Prediction of Length of Stay on the Intensive Care Unit Based on Bayesian Neural Network

Jiansheng Fang\textsuperscript{1,2,3}, Junlin Zhu\textsuperscript{3} and Xiaoqing Zhang\textsuperscript{2}

\textsuperscript{1}School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
\textsuperscript{2}Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
\textsuperscript{3}CVTE Research, Guangzhou 510530, China

Email: 11949039@mail.sustech.edu.cn

Abstract. Predicting length of stay (LoS) accurately in the intensive care unit (ICU) is important to improve care quality and resource utilization. However, for LoS prediction, existing methods are facing main challenges, including uncertain prediction, generalization, interpretability, etc. In this paper, we utilize Bayesian neural network (BNN) to alleviate the above main challenges. The BNN introduces prior knowledge on the weights of neural networks and estimates outcomes from the predictive distribution after calculating the posterior distribution of weights. Extensive experiments on the eICU collaborative research database (eICU-CRD) show that the proposed method is competitive and more capable of anti-overfitting.

1. Introduction

Length of stay (LoS) in the intensive care unit (ICU) is a common indicator for evaluating nursing quality and improving resource utilization rate [1]. The shorter is the length of stay, and the lower is the cost [2]. It is necessary to provide a decision-making tool for predicting LoS of the patient in the ICU to maximize the resource utilization. Currently, the Acute Physiology and Chronic Health Evaluation (APACHE) scoring system [3] has been widely used in the ICU to measure the efficacy of ICU, and its versions IV and IVa are utilized in the prediction of LoS. For example, the Philips eICU collaborative research database (eICU-CRD) [4] provides actual LoS of patients and the predicted outcomes with APACHE IV, IVa.

But the predictive performance of APACHE scoring system is poor. Recently, a variety of machine learning methods are applied to improve the accuracy of the prediction of LoS in the ICU [5-7] and can be summarized into two points. First, various techniques of data preprocessing are adopted to discover the valid data distribution, such as Box-Cox transformation [8] and missing value imputation [9]. Second, machine learning models are selected and improved based on specific tasks, such as random forest models [7, 9], linear regression models with a penalty item [10].

However, the main challenges of LoS prediction are still unsolved, including uncertain prediction, generalization, interpretability, etc. In essence, due to the data uncertainty caused by the complexity or noise of data, the prediction of LoS is a task of uncertain prediction. But the current methods do not consider probabilistic distributions of data. Moreover, the current methods without incorporating prior knowledge about the data limit the ability of generalization and interpretability of models. To address...
the main challenges, we adopt Bayesian Neural Network (BNN) model to predict LoS on the eICU-CRD. The main contributions of this work are as follows.

- We build a neural network model to explore the high-level features and non-linear relationship of data and incorporate Bayesian into the neural network model as a probabilistic model to solve uncertain prediction.
- We introduce priors on the network weights and infer the posterior of network weights with Laplace’s approximation method to enhance the ability of generalization and interpretability.
- We perform extensive experiments on eICU-CRD to validate the effectiveness of the BNN model and demonstrate its ability to alleviate overfitting.

2. Bayesian Neural Network

As figure 1 shows, we define a probabilistic model for the fully connected neural network. In the probabilistic model, the prior assumption is introduced on the weights, and the likelihood function is built on inputs, weights, noise.

First, the prior distribution over the network weights is assumed to be Gaussian:

\[ p(\omega \mid \alpha) \sim \mathcal{N}(0, \alpha^{-1}) \]  

where \( \alpha \) is the precision of Gaussian distribution, \( \omega \) denotes weight vector of hidden layers. Further, we assume the actual outputs of hidden layers can be represented as the combination of predicted outputs and noise:

\[ t = y(x, \omega) + \epsilon \]  

where \( t \) is the actual outputs, \( y(x, \omega) \) is the predicted outputs of model. The \( \epsilon \) denotes the noise, and is assumed to be Gaussian:

\[ p(\epsilon \mid \beta) \sim \mathcal{N}(0, \beta^{-1}) \]  

where \( \beta \) is the precision of Gaussian distribution.

