Diabetes Prediction with the TCN-Prophet model using UCI Machine Learning Repository

Tan Tianbo, Inwhee Joe
Hanyang Univ., Hanyang Univ.

UCI machine learning repository 사용한 TCN-Prophet 기반 당뇨병 예측

탄텐보, 조인휘
한양대학교,한양대학교
tianbo0928@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Abstract

Diabetes is a common chronic disease that threatens human life and health, and its prevalence remains high because its mechanisms are complex, further its etiology remains unclear. According to the International Diabetes Federation (IDF), there are 463 million cases of diabetes in adults worldwide, and the number is growing. This study aims to explore the potential influencing factors of diabetes by learning data from the UCI diabetes dataset, which is a multivariate time series dataset. In this paper we propose the TCN-prophet model for diabetes. The experimental results show that the prediction of insulin concentration by the TCN-prophet model provides a high degree of consistency, compared to the existing LSTM model.

Ⅰ. Introduction

Diabetes mellitus or diabetes is one of the incurable chronic diseases caused by lack or absence of a hormone called insulin. It is an essential hormone produced by the pancreas. The presence of high blood sugar levels in the blood is known as Hyperglycemia in medical terms. Therefore, by predicting the insulin level in a patient's body, it is possible to predict the risk of diabetes.

UCI Machine Learning Repository is a large, freely available database. 70 records of diabetic patients in the UCI database were used for the study. Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. For paper records, fixed times were assigned to breakfast, lunch, dinner, and bedtime.

In this study, we propose a hybrid approach that integrates the prophet and temporal convolutional networks (TCN) models to forecast insulin. In this experiment, to evaluate the predictive performance of our proposed model, TCN-Prophet model and the LSTM model are also applied for comparison. Experimental results show that the prophet model yields better results.

A. Prophet model

Prophet is a time series prediction algorithm based It can deal with some abnormal values and missing values and predict the future trend of time series almost automatically. The input is the known time series with time stamp and the length of the time series to be predicted. The output is the trend of future time series. Prophet decomposes the time series into trend items, periodic items, and residual items. For all $t \geqslant 0$, the decomposition of the sequence can be expressed as:

$$\mathrm{y}(t) = g(t) + s(t) + h(t) + \varepsilon(t) \qquad (1)$$

The expression of logistic regression function (nonlinear growth) is presented as:

$$g(t) = \frac{C(t)}{1 + \mathrm{e}^{-\left(k+\alpha(t)^{\mathrm{T}}\delta\right)\cdot\left(t-\left(m+\alpha(t)^{\mathrm{T}}\gamma\right)\right)}} \qquad (2)$$

The expression of piecewise linear function (linear growth) is presented as:

$$g(t) = \left(k + a(t)^{\mathrm{T}}\delta\right) t + \left(m + a(t)^{\mathrm{T}}\gamma\right) \qquad (3)$$

Ⅱ. Proposed Model

B. TCN model

Temporal convolutional networks (TCN) are generic convolutional networks proposed for sequence modeling tasks with causal constraint. A sequence modeling network is a nonlinear mapping functic $f\colon X^{T+1} \rightarrow Y^{T+1}$ which is obtained by supervised learning. When an input sequence the mean square error (MSE) is selected as the loss function

$$\mathrm{MSE} = \frac{1}{T}\sum_{u=1}^{T}(\hat{y}_u - y_u)^2 \qquad (4)$$

C. TCN-Prophet model

Routine insulin dose data are a series of complex discontinuous time series, which are affected by many factors. Firstly, the features of the series are extracted using TCN, and then the data are decomposed using Prophet model to predict the future series according to the additive model. Finally, the least squares method is used to fit the weight coefficients of the TCN and Prophet models so that the prediction results of the hybrid TCN-Prophet model have minimal differences from the real data. Define the output sequence of the TCN model as T (t) and the output sequence of the Prophet model as P (t), and the least squares method assigns the weights of both as $w1$,$w2$ the final prediction of the TCN-Prophet model is as follows.

$$\hat{Y}(t) = w_1 T(t) + w_2 P(t) \qquad (5)$$

$$w_1 + w_2 = 1. \qquad (6)$$

Ⅲ.Experimental Results

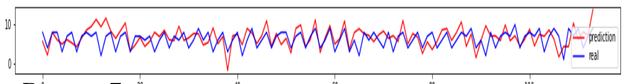
In this study, our dataset was derived from the UCI machine learning database. Among the 70 patients, data on the change in insulin concentration over time was extracted to make predictions. Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

