Developing an Early Warning System for Sepsis

Chloé Pou-Prom[1], Zhen Yang[1,2], Maitreyee Sidhaye[1,2], David Dai[1]

[1] St. Michael's Hospital, Toronto, Canada
[2] University of Toronto, Toronto, Canada

Abstract

Sepsis is a life-threatening condition that is caused by infection, and is estimated to affects an estimated 1.7 million adults in the United States and contributes to 265,000 deaths annually. Identifying sepsis before it happens and treating it earlier leads to decreased mortality and decreased lengths of stay. As part of the PhysioNet/Computing in Cardiology Challenge 2019, we developed an ensemble-based approach for the early detection of sepsis in ICU patients.

Our final model predicted sepsis using the previous 24 hours of data, and consisted of a combination of two convolutional neural networks and a random forest trained on different subsets of the data. In training our models, we experimented with random undersampling and cluster-based undersampling as a means for addressing severe class imbalance. On validation data, our final model achieved a utility score of 0.432 on hospital A (AUROC: 0.794, AUPRC: 0.101), 0.247 on hospital B (AUROC: 0.816, AUPRC: 0.056), and a utility of 0.375 on combined data from both hospitals (AUROC: 0.809, AUPRC: 0.089). On the heldout test data, the model obtained a utility score of 0.378.

1. Introduction

Sepsis is a serious medical condition that occurs when the body amounts an overwhelming immune response to an infection. The immune response cascades into systemic inflammation, causing restricted blood flow to organs and tissues, ultimately leading to organ damage [1–3]. Sepsis can happen to anyone, although it is more common in people with serious medical conditions and comorbidities [4]. Consequently, sepsis is a major concern for hospital inpatients, particularly since untimely treatment of sepsis can lead to an increased length of stay, morbidity, and mortality [5].

In the United States, sepsis affects an estimated 1.7 million adults and contributes to 265,000 deaths annually [6]. In 2013, sepsis-related hospital costs in the US totaled $24 billion, representing 6.3% of the total hospital costs [7]. These cost estimates have consistently increased in the last two decades, with recent estimates reporting a 5-fold increase in inflation-adjusted cost spending compared to 1997 [8].

Timely treatment of sepsis for hospital inpatients is important for patient prognosis, however detection of sepsis is difficult as its early presentation can resemble many other clinical conditions. As part of the PhysioNet/Computing in Cardiology 2019 Challenge, we used electronic health records data from over 40,000 patients to build models for identifying sepsis before its onset. In this paper, we present our approach for detecting sepsis and discuss a method for informed subsampling in the presence of severe class imbalance.

Imbalanced data is a ubiquitous problem in healthcare data. Approaches for dealing with class imbalance involve weighting the loss function, undersampling the majority class, or oversampling the minority class (e.g., SMOTE [9]). We sought to explore this further and focus on undersampling. Here, we present our approach for detecting sepsis using clinical data from the PhysioNet/Computing in Cardiology Challenge 2019 [10]. Our final model to the challenge consisted in an ensemble-based approach trained using random- and cluster-based undersampling.

2. Methods

2.1. Data and pre-processing

The challenge dataset consisted of time-invariant demographic features (e.g., age, gender), and vital signs (e.g., heart rate, temperature) and laboratory values (e.g., calcium, lactic acid) sampled at an hourly level from two different hospitals (hospital A and hospital B) [10].

We prepared the data into time windows, as shown in Figure 1. The time-dependent features had varying degrees of missingness, ranging from 13% to 100%. For each patient, missing values were imputed by applying last observation carried forward. Where that was not possible (i.e., missing values that occur before the first observation in the visit), the missing value was filled with a value of -1. Additionally, we created indicator variables for each fea-

