A Comparative Analysis of HMM and CRF for Early Prediction of Sepsis

Saman Noorzadeh, Shahrooz Faghihroohi, and Mojtaba Zarei
Institute of Medical Science and Technologies, Shahid Beheshti University, Tehran, Iran

Abstract

This study aims to detect and predict sepsis precisely in ICU patients according to the data published for Physionet challenge 2019. Sepsis prediction can help in early intervention and therefore less mortality rate. Hidden Markov Model (HMM) is applied with the independence assumption of features; however, to tackle this problem, Linear-chain conditional random field (CRF) is implemented and the results are compared to HMM. The results show that CRF outperforms HMM in the early prediction of sepsis. The team of the authors, named IMSAT, ranked 50 in the mentioned challenge by gaining a utility score of 0.19.

1. Introduction

Sepsis is a dangerous crisis caused by infection in the body. This life-threading situation is among the leading causes of death in the intensive care unit (ICU) [1]. If sepsis progresses it can turn into sever sepsis, followed by organ failure and death. With early diagnosis, the rate of mortality can be reduced and the clinicians can apply early treatments.

Currently there are scoring methods for identifying and predicting sepsis patients like SOFA, qSOFA, MEWS, etc [2]. These methods are mainly based on analytic methods, they use a small pre-defined dataset, and lack generalizability [3]. That is why, the recent studies have focused on machine learning methods and they have shown that these methods can outperform the existing scoring systems [3] [4].

Some studies have used svm in this regard. For example in [5] lab data and biomedical signals and SIRS scores were used to create an SVM model to predict sepsis between 24 hours before diagnosis. In [6] svm is used to predict if a septic patient will develop sever sepsis. Logistic regression is another approach used in [7]. In another study it is shown that using Factor analysus to extract features before logistic regression perform better in the prediction of sepsis [8]. Neural networks are another recent approach which is used in [7]. More recently Fuzzy modeling is used for this aim. In [9] probabilistic Fuzzy models are used to predict the mortality in sepsis shock and they are proved to perform better than logistic regression and Neural Networks. As integration of models is of a great interest recently, some has used ensemble models to show that it can perform better than the single models [10]. Finally, a brief review of the existing methods can be found in [11].

In this study, we examined two popular probabilistic models to predict sepsis from the clinical data of physionet challenge 2019 [12]. The first model is Hidden Markov Model (HMM) combined with Logistic Regression (LR) to model the sequence of the multi-dimensional sepsis data. The other method is Conditional Random Field (CRF) which handles the multi-dimensionality of the data in a more efficient way.

The rest of the paper is organized as folows: in section 2, first the data is described and then we explain how the data is processed to be given to the mentioned methods. This includes the Feature selection, treating NANs, normalization, and data splitting. Afterwards, HMM and CRF are explained, their pros and cons are described, and their differences are compared methodologically, and their application on the data is studied. In section 3 the results are discussed by 10-fold cross validation and also the method is tested on a new dataset. All these are finally wrapped up in section 4.

2. Method

In this section, HMM and CRF are explained and their differences are described, and finally they are applied for the prediction of sepsis. Before that, the dataset used here is explained and the pre-process of the data and feature selection is explained.

2.1. Data

The data is taken from the PhysioNet Computing in Cardiology Challenge 2019 [12]. The dataset that is provided is from two hospitals and includes 40 features as 6 Demographics, 8 Vital Signs, and 36 Laboratory values of patients who are entered to the hospital Intensive care unit (ICU). The first dataset consists of 20336 patients, 1790 of whom had sepsis, and the second included 20000 patients with 1142 sepsis patients. For each patients the data consists of hourly records, and whether the patient has sepsis or not is labeled also at every hour by 1, and 0, respectively. For a detailed description of data please consult [12].