Based on the prior assumption of weights and noise, we can build the likelihood function:

\[ p(t \mid \omega, \beta) \sim \mathcal{N}(t \mid y(x, \omega), \beta^{-1}) \]  

where \( y(x, \omega) \) is the mean of Gaussian distribution. In the actual application scenario, the prior can be assumed to be other exponential distribution and assigned different precision for different parameters. Last, according to the Bayes rule and Laplace approximation method, we calculate the posterior distribution for the weights. The process is as follows:
The log posterior function is:

$$\ln p(\omega | t) \propto \ln p(t | \omega, \beta) + \ln p(\omega | \alpha)$$

(5)

The cost function is:

$$E(\omega) = -\ln p(\omega | t) = \frac{\beta}{2} \sum_{n=1}^{N} [t_n - y(x_n, \omega)]^2 + \frac{\alpha}{2} \omega^T \omega$$

(6)

where $N$ is the number of sample. Based on the cost function, the optimal posterior can be calculated after network forward and backpropagation.

Suppose $\omega_{map}$ denotes the optimal posterior and $S$ denotes the second derivative of the cost. Function $E(\omega)$ with respect to $\omega$. The posterior distribution is:

$$p(\omega | t) \sim \mathcal{N}(\omega_{map}, S^{-1})$$

(7)

Based on the posterior distribution, the predictive distribution is:

$$p(\hat{t} | x, \omega_{map}) = \mathcal{N}(\hat{t} | y(x, \omega_{map}), g^T S^{-1} g)$$

(8)

where $\hat{t}$ is the predicted value, $x$ is the new data, and $g$ is the first derivative of the model $y(x, \omega)$.

Different from the maximum likelihood (ML) method, BNN adopts maximum a posterior (MAP) method to estimate the weights with prior knowledge. More concretely, the outputs of hidden layers and output layer obey Gaussian distribution but not an optimal local value, this probabilistic output meets uncertain prediction. Naturally, the loss of BNN in the process of numerical optimization iteration contains two parts: Kullback-Leibler divergence between prior and posterior of weights, the difference between the predicted value of Gaussian density function and the actual value.

3. Experiments

3.1. Experimental Settings

Dataset. The eICU-CRD contains the data records of more than 200,000 ICU inpatients in the United States from 2014 to 2015 [4]. In eICU-CRD, the Acute Physiology Score (APS) table contain the inputs for model training, and the patient results table contains the predicted LoS with APACHE as well as actual LoS. The input data type contains the measured value, the binary flag, and the diagnostic text. To handle multiple types of data, the classification feature was coded by the one-hot encoding, the numerical features were transformed into a range between zero and one, and the text features were mapped into the index as a categorical feature. After data preprocessing and missing data imputation, the inputs with 52 features were split into 80% (108,988 records) for training, and 20% (27,248 records) for testing.

Metrics. Evaluation of metrics of explained variance (EV), mean absolute error (MAE), R-squared score ($R^2$) were adopted to assess the regression performance on LoS prediction.

(1) MAE (mean absolute error) is the measurement between the predicted value and observed value [11]. It is suitable for evaluating data distribution with skewness. It is nonnegative. The closer the value is to 0, the closer the predicted value is to the real value. The computation of MAE is as equation (9).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

(9)

where $n$ denotes the number of test data, $y_i$ is the $i$-th predicted value, and $\hat{y}_i$ is the $i$-th actual value.
(2) $R^2$ measures the prediction ability of regression model. It ranges from -1 to 1, where the higher the value, the higher the correlation [12]. The calculation method is present as equation (10).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ represents the average LoS of test data.

(3) EV is used to measure the differences between predicted and actual data. [13]. It has a similar interpretation with $R^2$ in regression, and is computed as equation (11).

$$EV = 1 - \frac{\text{var}(y - \bar{y})}{\text{var}(y)}$$

where $y$ is the actual value, $\bar{y}$ denotes the predicted value and $\text{var}(\cdot)$ represents variance.

