Sepsis Detection in Sparse Clinical Data Using Long Short-Term Memory Network with Dice Loss

\*Abstract

Feature normalization result : 0.281 normalized utility

1. Introduction

Machine learning Method used :

SVM

Hidden Markov Model

XGBoost

random forest

LSTM

NLP

clinical time-series with multi-output Gaussian processes and fed latent variables into a RNN to classify the patient as septic or non-septic.

static information extracted from information about patient with fully connected (FC) network and dynamic information extracted from time-series with a combination of

convolutional neural network and LSTM

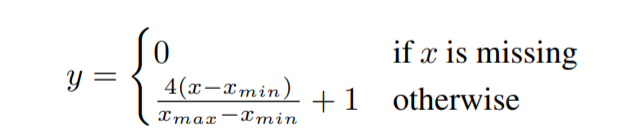
Models are Deal with a high portion of missing values in the provided dataset and highly imbalanced dataset

1. Methods

2.1 Missing Value Problem

feature normalization into the fixed range of values is applied including the replacement of missing values with numerical representation from outside the normalized range.

Range of all features = [1,5]



Preprocessing observation:

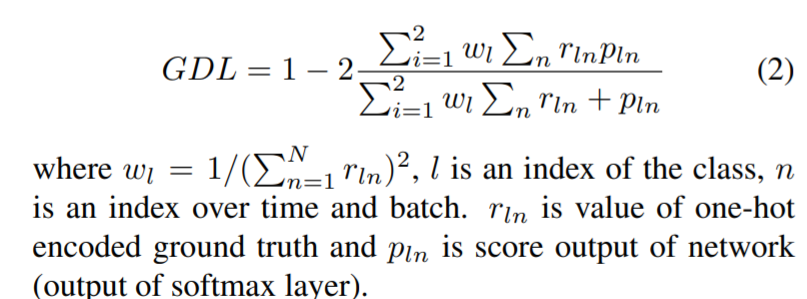
it was noticed that for some clinical features there are values not corresponding to physiological possibilities of the human body were removed

However, this data preprocessing did not provide any improvement. Even though it is believed that there should not be present data which do not make sense from the physiological view.

2.2 class imbalance problem

weighted loss function such as weighed cross-entropy or Generalized Dice Loss (GDL)

Math:



2.3 Network Architecture and Implementation

7 blocks each of LSTM layers followed by 3 fully connected layers.

Inspired by ResNet [16] and DenseNet [17], we add also residual skip connections

Input: Concatenation of output of the previous block, skip over the previous block and network input

Every fully connected block is followed by ReLU and dropout layer (with 0.5 drop probability)

Output softmax layer ensures the mapping of the output values into the range 0-1.

2.4 Threshold Adjustment

grid search, where utility measure was maximized for the validation set.

(AS loss function does not guarantee that best threshold of output score with respect to the utility score will be 0.5. )

2.5 Training details and implementation

Listed clearly

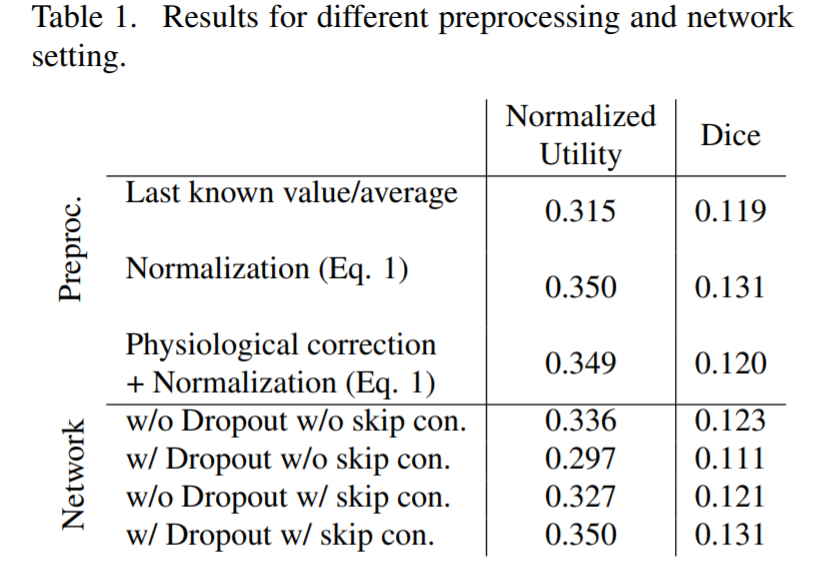
2.5 Result

The dataset contains data from 40336 patients, where 40 clinical features were recorded for every patient by 1-hour interval. The training dataset was randomly divided into internal training (90 %) and validation (10 %) sets.

Sepsis labels are shifted by 6 hours, thus algorithm should predict sepsis 6 hours before it starts. Results were evaluated in terms of dice coefficient and normalized utility (official challenge metric).

0.350 utility and 0.131 dice on our validation set, and 0.372 utility on challenge official partial test set A

0.350 utility and 0.131 dice on our validation set, and 0.372 utility on challenge official partial test set A



use of the network with both dropout and skip connections leads to significant improvement. This comparison was performed for normalization by Eg. 1 and for the network with 7 blocks

**Github Code:**

<https://github.com/ECGuru10/PhysioNetCHallenge2019>