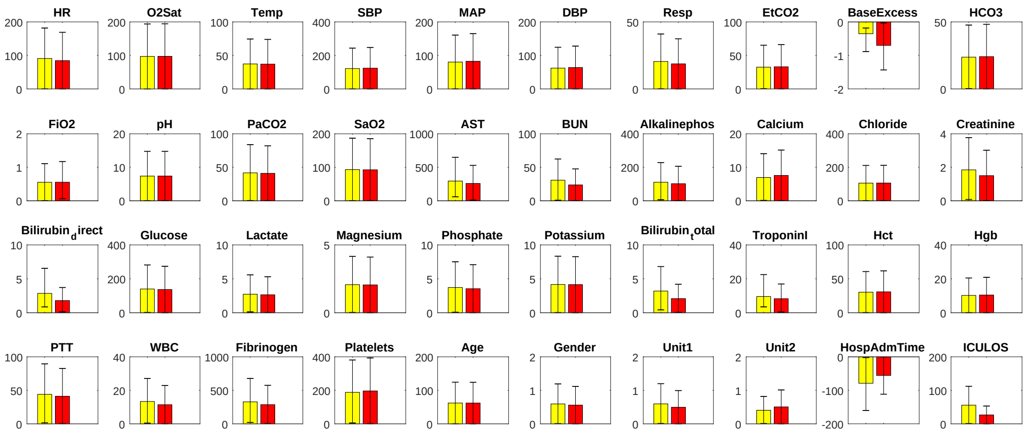

Figure 1. The comparison of the sepsis and non-sepsis data for all 40 features. The yellow and red bars represent the mean of the features of sepsis and non-sepsis features, respectively. A %95 confidence interval is also shown for every bar. 17 most significant ones are selected.

2.2. Feature Selection and Data Process

A feature reduction is first done to reduce the computational complexities. Consider the data for each subject i as $X^i = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_i}]^T$ where N_i is the number of hourly records for that patient, which can be different for different patients, and $\boldsymbol{x}_j$ is defined as $[x_1, \cdots, x_F]^T$, where F is equal to the number of features. For every feature, the data of sepsis and non-sepsis patients were compared using ttest. Supposing that the whole data includes P patients is denoted as $X = [X^1, ..., X^P]$. For every feature f, X^i_f is divided into sepsis and non-sepsis data and these two are compared by ttest. Since the number of sepsis patients are much fewer than the non-sepsis patients, non-sepsis patients were randomly sampled 1000 times, and ttest is done 100 times on these two sets and the p-values were averaged. Then 17 features with pvalues less than 0.1 were selected. A comparison of the features is shown in Fig. 1.

The 40 features of data which were collected hourly contains a considerable amount of missing data in the records. This can be due to the fact that the collection of all features at every hour is not an easy task for the hospital or that they are sometimes unnecessary according to the patients conditions. Therefore, for each patient's records in each feature, the first and last missing data were replaced by the first or last existing data, respectively, and the rest were linearly interpolated.

The data is then normalized. Let μ_i and σ denote the mean and the standard deviation of i^{th} feature ($i \in [1 \cdots, F]$) in the data X. Then for n^{th} patient, its corresponding data elements are normalized as follows:

$$X^n(h,i) \leftarrow \frac{X^n(h,i) - \mu_i}{\sigma_i},\ h \in [1, N_i] \text{ and } i \in [1, F] \tag{1}$$

The last stage is data split in which we split the patient's data into S segments with O overlap. The prediction of y_t is done in a way that the prediction at time t depends only on $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t]$. Meaning that the prediction does not take into account the future data. That is why we also have small observation sequence for the test phase. Therefore, data split is done so that the train data sequences size are not very high comparing the test data sequence. If the performance of the method calculated by crossvalidation, is considered a function of S and O, then these two variables can be calculated by maximizing such function. All the explained stages are depicted in Fig. 2.

2.3. Prediction of Sepsis

The valid assumption in the prediction of sepsis is the dependency of sepsis states in consecutive time frames. One of the most prominent group of statistical models which considers this property is Markov random fields (MRF). We have examined two well-known members of the MRF including Hidden Markov Model (HMM) and Conditional Random Fields (CRF). In the following section we briefly describe the basics of each method.

