Predictive Modeling for Differential Diagnosis and Mortality Risk Assessment

Tony Lindsey PhD[1], Sena Veazey[2], Saul Vega[2] and Jose Salinas PhD[2]

[1]NASA Ames Research Center, Moffett Field, CA
[2]US Army Institute of Surgical Research, San Antonio, TX

ABSTRACT

The prevalence of electronic health record (EHR) storage systems has created prodigious biomedical informatics opportunity. Automated machine learning methods are effective at analyzing such data and have become common tools for healthcare predictive modeling. Medical informatics researchers have investigated the potential of deep learning and classical models applied to emergent care scenarios. In particular, differential diagnosis (DDX) prediction for admitted patients has proven useful in reducing superfluous lab tests and improving inpatient triage decision-making. *BICEPS is the current US military treatment route and emphasizes DDX by severity of combat stress reaction symptoms. Moreover, identification of high-risk mortality patients in extreme environments such as combat support hospitals is vitally important for cost-effective medical resources allocation.

The Medical Information Mart for Intensive Care (MIMIC-III) is an openly available dataset developed by the MIT lab for computational physiology and comprises de-identified critical care inpatient data. The repository was utilized in our study, contains hospital patient laboratory measurements, pharmacologic prescriptions, diagnostic data and procedure event recordings. When considering adult patients and discounting admissions with ICU length of stay less than 24 hours, there were 37,787 unique admissions and 30,414 total patients.

We examined the top 25 most frequent ICD-9 group-level disease specificities in MIMIC-III using a multi-label classification model. In-hospital mortality was modeled as a binary classification task with 4,155 (13%) of the adult patient population expiring, from which 3,138 (75.5%) resided in the ICU setting. The metrics AUC, F1 score, sensitivity and specificity values measured prediction performance and were calculated for each disease label.

The usage of ICD-9 group codes reduced feature dimensions from 14,567 to 942 and greatly improved distribution of patient diagnostic categories. Disease temporal patterns were captured by considering the 6 most frequently sampled vital signs and top 13 commonly sampled laboratory values. Missing data was replaced at each time-stamp by a form of hot-deck imputation called “last observation carried forward”. Time-series raw hourly average values were converted into 5 summary features (mean, standard deviation, number of observations, min & max values). Patient demographic variables such as age, gender, marital status and ethnicity were also factored into the modeling.

Choi et al showed that contextual embedding of medical data, diagnostic and procedural codes alone can predict future diagnoses with sensitivity as high as 0.79. We utilized an embedding technique called *word2vec* which allowed sparse representations of medical history to be transformed into dense word vectors. The mappings captured contextual information by treating each admission as a sentence and learning the most likely neighboring words in a sliding window fashion.

Binary and multi-label classification was achieved via collapse models, which do not consider temporal information, as well as recurrent neural networks (RNN) with regularization, *Softmax* output layer activation together with categorical cross-entropy as the loss function. See result Table 1 & 2 below.

DDX	Model	AUC / F1	Model	AUC / F1
Pathology	**Collapse**		**RNN**	
Coronary Artery Disease	MLP w/ x48	0.796 / 0.520	CNN w/x19+demo	0.792 / 0.480
Atrial Fibrillation	MLP w/ x48	0.745 / 0.400	LSTM w/x19+demo	0.768 / 0.341
Acute Kidney Failure	MLP w/ x48	0.885 / 0.505	CNN w/x19	0.862 / 0.485
Type II Diabetes Mellitus	MLP w/ x48	0.740 / 0.200	LSTM w/x19+demo	0.745 / 0.144
Hyperlipidemia	MLP w/ x48	0.750 / 0.170	CNN w/x19+demo	0.748 / 0.173

AUC – Area under receiver operating characteristic curve; F1 – Weighted average of precision & recall
MLP – Multi-layer perceptron; x48 – Last 48 hours of physiologic feature values; demo – demographics
CNN – Convolutional neural network; LSTM – Long short-term memory

Table 1 DDX - Five pathologies selected by frequency of occurrence.

Rank	Model	AUC	F1	SN	SP
Collapse models					
1	MLP w/ W48	0.852	0.544	0.875	0.833
2	RF w/ W48	0.841	0.520	0.862	0.820
3	GBC w/ W48	0.771	0.435	0.757	0.784
RNN models					
1	LSTM w/ x19 + h2v	0.948	0.621	0.881	0.885
2	CNN-LSTM w/ x19	0.938	0.632	0.851	0.893
3	CNN-LSTM w/ x19	0.931	0.585	0.852	0.866

RF – Random forest; GBC – Gradient boost classifier; SN – Sensitivity; SP – Specificity; W48 – Diagnostic history *word2vec* embedding + x48; x19 + h2v – Combined patient visit and demographic information-level representations.

Table 2 Mortality risk top model predictors.

Our results indicate that RNN models are most suitable for in-hospital mortality predictions, where temporal patterns of simple physiologic features are sufficient to capture mortality risk. Deep models in general out-perform collapse models for the differential diagnostic task. However, temporal information from RNN models didn't provide additional benefit when compared to MLP.

*BICEPS – brevity, immediacy, contact, expectancy, proximity, and simplicity.