Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200406

Feature Set for a Prediction Model of Diabetic Kidney Disease Progression

Masaki ONO [a,1], Takayuki KATSUKI [a] Masaki MAKINO [b] Kyoichi HAIDA [c]
Atsushi SUZUKI [b] and Reitaro TOKUMASU [a]
[a] *IBM Research - Tokyo*
[b] *The Dai-ichi Life Insurance Company, Limited*
[c] *Fujita Health University*

Abstract. In this paper, we propose feature extraction method for prediction model for at the early stage of diabetic kidney disease (DKD) progression. DKD needs continuous treatment; however, a hospital visit interval of a patient at the early stage of DKD is normally from one month to three months, and this is not a short time period. Therefore it makes difficult to apply sophisticated approaches such as using convolutional neural networks because of the data limitation. The propose method uses with hierarchical clustering that can estimate a suitable interval for grouping inputted sequences. We evaluate the proposed method with a real-EMR dataset that consists of 30,810 patient records and conclude that the proposed method outperforms the baseline methods derived from related work.

Keywords. disease risk prediction model, diabetic kidney disease

1. Introduction

We constructed a risk prediction model for diabetic kidney disease (DKD) that receives the medical events of a patient in the past 180 days and predicts whether the patient will stay at the 1st stage after 180 days. A patient at the early stage of DKD needs continuous treatment; however, a hospital visit interval is from one month to three months, and this is not a short period. We focused on this issue, and the proposed method derives time series variation with hierarchical clustering that can estimate a suitable interval for grouping inputted sequences.

The number of studies on risk prediction models has increased and most of the studies applied logistic regression or cox regression to the latest medical events. Some studies focuses on patients that frequently visit a hospital, so records with many medical events in a certain period of time are available [1][2] and use CNN to distill a time series variation of a given patient.

The proposed method transforms a value sequence into groups on the basis of the result of the hierarchical clustering for a corresponding timestamp sequence. We use the hierarchical clustering method that has the constraint not on the number of the clusters but on the distances among the clusters. Then, we derive the time series variation from

[1] E-mail: moono@jp.ibm.com

Table 1. The AUC in the prediction for the stage of DKD of the patient after 180 days. "LR" stands for the logistic regression.

ID	Model	AUC	ID	Model	AUC
1	LR without time series variation	0.625	4	CNN #1 (interval 30 days)	0.632
2	LR with time series variation	0.722	5	CNN #2 (interval 14 days)	0.660
3	CNN #1 (interval 14 days)	0.666	6	CNN #2 (interval 30 days)	0.635

the group. For example, we calculate the averaged value and the standard deviation from the averaged cluster's values. The key factor is the hierarchical clustering that can estimate a suitable interval for grouping inputted sequences. We use logistic regression as a statistical model for the prediction because logistic regression can provide interpretable results. Interpretation is one of the important factors in the field of medical informatics.

2. Results and Discussion

We evaluate the proposed method with a real-EMR dataset from a Japanese hospital. We collected the medical history of a patient at the 1st stage. The number of patients is 30,810, where half of the patients remain at the 1st stage after 180 days. We focus on 12 kinds of laboratory tests as feature sets on, for example, uric protein and eGFR.

We conducted a 5-fold cross-validation test in the experiment and measured the AUC, which is the standard evaluation metric for a classification problem. And we used the baselines, the first of which is a logistic regression that handles the latest values of the feature sets and does not handle the time series variations that the proposed method creates. The other baselines are two kinds of CNNs that consist of 5 layers and handle feature matrices created in a previous work [1] [2]. We created two kinds of feature matrices created with different the interval for CNNs, 14 days and 30 days.

Table 1 shows the experimental results. The best performance is 0.722, which was achieved by the logistic regression with the time series variation that the proposed method creates. The experimental results point to the following conclusion: (1) understanding the time series variation improves the prediction performance of the risk prediction model of diabetic nephropathy and (2) the proposed method outperforms the baselines. The first conclusion is reached on the basis of the comparison between the ID 1 result and the others; the AUC of ID 1 is lowest and is derived from the logistic regression without the time series variations. The second conclusion is reached on the basis of the comparison between the ID 2 result and the others. The ineffectiveness of CNN for this task is caused by the missing value in the feature matrix. A low frequency of hospital visits causes the missing value in the feature matrix for CNN.

References

[1] Y. Cheng, F. Wang, P. Zhang and J. Hu, Risk Prediction with Electronic Health Records: A Deep Learning Approach, in: *SDM*, 2016.

[2] Z. Huang, W. Dong, H. Duan and J. Liu, A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records, *IEEE Transactions on Biomedical Engineering* **PP**(99) (2017), 1–1. doi:10.1109/TBME.2017.2731158.