Using Features Extracted From Vital Time Series for Early Prediction of Sepsis

Qiang Yu, Xiaolin Huang*, Cheng Wang, Yun Ge

School of Electronic Science and Engineering, Nanjing University, China

*Corresponding author: xlhuang@nju.edu.cn

We trained a basic random forest model using set A, then tested it on set B and got a $U_{normalized}$ of 0.174, a F measure of 0.102, an AUROC of 0.67, an AUPRC of 0.037 and an Accuracy of 0.926.

Main specific processing includes:

Firstly, as for NAN value replacement, we took two complementary options: when there is a rational value for the feature within three hours prior to or posterior to this moment, we took the most recent value as the replacement; otherwise, the mean value was used.

Secondly, besides the original 40 features, we extracted eight one-hour-change features from vital signs, which were regularly sampled per hour. Thus, the random forest model has 48 input features in total.

Thirdly, for each file in the training set, we located the moment at which the original label changes from 0 to 1 (if there is), and then, we modified the preceding one to six hours' label to 1, and the seven to ten hours' label to 0.8, 0.6, 0.4 and 0.2 respectively. The modified labels were used as the target in training.

Finally, in test, the trained random forest outputs the probability that the input is assigned to sepsis. To get a 0 or 1 label, we set the threshold as 0.1.

Although current scores are not good enough, there are much we can try to improve:

First, except column 40, demographic features are time-nondependent, which is quite different from what vital signs and laboratory values are. Therefore, in a more efficient model, time-dependent features and time-nondependent features should be treated differently.

Second, more information about the time evolvement of vital signs should be extracted and efficiently used.

Last but not least, a more specialized classifier model as well as an appropriate optimization method should be constructed and used.