Diagnosis of Sepsis Using Ratio Based Features

Shivnarayan Patidar
National Institute of Technology Goa, Ponda, India

Abstract

Early prediction of sepsis is of utmost importance to provide optimal care at an early stage. This work aims to use machine learning for early prediction of sepsis using ratio and power-based feature transformation. The feature transformation and feature selection process is optimized by applying a genetic algorithm (GA) based approach to extract the information specific to the sepsis from the given raw patient covariates that maximizes the underlying classification performance in terms of utility score. The proposed method begins with filling the missing values in the training dataset. Then, GA is applied strategically to identify influential ratio and power-based features from the raw patient covariates. The utility score is maximized as an objective of the optimization. RusBoost is used with default settings for underlying classification during optimization. Subsequently, an optimal RusBoost model is developed with a set of 55 identified features. Independent performance evaluation of the proposed method with the 2019 PhysioNet/CinC Challenge dataset has officially achieved 19^{th} rank with a utility score of 30.9% on the full hidden test data. This work appears as Shivpatidar on the leaderboard. The proposed early warning system has potential clinical value in critical care clinics.

1. Introduction

Sepsis is a potentially fatal condition that occurs when the host response to infections lead to tissue damage, organ failure, or even death [1]. Anyone suffering from an infection can develop sepsis, however elderly people, pregnant women, neonates, hospitalized patients, and people with HIV/AIDS, liver cirrhosis, cancer, kidney disease, autoimmune diseases and no spleen, are more prone to get sepsis. [2].

Monitoring the physiological status of a patient who can develop sepsis in future is of utmost importance in critical care clinics [3]. Nevertheless, such monitoring can assist in early diagnosis and then immediate treatments of sepsis in turn ensuring higher rates of survival and lower healthcare costs. Each hour of delay in the treatment for septic patients causes a 4-8% increase in mortality [4, 5].

Globally, around 30 million people get sepsis and as a result of which 6 million people lose their lives every year. Paediatric sepsis cases amounting to 4.2 million newborns and children poses one of the greatest challenges to paediatric critical care medicine.

In the U.S. itself, around 1.7 million people develop sepsis and nearly 270,000 patients die from it each year. Moreover, more than 33% of the people who die in the U.S. hospitals have sepsis. The economic burden of sepsis is surpassing any other illness with $24 billion which is 13% of U.S. healthcare expenses per year. It is noteworthy that a large part of this burden is due to the sepsis patients that were not diagnosed even after hospitalization [6]. The global economic burden of sepsis is even higher with the developing world at most risk. In a nutshell, severe sepsis and septic shock can cause significant morbidity, mortality, and healthcare expenses and therefore the need for the hour is to detect it as early as possible. Early prediction and immediate antibiotic treatment for sepsis are crucial needs for improving the conditions of sepsis patients.

Despite the intensive research for the management of severe sepsis and septic shock, there is a lack of screening tool capable of continuously monitoring its development [7]. Against this backdrop, this study proposes ratio and power-based feature transformation and RusBoost based classification. By applying a genetic algorithm (GA) based optimization, clinically significant features are optimally selected after said transformation of the given raw patient's covariates. Specifically, from the cross-validation data, a set of ratio and power-based features are tried and tested during GA based optimization to maximize the underlying classification performance. Such features are then used to design an optimal RusBoost classifier architecture for clinical decision making.

2. Methodology

As shown in Figure 1, the proposed methodology for early diagnosis of sepsis involves the following main subsections:

2.1. Pre-processing

In clinical settings, some tests cannot be carried out because either the hospital lacks the necessary medical device, or some medical tests may not be appropriate for cer-

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.207

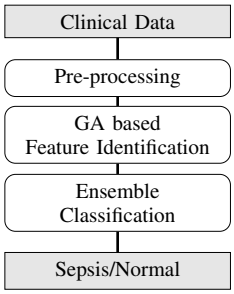

Figure 1. The proposed system.

tain subjects. As a result, the collected patient data may contain a large number of missing values. Missing or unknown values in the data can degrade the performance of pattern recognition techniques and therefore need to be dealt with when solving real-life classification problems. There are many ways to deal with missing values such as (a) case deletion, (b) imputation, (c) use of model-based procedures like expectation-maximization algorithm, (d) use of machine learning procedures, where missing values are incorporated into the classifier [8]. In this work, in order to replace the missing observations, linear interpolation of neighbouring two non-missing values is used to fill all the not a number of values for improving the overall performance of the algorithm.