2.3.1. HMM and LR

Here, HMM is first order, and there are two states for sepsis and non-sepsis, and the observed sequence is the features of data. An HMM model consists of transition matrix, whose $(i, j)^{th}$ element defines the probability of going from state i to state j, the emission matrix, whose $(i, j)^{th}$ element defines the probability of being at state i when observing the observation j, and the initial vector, whose i^{th} element defines the probability that the initial

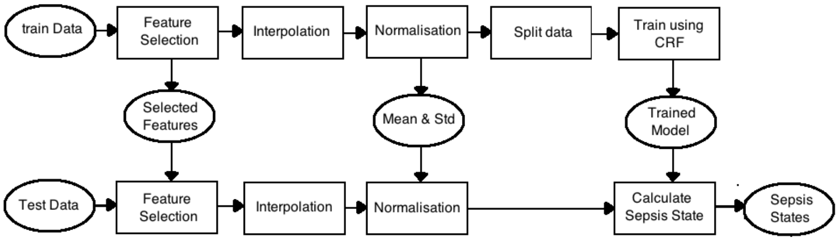

Figure 2. Block diagram of sepsis prediction using CRF method.

state of the sequence is state i. To predict the states the Viterbi algorithm is used. This algorithms finds the best state sequence given the model and the observations. The description of the Viterbi algorithm can be found in [13]. The transition matrix is estimated from the data, however the emission matrix is difficult to calculate because of the high dimensionlity of the data. Therefore, LR is applied to estimate the emission [14]. The LR model uses the logistic function to map the input features to values between 0 and 1. The logistic function is defined as logistic$(n) = \frac{1}{1+e^{-n}}$. In linear regression the relation between the output and the feature is modeled by $\hat{y}_i = \alpha_0 + \alpha_1 x_1^i + \cdots + \alpha_f x_f^i$, for i^{th} data sample, when the data has f features. For HMM, we need probabilities, so we use LR to force the outcome to be between 0 and 1:

$$p(y_i = s_1|\boldsymbol{x}_t) = \frac{1}{1 + e^{-\left(\alpha_0 + \alpha_1 x_1^i + \cdots + \alpha_f x_f^i\right)}} \qquad (2)$$

Where y_i denotes the state at time i which can be sepsis (denoted as s_1), or non-sepsis (denoted as s_2). Here because of the very large number of observations we consider the observations to have a uniform distribution meaning that $P(\boldsymbol{x}_t) = 1$. We also assume that the probability of states are the same $P(s_1) = P(s_2)$. So, the emission matrix is then calculated as $P(\boldsymbol{x}_t|y_t = s_j) = \frac{P(y_t=s_j|\boldsymbol{x}_t)P(\boldsymbol{x}_t)}{P(s_j)}$ and considering the assumptions: $P(\boldsymbol{x}_t|y_t = s_j) \sim P(y_t = s_j|\boldsymbol{x}_t)$.

2.3.2. CRF

Similar to the HMM, CRF are categorized as the family of sequence modeling methods. However, there is a major differences between them. The main distinction of CRF with respect to HMM is that HMM is a generative model which computes the actual distribution of each class. However, the CRF belongs to the discriminative class of models which finds the decision boundary between classes [15]. Mathematically, CRF computes the conditional probability $p(y|\boldsymbol{x})$ rather than the joint probability $p(\boldsymbol{x}, y)$ computed by HMM. This will be helpful to avoid computation of prior probability $p(\boldsymbol{x})$ which is not straightforward [16].

The type of CRF method used for the prediction of the sepsis is linear-chain CRF which is an extension of logistic regression for the classification of a sequence. In this approach, the feature functions exponents are multiplied to make the conditional probability distribution:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_0} exp(\sum_{i=1}^{n}\sum_{k=1}^{m} \alpha_k f_k(y_{i-1}, y_i, \boldsymbol{x}) + \sum_{i=1}^{n}\sum_{k=1}^{m} \beta_k g_k(y_i, \boldsymbol{x})) \qquad (3)$$

In this equation, Z_0 is the normalisation factor computes from all state sequences, $f_k(y_{i-1}, y_i, \boldsymbol{x})$ and $g_k(y_i, \boldsymbol{x})$ are feature functions, and α_k and β_k are weight parameters. The feature functions which are corresponding to the state and emission functions in HMM, are usually chosen as a binary function. The weight parameters of the model are learnt from the training data by maximizing the conditional log likelihood [17].

The algorithms such as forward-backward and Viterbi used for HMM, can also be applied to CRF. We also utilized the library of pystruct for the implementation of linear-chain CRF model.