Date	Time	Code	Value
03-01-1991	08:00	58	281
03-01-1991	08:00	33	007
03-01-1991	08:00	34	022

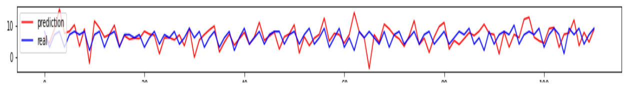
The Code field is deciphered as follows:

33	Regular insulin dose
35	Ultra Lente insulin dose
57	Unspecified blood glucose measurement
58	Pre-breakfast blood glucose measurement
59	Post-breakfast blood glucose measurement
60	Pre-lunch blood glucose measurement
61	Post-lunch blood glucose measurement
62	Pre-supper blood glucose measurement
63	Post-supper blood glucose measurement

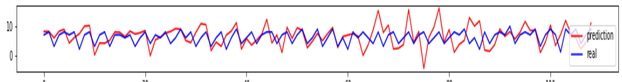
In the LSTM prediction model,70% of the data is used as training and 30% as testing. In the paper, a stochastic gradient descent (SGD) approach is taken every 100 epochs to train each LSTM unit. To counteract gradient explosion, the paper uses a scaled gradient parametric approach, uses a weight decay method, and selects two hyperparameters through the validation set. we debug the parameters of range to find the prediction data that best matches the true value when the range values are 7,14,21, and the results are as follows



Range=7

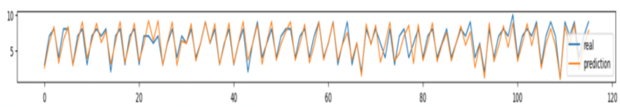


Range=14



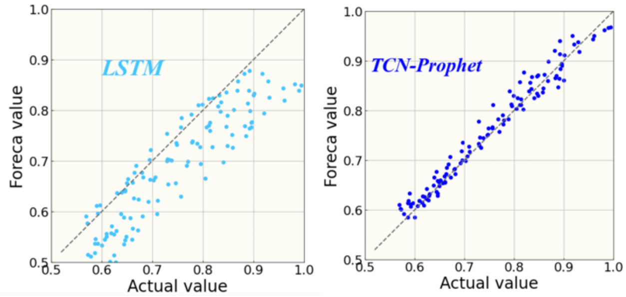
Range=21

In the TCN- Prophet model ,we performed the same preprocessing as the LSTM model. Since the clinical dataset of 70 patients is relatively large, the obtained results based on these samples are credible and persuasive due to the diversity and richness of data. Through observation, it is found that when the model acts on the sample with the overall insulin dose within the normal level, the model can still maintain accurate prediction effect even if there are insulin dose fluctuations.



TCN-prophet model

It can be seen from the two picture that the LSTM performs better in models based on the recurrent neural network prediction. the TCN-Prophet model proposed in this paper all perform the best on the insulin dose datasets. The experimental results show that the TCN-Prophet model has much higher forecasting accuracy than the classical models in time series forecasting.



Also, by comparing the two models, it can be observed that the TCN-Prophet model has better results than the LSTM model

III. Conclusion

In this paper, the Prophet method combined with TCN is used to process the time series data of insulin to solve the prediction problem. The UCI database from a total of 70 diabetes patients were used to train and test the model. By analyzing the overall experimental results, we conclude that the Prophet-TCN can achieve better prediction accuracy. However, the implementation of Prophet model has high requirements for timestamp data, dilated coefficient and the network structure is not concise enough, which makes the model training time longer. So, in the future we will optimize the model and collect more patient data to make better predictions about diabetes.

References

[1] R. Bellazzi, A. Abu-Hanna, "Data Mining Technologies for Blood Glucose and Diabetes Management," Journal of Diabetes Science and Technology, vol. 3, pp. 603-612, May 2009.

[2] L. Xie and Q. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models, " IEEE Transactions on Biomedical Engineering, vol. 67, no. 11, pp. 3101-3124, 2020.

[3] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, "Using LSTMS to learn physiological models of blood glucose behavior, " in 2017 39th Annual International Conference of the IEEE, pp. 2887-2891, 2017.