2.2. Ratio and Power-based features

In literature, ratio-based features have found to improve the capabilities of anomaly detection systems [9]. In view of this, we have explored all possible ratio-based features up to order of three as given by the set:

$$R = \left\{ \frac{x^k}{y^m z^n} : x, y, z \in P; -9 \leq k, m, n \leq 9 \right\},$$

where, P is the vector containing 38 of the 40 given patient signs ignoring the two values of administrative identifier for ICU units. It is worth to note that the exploration space is not limited to the above-given space in general.

2.3. RusBoost based Classification

Ensemble classification [10, 11] strategically creates a set of weak classifiers and group them to get enhanced overall performance for machine learning applications. RusBoost is one of the popular methods to handle the class imbalance problem. It belongs to the class of boosting-based ensemble classifiers that deploys techniques for data preprocessing into boosting algorithms. In order to alleviate the class imbalance in data, RUSBoost removes instances of the majority class by random undersampling in each iteration. The undersampling technique makes RusBoost quite faster in terms of model building time as it uses fewer instances to construct each classifier in the ensemble. In fact, for every iteration during training, the method changes and bias the weight distribution toward the minority class for the next classifier or weak learner. After a model is trained, the weights of the original dataset instances are updated for the next sampling phase. The process of modifying the weight distribution repeats up to the number of weak learners considered ensuring more diversity in the training data which ultimately benefits the ensemble learning. The predictions of the individual weak learners are then merged using the majority vote to obtain the final decision.

2.4. GA based optimization

In 1962, Holland pioneered the development of GA as a stochastic optimization algorithm. The underlying principles of GA are an imitation of the process of natural selection and the evolution of genetic materials in living beings. In the following section, GA based feature transformation and selection approach are detailed to estimate the influential clinical features for detecting sepsis by constrained maximization of the normalized utility function as discussed earlier in Section 2.4. In order to maximize normalized utility function using GA, the search space for the model parameters needs to be specified properly beforehand. This step, in turn, can facilitate the faster convergence of the GA with an appropriate choice of the search space. Normally, the search space can be determined by analyzing the characteristics of the parameters. In this work, the entire feature search space is empirically chosen and it can be simply described as follows:

$$\boldsymbol{\theta} : \left\{ \theta : \theta^L \leq \theta \leq \theta^U \right\},$$

where, L and U denotes the related lower and upper ranges. Now, with GA the global or near-global estimate of the ratio and power-based features θ^* can be derived as follows:

$$\boldsymbol{\theta^*} = \arg\max_{\theta \in R} \{ U_n \}$$

where, U_n is the normalized utility score which measures the performance of the algorithm for its binary classification as given by the challenge organizers [7].

3. Results and discussion

In this study, we sought to design a screening tool for sepsis using machine learning methods and 2019 PhysioNet/CinC Challenge dataset [7]. The dataset used is comprised of three distinct US hospitals, arbitrary named A, B, and C. Part of dataset A and B forms the training

set, and C is completely kept hidden for testing purposes during the challenge.

The training set contains 1,552,210 records of 40,336 patients comprising vital signs, laboratory values, and demographics. Each observation has 40 attributes. The training set has two groups of records. Group A has 790,215 records of 20,336 patients. Group B has 761,995 records of 20,000 patients. The combined data has 1,552,210 records for 40,336 patients. Group A covers 1790 patients with 2.17% sepsis records amounting to 8.8% sepsis patients. The group B has 10,780 records of 1142 patients with 1.41% sepsis records amounting to 5.71% sepsis patients. The combined data has 27,916 records of 2932 patients with 1.80% sepsis records amounting to 7.27% sepsis patients.

