We build an **ensemble** model to predict sepsis in the **PhysioNet/CINC Challenge 2019** dataset. Our model is trained with different undersampling methods and achieves a **utility score of 0.378** on the heldout evaluation data.

Developing an Early Warning System for Sepsis

Chloé Pou-Prom [1*]
poupromc@smh.ca
Zhen Yang[1,2] Maitreyee Sidhaye[1,2] David Dai[1]

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

1 Introduction

– **Sepsis** is a life-threatening condition that is caused by infection. Identifying sepsis before it happens and treating it earlier leads to decreased mortality and decreased lengths of stay.

– Imbalanced data is a ubiquitous problem in healthcare data. We explore this further and focus on **undersampling**.

– Our submission to the **PhysioNet 2019** challenge is an ensemble model trained using random- and cluster-based undersampling. We achieve a **utility score of 0.378** on the evaluation data.

2 Methods

Data and Pre-processing

– Challenge datasets consist of demographics, vitals, and laboratory values sampled at an hourly level from two different hospitals (hospital *A* and hospital *B*) (Reyna et al. 2019).

– We impute missing data with **last observation carried forward**, and fill remaining missing values with -1.

– We create **indicator variables** to differentiate measured features from imputed features.

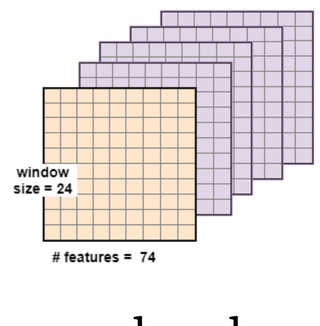
– For each row, we create **24 hour** windows of data. Earlier windows are filled with 0's.

Patient ID	ICULOS	Var1	Var2
2	3		
2	4	0.4	
2	5		
2	6	0.5	
2	6		1.2

↓

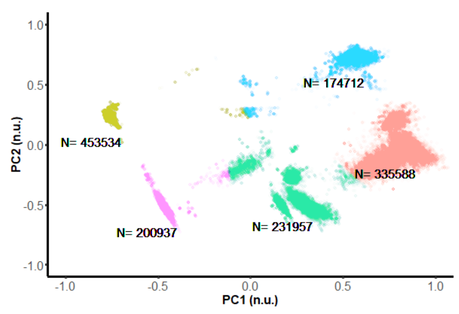
Patient ID	ICULOS	Var1	Var2	Var1_m	Var2_m
2	3	-1	-1	0	0
2	4	0.4	-1	1	0
2	5	0.4	-1	0	0
2	6	0.5	-1	1	0
2	6	0.5	1.2	0	1

↓



Clustering and undersampling

– To address **class imbalance**, we undersample the majority class (i.e., windows that don't experience sepsis) by sampling **randomly** or based on **clusters**.



– For cluster-based undersampling, we train **k-means** on the first two **PCA** components, and we sample equal number of data from each cluster of the majority class.

– *Intuition*: Data from the same cluster are similar to each other and we want an adequate representation of the majority class.

Models

– We train **convolutional neural network** (CNN) and **random forest** (RF) models on different subsets of the data, varying in sampling method (random vs. cluster) and ratio of sepsis:non-sepsis.

– Our final model is an ensemble (**logistic regression**) which takes as input the scores of the following models: 1) RF trained on *random* subset, 2) CNN trained on *random* subset, and 3) CNN trained on subset sampled *based on K-means clusters*.

– The data for training the models were all sampled at a ratio of *1:2*.

3 Results

– Results on validation data (80/20 split).

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
Heldout	–	–	–	–	0.378

– We run different experiments with undersampling and find that *random-based undersampling performs better than cluster-based undersampling*, however *including models trained on a cluster-based sample of the data in our ensemble lead to better results* (refer to preprint for details).

4 Conclusion

– We use an ensemble-based approach for predicting sepsis in ICU hospital patients.

– For the PhysioNet dataset, cluster-based undersampling is useful as part of an ensemble, but not on its own.

– **Future work**: Account for distance in cluster-based sampling.

Acknowledgements

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. Poster was made with Posterdown (Thorne 2019).

References

Reyna, Matthew A., Chris Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. 2019. "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019." *Critical Care Medicine* In press.

Thorne, W. Brent. 2019. *Posterdown: An R Package Built to Generate Reproducible Conference Posters for the Academic and Professional World Were Powerpoint and Pages Just Wont Cut It.* Vol. 0.1.2. https://github.com/brentthorne/posterdown.