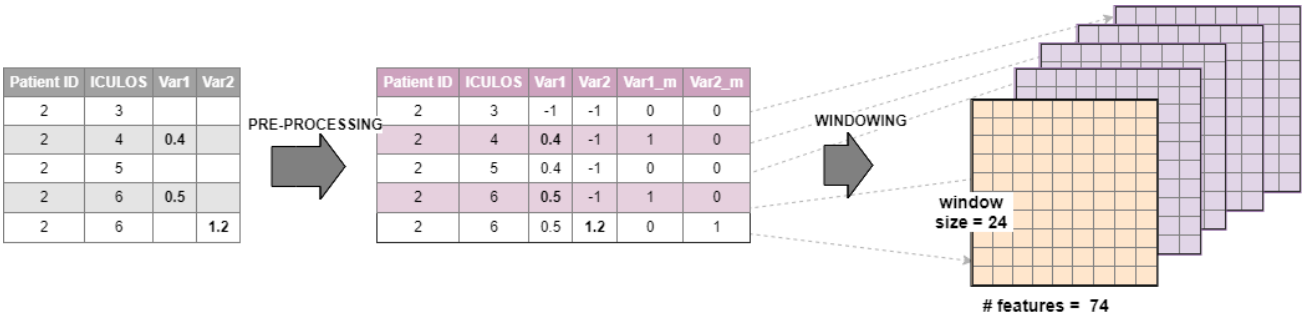

Figure 1. Data preparation pipeline. For each patient, we pre-processed the data with 'last observation carried forward' and filled remaining missing values with -1. We also created indicator variables for each feature, to identify if a feature was measured or imputed in that time interval. Finally, for each row of data, we create windows consisting of the previous 24 hours of data. Each data point consists of a window of dimension (window size = 24) by (# features = 74).

ture, denoting whether the value was observed or imputed within each hour (i.e., a 1 if the value was observed and a 0 if it was imputed). Finally, we created running windows of data (with window size = 24), with earlier rows filled in with 0's (i.e., each data point includes up to the previous 24 hours of the patient's visit information).

2.2. Unsupervised clustering and subsampling

To explore the heterogeneity of the hospital inpatient population, we applied K-means clustering to the set of demographic and physiological features of all patients. Since K-means is based on Euclidean distances and distance metrics may not be meaningful in high-dimensional space [11], we applied Principal Components Analysis to the original features to obtain a lower dimensional representation of the data. We applied K-means clustering (with k = 5) to the first two principal components (PC1 and PC2, representing 70% of the total variance) for each observation. Figure 2 visualizes the distribution of the five clusters in the 2-dimensional space of PC1 and PC2. The clusters were then used to inform sampling.

At a patient-level, sepsis occured in 8.8 % of patients in hospital A and 5.7 % of patients in hospital B (combined: 7.3 %). When we accounted for the temporality of the data and looked at the window-level, sepsis occurred in 2.2 % of windows in hospital A and 1.4 % of windows in hospital B (combined: 1.8 %). Given the large size of the dataset and the severe class imbalance in the outcome, we subsampled the majority class to increase model training speed and to improve model representation for the minority class.

After merging the datasets from hospital A and hospital B, we then split the combined data into training and validation data using an 80/20 split. We used two different undersampling techniques and created various subsets of the training data.

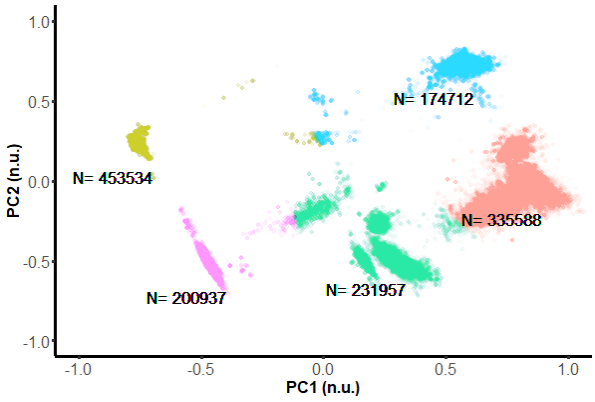

Figure 2. Cluster visualization in Principal Component space. We applied Principal Component Analysis (PCA) to the data and then clustered the top two components (PC1 and PC2) - accounting for 70% of the total variance - using k-means clustering (k = 5).