Using GA in Matlab, an optimal sepsis detection framework is designed for classifying the records into non-sepsis and sepsis classes. As the given data has a high density (approximately 20%) of missing values, therefore the algorithm begins with linear interpolation of neighbouring two non-missing values to fill all the not a number values. 90% of the training set is used for cross-validation and remaining is used for testing purposes. In order to find the diagnostically useful features, GA based optimization with feature transformation and feature selection process from the given patient signs is applied to characterize sepsis. Optimization is done with an objective to maximize the underlying classification performance in terms of utility score. It is to be noted that only indices of the possible combinations of the ratio based features are passed to the objective function. And actual computation of GA based selected ratios and their selection is decided within the objective function. A supervised ensemble machine learning model called RusBoost is used with default settings for the underlying classification of the sepsis and non-sepsis records. Nevertheless, a 3-fold cross-validation scheme is implemented to get robust performance while doing optimization. Due to involved computational complexity, only the carefully chosen set of ratios as described earlier in Section 2.2 were explored. As a result, after GA based optimization, 17 most influential ratio and power-based clinical features are identified as mentioned in Table 1. However, it is to be noted that the optimal feature set may vary depending upon the GA settings. The obtained features are subsequently clubbed with given 38 signs to form the final feature set consisting of 55 values for classifying non-sepsis and sepsis records. An optimal ensemble system of RusBoost is designed as a classifier model for the early prediction of sepsis on test data. Table 2 shows the performance statistics of the proposed method in the context of clinical decision making at specified time thresholds in hours. This utility function is used in its original form as given by the challenge organizers for tabulating the statistics in Table 2. It rewards the parameter 1.0 to the classifiers for early predictions of sepsis if it predicts sepsis between intervals ranging from 12 hours before to 3 hours after the actual onset of sepsis (t_{sepsis}). The classifier is penalized if it does not predict sepsis or predict sepsis more than 12 hours before t_{sepsis}. The maximum penalties for very early detection and late detection are parameter 0.05 and -2.0 respectively. The classifier that predicts sepsis for non-sepsis cases is penalized with a parameter 0.05 which is the same as the very early detection penalty. The classifier is neither rewarded nor penalized if it does not predict sepsis. Moreover, the performance particular to sepsis cases is tabulated in Table 3.

The algorithm is finally evaluated officially with the 2019 PhysioNet/CinC Challenge dataset and obtained the utility score of 30.9% for the full hidden test data. Table 4 depicts the detailed performance of this work on hidden test sets. The reported execution time on hidden test set A in h:m:s is 15:57:22. This work appears in the ranking table as Shivpatidar and it has been ranked 19^{th}. The results obtained on a hidden set were consistent at least hospital-wise with that of the training. The proposed early warning system has potential clinical value in critical care clinics. The proposed method is very fast and in a real monitoring setting, this can increase the efficacy of the treatment of sepsis. This work reveals that the ratio and power-based features derived using patients signs are quite promising for better characterization of patient records to detect sepsis. The proposed method is robust to missing data even at a very lower density of 20% of the data. In order to analyze the effect of training data size on the performance of the proposed system, the model is trained and tested with varying training data and a fixed subset of test data respectively. And the resultant performance has been observed to be incremental with size of training data as depicted in Figure 2. It is noteworthy that the prognostic ability of the proposed method can be enhanced further by introducing more training instances while developing the involved model.