Baselines. Except for APACHE IV and IVa, we also compared BNN with the following methods:

- Least square. Least square (LS) is a state-of-the-art linear regression model and has been widely applied in the LoS prediction [5, 10]. LS with Lasso regularization (LS+L1) is designed to get the sparse solution. And LS with Ridge regularization (LS+L2) is usually used to promote robustness.
- Random forest. Random forest (RF) is an ensemble learning method for classification and regression and has achieved better performance in LoS prediction [7, 9].
- Deep neural network. Deep neural network (DNN) is an artificial neural network with multiple layers between the input and output layers. DNN is able to capture composite relations and mine the high-level features but prone to overfitting. Deep learning has been widely applied in the medical domain and outperforms conventional methods [14-15].

3.2. Results

We implemented BNN based on Pytorch. We applied the technique of grid search and cross-validation to train LS and RF. We try our best to find the best value of dropout for DNN. For all models, the best performance given corresponding parameters on the test set of eICU-CRD was reported in table 1 on the same experimental device. As table 1 shows, the best performance of MAE was achieved by BNN, and the highest value of $R^2$ and EV were achieved by RF. On the whole, BNN is competitive with state-of-the-art models. Next, we mainly observe the effectiveness of anti-overfitting.

First, we compare the performance of overfitting on LS, RF, BNN by observing the discrepancy of MAE between train set and test set. As the right image of figure 2 shows, the MAE differences of three models all declines with the sample size growing, and BNN (red line) is better than LS (green line) when the sample ratio is greater 0.5. Obviously, the MAE difference of RF (gray line) is largely higher than other models. This is coincidence with the outcomes of table 1, higher EV and $R^2$ mean better fitting. The experiment demonstrates that the ability of generalization (anti-overfitting) of BNN is better than other models, lowest MAE, and better EV and $R^2$ (not the highest).

Second, we observe the performance of overfitting on DNN, BNN by setting different weight sizes of the network. The tendency of MAE performance on the test set makes clear that serious over-fitting occurs (when weight size = 32x8) in DNN (green line) with model complexity growing, as shown in the left image of figure 3. as the right image of figure 3 shows, the loss of BNN with different weight sizes on the test set all can converge to a stable value.

Upon the experimental results, we focus on discussions on the efforts of BNN in alleviating the main challenges of LoS prediction.
Table 1. Performance of LoS prediction on the eICU-CRD.

| Method   | MAE  | $R^2$ | EV    |
|----------|------|-------|-------|
| Apache IV| 2.454045 | 0.027745 | 0.065887 |
| Apache IVa| 2.250560 | 0.067626 | 0.076282 |
| LS+L1    | 2.011869 | 0.091121 | 0.091173 |
| LS+L2    | 2.010006 | 0.093095 | 0.093121 |
| RF       | 1.964103 | 0.110570 | 0.110646 |
| DNN      | 2.019279 | 0.092554 | 0.092554 |
| BNN      | 1.955044 | 0.097909 | 0.098971 |

Figure 2. Overfitting performance with sample size growing: Left. MAE on train set and test set; Right. Difference of MAE between train set and test set.

Figure 3. Overfitting performance with model complexity growing: Left. MAE of DNN and BNN on test set; Right. Loss convergence of BNN.

4. Discussions

First, the predictions of BNN obey Gaussian distribution, which is determined by the posterior weights (mean) and the precision of noise. Both mean and precision of Gaussian distribution are solved by training networks with numerical optimization. The predicted value sampled from a distribution satisfies the requirements of uncertain prediction of LoS, and more likely to approach the optimal global solution. The best performance of MAE reaches 1.86, which is greater than the average (1.95) of 200 samples. This value can be viewed as the most approximated to the optimal global solution on the eICU-CRD dataset.

Second, the weights with prior constraints amount to the regularization of LS for alleviating overfitting and improving the ability of generalization. The performance of anti-overfitting and generalization of BNN has been demonstrated the best in our experiment. By observing $R^2$ and EV, it can be seen that RF is the best fit for training data but lower MAE than BNN. Although we apply grid search and cross-validation to train RF, RF still occurs more serious overfitting than other methods. With the size of train data growing, the overfitting of BNN is weaker than LS with regularization. With the complexity of the model growing, BNN is much better than DNN with dropout.