- **Random sampling**: From the non-septic windows of the training data, we sampled at random.
- **Cluster-based sampling**: Each non-septic window of the training data was assigned to a cluster based on our k-means cluster model trained on the first and second principal components. We sampled randomly and with replacement *from each cluster*, until we had an equal number of data points from each cluster. Our intuition was that windows within the same cluster would be similar to each other. As such, when undersampling, we wanted to ensure we were sampling in a way that gave an adequate representation of the majority class.

All data from the minority class (i.e., windows that experience sepsis) were retained. We tested sampling ratios of 1:2, 1:5, 1:10 and 1:20. For example, for a ratio of 1:2, we ensured that there were twice as many windows without sepsis (majority class) than windows with sepsis (minority class).

Dataset	AUROC	AUPRC	Accuracy	F-measure	Utility
A	0.794	**0.101**	0.761	**0.126**	**0.432**
B	**0.816**	0.056	**0.863**	0.094	0.247
Combined	0.809	0.089	0.772	0.105	0.375
Evaluation	–	–	–	–	0.378

Table 1. Validation results results. We report the area under the receiver-operator curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, F-measure, and utility score. Metrics are for hospital A, hospital B, the combined data (i.e., hospital A and hospital B), and on the heldout evaluation data

2.3. Models

We trained convolutional neural network and random forest models on the different subsets of the training data (i.e., random sampling vs "cluster"-based sampling of majority class; different sampling ratios) as described above. We then ensembled the different CNN and RF models together through logistic regression.

Convolutional Neural Network (CNN): We implemenedt a convolutional neural network (CNN) in `keras` [12]. The CNN architecture consisted of up to three two-dimensional convolution layers (we experimented with 1, 2, and 3 layers) with kernel size 3×3 and ReLU activation . Max pooling and dropout were applied after the convolution layer(s) with one fully connected layer for the output. We trained the model with the Adam optimizer (using default parameters) and early stopping based on the validation loss. The CNN took as input the 24-hour window of pre-processed features (i.e., dimension of 24×74).

Random Forest (RF): The random forest consisted of 100 estimators and the `scikit-learn` default parameters [13]. The RF took as input a flattened representation of the 24-hour window (i.e., a vector of dimension $24 \cdot 74 = 1,776$).

Ensemble model - logistic regression (LR): The best CNN and RF models were combined using a logistic regression model. The output probabilities of each model were used as input features to the ensemble. We experimented with different combinations of RF and CNN trained on different subsets (varying in sampling approach and ratios). Our final submitted model was a logistic regression ensemble of the following:

1. RF trained on a subset of the training data, sampled *randomly* and with a *1:2* ratio
2. CNN with 1 convolution layer trained on a subset of the training data, sampled *randomly* and with a *1:2* ratio.
3. CNN with 1 convolution layer trained on a subset of the training data, sampled *based on K-means clusters* and with a *1:2* ratio.

3. Results

We used an 80/20 train/validation split on the data and report our ensemble model results in Table 1. Our model achieved the best utility score with data from hospital A, but had better AUROC and AUPRC with data from hospital B. On the heldout evaluation data, the ensemble model achieved a utility score of 0.378.

When building our final ensemble model, we explored different undersampling methods. For each method and sampling ratio, we sampled the training data 10 times. We report results of the 1-layer CNN and RF in Table 2. The results consist of the AUROC, AUPRC, and utility scores on the combined validation data averaged over the ten subsets. Generally, using a ratio of 1:2 achieved better result and the random undersampling performed better than the cluster-based sampling in all scenarios.

4. Discussion and Conclusions

We present our ensemble model for the PhysioNet/Computing in Cardiology 2019 challenge. Our final model consisted in an ensemble of a random forest and two convolutional neural networks trained on different subsets of the data. In our experiments, we tried random undersampling and a cluster-based sampling approach, to address the class imbalance.

We hoped that using cluster-based sampling would better capture the population from the majority class. In this set of experiments, using cluster-based sampling alone did not yield the best results. However, we found using it in combination with other models helped in the final ensemble. Our cluster-based sampling was fairly simple - each data point was assigned to a cluster. A better approach would be to take into account cluster distance, since our current undersampling technique did not account for within-cluster differences. In the future, we would like to explore methods that better leverage the distance information obtained by unsupervised clustering.