3. Results

We dispose two datasets that we denote as dataset A and B which are from two hospitals. The sepsis should be predicted 6 hours before the onset of sepsis. First, HMM-LR and CRF are both applied on dataset A and the results of a 10-fold cross validation is shown in Table. 1. The results show that the accuracy of HMM is higher, however this is due to the fact that the two class population is very unbalanced and the higher sensitivity shows the great number of true negative. However by comparing sensitivity and F-measure we can conclude that CRF is a more trusted method in this regard. Since the prediction of sepsis in time (TP) is much more important than a false prediction of this crisis (FP), the challenge has proposed to use another score called normalized utility score in which the early prediction of sepsis is rewarded and the late prediction is penalized [12].

	HMM	CRF
Accuracy	0.714	0.52
F measure	0.08	0.28
Sensitivity	0.57	0.77
Specificity	0.72	0.39

Table 1. Results compared between HMM and CRF for 10-fold cross validation.

It has to be added that the authors competed with others in the Physionet challenge of 2019 [12]. The challenge tested the models on a full dataset of A, B and another one called C. Our team named IMSAT gained an official utility score of 0.19 on the full dataset which ranked the team 50^{th}. However, through this study, and the comparison made between HMM and CRF we expect a better performance for the CRF method.

4. Conclusions

In this study HMM and CRF are used separately to predict sepsis even hours before its onset in a clinical data. After the feature selection, interpolation of data for handling the missing data, normalization and data split the data is modeled by HMM and LR is used to have a better estimation of emission. Linear chain CRF is another sequence modeling method which is applied on the data. The results show that CRF is performing better than HMM because of the higher rate of specificity and F-measure.

Acknowledgements

This work was supported in part by grants from the Iran's National Elite Foundation.

References

[1] Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, Angus DC, Reinhart K. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. American journal of respiratory and critical care medicine 2016;193(3):259–272.

[2] McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. Annals of translational medicine 2016;4(24).

[3] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. Academic emergency medicine 2016;23(3):269–278.

[4] Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: a meta-analysis. Computer methods and programs in biomedicine 2019;170:1–9.

[5] Kim J, Blum J, Scott C. Temporal features and kernel methods for predicting sepsis in postoperative patients. Technical report, Citeseer, 2010.

[6] Wang SL, Wu F, Wang BH. Prediction of severe sepsis using svm model. In Advances in computational biology. Springer, 2010; 75–81.

[7] Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Critical care 2005;9(2):R150.

[8] Ribas VJ, Vellido A, Ruiz-Rodríguez JC, Rello J. Severe sepsis mortality prediction with logistic regression over latent factors. Expert Systems with Applications 2012; 39(2):1937–1943.

[9] Fialho AS, Vieira SM, Kaymak U, Almeida RJ, Cismondi F, Reti SR, Finkelstein SN, Sousa JM. Mortality prediction of septic shock patients using probabilistic fuzzy systems. Applied Soft Computing 2016;42:194–203.

[10] Mitra A, Ashraf K. Sepsis prediction and vital signs ranking in intensive care unit patients. arXiv preprint arXiv181206686 2018;.

[11] Vellido A, Ribas V, Morales C, Sanmartín AR, Ruiz-Rodríguez JC. Machine learning for critical care: An overview and a sepsis case study. In International Conference on Bioinformatics and Biomedical Engineering. Springer, 2017; 15–30.

[12] Reyna M, Josef C, Jeter R, Shashikumar S, M BWM, Nemati S, Clifford G, Sharma A. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Critical Care Medicine 2019;.

[13] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 1989;77(2):257–286.

[14] Springer DB, Tarassenko L, Clifford GD. Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2015;63(4):822–832.

[15] Bouchard G, Triggs B. The tradeoff between generative and discriminative classifiers. In 16th IASC International Symposium on Computational Statistics (COMPSTAT'04). 2004; 721–728.

[16] Ayogu II, Adetunmbi AO, Ojokoh BA, Oluwadare SA. A comparative study of hidden markov model and conditional random fields on a yoruba part-of-speech tagging task. In 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE, 2017; 1–6.

[17] Sutton C, McCallum A, et al. An introduction to conditional random fields. Foundations and Trends in Machine Learning 2012;4(4):267–373.

Address for correspondence:

Shahrooz Faghihroohi
IMSAT, Shahid Beheshti University, Tehran, Iran
shahrooz.faghihroohi@gmail.com