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*A

Zhen Yang*B

Maitreyee Sidhaye*C

David Dai*A

*St. Michael’s Hospital, Toronto, Canada

†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.

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.

3 Results

Results on validation data (80/20 split).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Accuracy</th>
<th>F-measure</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.794</td>
<td>0.101</td>
<td>0.761</td>
<td>0.326</td>
<td>0.432</td>
</tr>
<tr>
<td>B</td>
<td>0.816</td>
<td>0.056</td>
<td>0.863</td>
<td>0.094</td>
<td>0.247</td>
</tr>
<tr>
<td>Combined</td>
<td>0.809</td>
<td>0.089</td>
<td>0.772</td>
<td>0.105</td>
<td>0.375</td>
</tr>
</tbody>
</table>

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, Jamie Moore, Greg Arbour, Michaelia Young, Kasthuri Karunanithi, and Neal Kao for insightful discussions on the challenge. This work is funded by the Li Ka Shing Foundation. Poster was made with Posterdown [Thorne 2019].

References