Furthermore, our current approach was limited in that we only used 24 hours of data to make a prediction. We reasoned that measures from within a day would be sufficient to capture enough information.

		CNN			RF		
Sampling	**Ratio**	**AUROC**	**AUPRC**	**Utility**	**AUROC**	**AUPRC**	**Utility**
Random	1:20	0.790 (0.005)	0.079 (0.003)	0.352 (0.009)	0.753 (0.004)	0.064 (0.001)	0.238 (0.005)
	1:10	0.787 (0.007)	0.078 (0.003)	0.346 (0.014)	0.751 (0.004)	0.064 (0.001)	0.267 (0.005)
	1:5	0.797 (0.007)	0.080 (0.005)	0.360 (0.011)	0.747 (0.004)	0.064 (0.001)	0.261 (0.006)
	1:2	0.797 (0.009)	0.080 (0.005)	0.355 (0.011)	0.743 (0.002)	0.067 (0.001)	0.271 (0.004)
Cluster	1:20	0.779 (0.005)	0.077 (0.004)	0.339 (0.009)	0.743 (0.006)	0.063 (0.001)	0.254 (0.005)
	1:10	0.784 (0.010)	0.078 (0.004)	0.343 (0.015)	0.742 (0.004)	0.063 (0.001)	0.246 (0.006)
	1:5	0.783 (0.008)	0.074 (0.004)	0.340 (0.014)	0.734 (0.002)	0.063 (0.002)	0.249 (0.008)
	1:2	0.786 (0.007)	0.078 (0.003)	0.345 (0.012)	0.732 (0.001)	0.064 (0.001)	0.225 (0.005)

Table 2. Experiments across sampling techniques and with different ratios. The ratios correspond to the prevalence of non-septic 24-hour windows of data to septic 24-hour windows of data. For each sampling technique and sampling ratio, we sample 10 times and report the averaged AUROC, AUPRC, and utility scores (with standard deviation in parentheses).

Acknowledgments

We thank Sebnem Kuzulugil, Joshua Murray, Greg Arbour, Michaelia Young, Kasthuri Karunanithi, and Neal Kaw for insightful discussions on the challenge. This work is funded by the Li Ka Shing Foundation.

References

[1] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):762–774.

[2] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):801–810.

[3] Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, Angus DC, Rubenfeld GD, Singer M. Developing a new definition and assessing new clinical criteria for septic shock: for Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):775–787.

[4] Gaieski D, Edwards J, Kallan M, Carr B. Benchmarking the incidence and mortality of severe sepsis in the United States. Critical Care Medicine May 2013;41(5):1167–1174.

[5] Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M. Early goal-directed therapy in the treatment of severe sepsis and septic shock. New England Journal of Medicine 2001; 345(19):1368–1377. PMID: 11794169.

[6] Dantes RB, Epstein L. Combatting Sepsis: A Public Health Perspective. Clinical Infectious Diseases 05 2018; 67(8):1300–1302. ISSN 1058-4838.

[7] Healthcare cost and utilization project - statistical brief 204. URL https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.jsp.

[8] Healthcare cost and utilization project - facts and figures 2009. URL https://www.hcup-us.ahrq.gov/reports/factsandfigures/2009/exhibit4_1.jsp.

[9] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 2002;16:321–357.

[10] Reyna MA, Josef C, Jeter R, Shashikumar SP, M. Brandon Westover MB, Nemati S, Clifford GD, Sharma A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine 2019;In press.

[11] Aggarwal CC, Hinneburg A, Keim DA. On the surprising behavior of distance metrics in high dimensional space. SpringerLink Jan 2001;URL https://link.springer.com/chapter/10.1007/3-540-44503-X_27.

[12] Chollet F, et al. Keras. https://keras.io, 2015.

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

Address for correspondence:

Chloé Pou-Prom
St. Michael's Hospital, 2 Queen Street East
PoupromC@smh.ca