Last, interpretability is the hot research point in both medical applications and neural networks. The prior knowledge introduced in the model is beneficial to enhance the interpretability of neural
networks. However, it is insufficient to show the effectiveness of prior knowledge in this work. In actual. The hyper-parameters of prior distribution should be different for each weight and solved in the network training. Then the contribution of each weight can be measured for the predicted value. Furthermore, the technique of visualizing data can be applied to observe the distribution of weights, such as T-Distribution Stochastic Neighbour Embedding (T-SNE) [16].

5. Conclusions
In this work, we proposed Bayesian Neural Network (BNN) to predict LoS on the eICU-CRD by incorporating the probabilistic model into the neural network to alleviate the main challenges of LoS prediction. The BNN introduces prior knowledge on the weights of neural networks and estimates outcomes from the predictive distribution after calculating the posterior distribution of weights. Extensive experiments demonstrate that BNN is qualified for uncertain prediction of LoS, and the generalization of BNN is better than the state-of-the-art methods by observing overfitting. In the future, we need to optimize BNN further to solve the hyper-parameter of prior distribution to enhance the interpretability of the model.

References
[1] Halpern N A and Pastores S M 2015 Critical care medicine beds, use, occupancy and costs in the united states: A methodological review Critical Care Medicine 4 2452.
[2] Kahn J M, Rubenfeld G D, Rohrbach J and Fuchs B D 2008 Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients Medical Care 1226-33
[3] Strand K and Flaatten H 2008 Severity scoring in the ICU: A review Acta Anaesthesiologica Scandinavica 5 2467-478.
[4] Pollard T J, Johnson A E, Raffa J D, Celi L A, Mark R G and Badawi O 2018 The EICU collaborative research database, A freely available multi-center database for critical care research Scientific Data 5.
[5] Verburg I W, de Keizer N F, de Jonge E and Peek N 2014 Comparison of regression methods for modeling intensive care length of stay PloS One 9 e109684.
[6] Verburg I W M, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, Peek N and de Keizer N F 2017 Which models can I use to pre- dict adult icu length of stay? A systematic review Critical Care Medicine 45 e222-e231.
[7] Daghistani T A, Elshawi R, Sakr S, Ahmed A M, Al-Thwayee A and Al-Mallah M H 2019 Predictors of in-hospital length of stay among cardiac patients: A machine learning approach International journal of cardiology 288 140-147.
[8] Langan R C, Huang C C, Mao W R, Harris K, Chapman W, Fehring C, Oza K, Jackson P G, Jha R, Haddad N, et al. 2016 Pancreateicoduodenectomy hospital resource utilization in octogenarians The American Journal of Surgery 211 70-75.
[9] Balkan B and Subbian V 2018 Evaluating icu clinical severity scoring systems and machine learning applications: Apache iv/iva case study 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp 4073-76.
[10] Li C, Chen L, Feng J, Wu D, Wang Z, Liu J and Xu W 2019 Prediction of length of stay on the intensive care unit based on least absolute shrinkage and selection operator IEEE Access 7 710-721.
[11] Chai T and Draxler R R 2014 Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding rmse in the literature Geoscientific Model Development 7 1247-50.
[12] Nagelkerke N J, et al. 1991 A note on a general definition of the coefficient of determination Biometrika 78 691-692.
[13] Achen C H 1990 What does “explained variance “explain?: Reply,” Political Analysis 2 173-184.

[14] Liu X, Chen Y, Bae J, Li H, Johnston J and Sanger T 2019 Predicting heart failure readmission from clinical notes using deep learning arXiv preprint arXiv:1912.10306.

[15] Fu H, Xu Y, Lin S, Wong D W K, Baskaran M, Mahesh M, Aung T and Liu J 2019 Angle-closure detection in anterior segment oct based on multilevel deep network IEEE Transactions on Cybernetics.

[16] Maaten L V D and Hinton G 2008 Visualizing data using T-SNE Journal of Machine Learning Research 9 2579-2605.