4. Conclusion

In this work, we have explored the strength of ratio and power-based features with RusBoost based classification for automated diagnosis of sepsis. The feature transformation and selection process are optimized to extract the information specific to the sepsis that maximizes the underlying classification performance in terms of utility score. The proposed framework for sepsis has shown significant performance on a large and diverse dataset. Further, the low-complexity execution makes it a suitable candidate for the detection and early prediction of sepsis in real-time clinical environments. The performance of the proposed method can be enhanced further by adding more train-

Table 1. List of identified influential features

S. No.	Component of Features	Type
1	End tidal CO_2/Partial thromboplastin time (s)	x/y^2
2	Diastolic BP/Gender	x^4/y
3	Diastolic BP/Gender	x/y
4	Heart Rate/ Age	x/y
5	Age/ Gender	x/y
6	Heart Rate/(Systolic BP*Age)	x/yz
7	Heart rate (beats per minute)	x^5
8	Temperature (deg C)	x^4
9	Mean arterial pressure (mm Hg)	$1/y^2$
10	Diastolic BP (mm Hg)	x^2
11	End tidal CO_2 (mm Hg)	x^4
12	FiO_2:Fraction of inspired oxygen (%)	x^8
13	Alkalinephos (IU/L)	x^6
14	Creatinine (mg/dL)	$1/y^3$
15	Fibrinogen (mg/dL)	x^4
16	Age	x^7
17	ICULOS:ICU length of stay	1/y

Table 2. Performance statistics at specified time thresholds using the training data.

Threshold Hours	AUROC	AUPRC	F1-Score	Accuracy	Utility Score
6	85.22	13.32	13.38	87.92	40.00
12	84.00	16.87	18.04	87.73	35.10
18	81.31	11.23	16.21	87.73	30.10
24	80.01	10.12	14.79	87.56	28.51

Table 3. Performance statistics of sepsis cases using the training data.

Threshold Hours	AUROC	AUPRC	F1-Score	Accuracy	Utility Score
6	56.13	22.31	28.15	45.01	60.10

Table 4. Performance statistics on the hidden test data.

Test Sets	Utility Score	F1-Score	Accuracy
A	39.0	13.2	82.9
B	38.6	13.2	89.0
C	-21.2	4.1	73.2
Overall	30.9	NA	NA

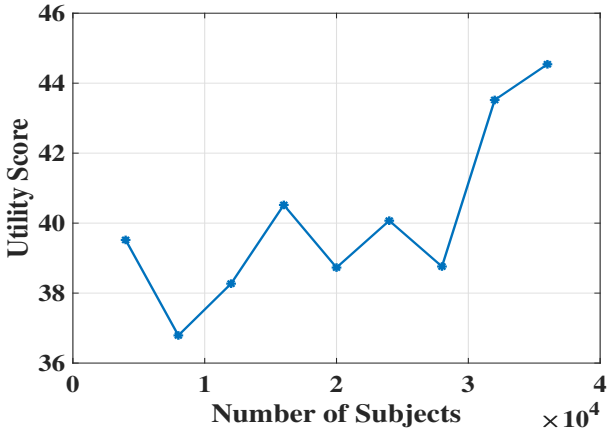

Figure 2. Effect of training data size on performance while training.

ing instances during model preparation. Exploration of the feature space beyond the considered limits including the derivative terms can be done to improve the predictive power of the said model for sepsis.

References

[1] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016;315(8):801–810.

[2] Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. British Medical Journal 2016;353:i1585.

[3] Bravi A, Green G, Longtin A, Seely AJ. Monitoring and identification of sepsis development through a composite measure of heart rate variability. PloS One 2012; 7(9):e45666.

[4] Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine 2017;376(23):2235–2244.

[5] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine 2006; 34(6):1589–1596.

[6] Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and costs of sepsis in the united statesan analysis based on timing of diagnosis and severity level. Critical Care Medicine 2018;46(12):1889.

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

[8] García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications 2010;19(2):263–282.

[9] Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Annals of Emergency Medicine 2019;73(4):334–344.

[10] Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 2006;6(3):21–45.

[11] Breiman L. Bagging predictors. Machine Learning 1996; 24(2):123–140.

Address for correspondence:

Dr Shivnarayan Patidar
Department of Electronics and Communication Engineering
National Institute of Technology Goa, Ponda
India, 403401
shivnarayan.patidar@nitgoa.ac.in