
All H	0.791$\pm$0.004	0.792$\pm$0.003
Stepwise H	0.790$\pm$0.003	0.791$\pm$0.004
All $\mathrm{H} \cup \mathrm{P}$	0.821$\pm$0.002	0.822$\pm$0.002
Stepwise H $\cup$ P	**0.823$\pm$0.002**	**0.823$\pm$0.002**

H = clinical history feature set
P = initial physiological state feature set

Imputation by means, SVD, and BPCA yielded identical AUCs

Septic Shock Prediction Features

Clinical history feature set (H)

Physiological state (recent and baseline values)
Cardiac: Systolic blood pressure, heart rate, pulse pressure[1]
Respiratory: Respiratory rate, $\mathrm{SpO_2}$
Other: Temperature, shock index[2] (recent only)

Laboratory history (recent value)
White blood count
Arterial pH

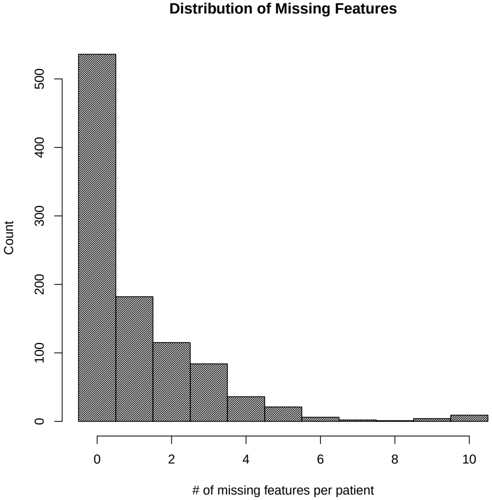
$\sim$ 6.7% of all observations are missing

$\sim$ 47.2% patients have at least one missing observation

[1] pulse pressure = systolic blood pressure - diastolic blood pressure
[2] shock index = heart rate / systolic blood pressure

Physiological Measurement Availability

Feature	% missing
RR	0.30
SpO_2	0.40
TEMP	1.10
Mean SBP	1.20
Mean PP	1.20
Mean HR	1.20
Mean RR	1.71
Mean SpO_2	2.11
Weight	4.92
WBC	5.82
Mean TEMP	6.43
SOFA	25.20
SAPS-I	28.92
APH	29.62

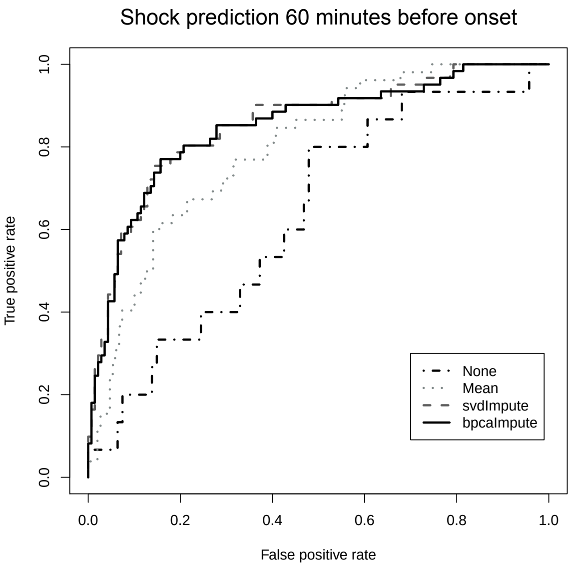


Shock Prediction Results with Logistic Regression Models

AUC comparison using varying imputation methods and prediction times

	Time t before onset		
Impute Method	30 min	60 min	120 min
None	0.764±0.050	0.728±0.041	0.699±0.045
Mean	0.822±0.026	0.798±0.025	0.751±0.029
SVD	0.854±0.010	0.848±0.007	0.831±0.018
PPCA	**0.855±0.009**	**0.852±0.007**	**0.828±0.016**
BPCA	0.853±0.010	0.847±0.007	0.829±0.017

Imputation Effects on Shock Classification Performance



Shock Prediction Results with Various Models

Three basic models

1. Full-featured Logistic Regression (LogR) with no regularization
2. Support Vector Machine (SVM) using a linear kernel
3. Recursive partitioning (RPART) or classification tree

AUC comparison of models 60 minutes before onset

	Model Type		
Impute Method	LogR	SVM	RPART
None	0.728±0.041	0.599±0.326	0.613±0.063
Mean	0.798±0.025	0.861±0.209	0.731±0.026
SVD	0.848±0.007	0.793±0.247	0.869±0.026
PPCA	**0.852±0.007**	**0.874±0.218**	**0.873±0.018**
BPCA	0.847±0.007	0.882±0.190	0.867±0.020

Conclusion

Accurate prediction of sepsis and septic shock using

- noisy and intermittently gathered clinical data
- common and primarily non-invasive clinical observations

Missing data imputation improves model performance

Matrix factorization imputation yields better predictive accuracy

Previous / Related Work

Sepsis

1. Multilayered perceptron with clinical factors and expression levels of immune system related genes
2. SVM on high-resolution physiological waveform data
3. Decision tree leveraging difficult or overly invasive clinical variables

Septic Shock

1. Early warning system (EWS) based on logistic regression and invasively-gathered waveform data
2. RPART approach for hospitalized non-ICU patients using routine laboratory tests and specific vital signs
3. Neural network in conjunction with center of gravity-based imputation utilizing laboratory tests and invasive catheter measurements