Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-106

Risk Prediction of Diabetic Nephropathy via Interpretable Feature Extraction from EHR Using Convolutional Autoencoder

Takayuki KATSUKI[a,1], Masaki ONO[a], Akira KOSEKI[a], Michiharu KUDO[a], Kyoichi HAIDA[b], Jun KURODA[c], Masaki MAKINO[d], Ryosuke YANAGIYA[e], and Atsushi SUZUKI[d]
[a] *IBM Research – Tokyo, Japan*
[b] *Business Process Planning Department, The Dai-ichi Life Insurance Company, Limited, Japan*
[c] *IT Business Process Planning Department, The Dai-ichi Life Insurance Company, Limited, Japan*
[b] *Division of Endocrinology and Metabolism, Department of Internal Medicine, Fujita Health University, Japan*
[e] *Division of Medical Information Systems, Fujita Health University, Japan*

Abstract. This paper describes a technology for predicting the aggravation of diabetic nephropathy from electronic health record (EHR). For the prediction, we used features extracted from event sequence of lab tests in EHR with a stacked convolutional autoencoder which can extract both local and global temporal information. The extracted features can be interpreted as similarities to a small number of typical sequences of lab tests, that may help us to understand the disease courses and to provide detailed health guidance. In our experiments on real-world EHRs, we confirmed that our approach performed better than baseline methods and that the extracted features were promising for understanding the disease.

Keywords. Electronic Health Record, Risk Prediction, Diabetic Nephropathy, Kidney Disease, Convolutional Autoencoder, Feature Extraction

1. Introduction

Diabetic nephropathy (DN) is a kidney disease which is commonly complicated with diabetes mellitus [1]. For its risk prediction and detailed health guidance, growing attention is being paid to analyzing the electronic health records (EHRs) [2,3]. While most of the previous studies focused on past medical histories or lab tests at certain time points [4], using the long-term information available in EHR may be of help in risk prediction and lead to better understanding of the disease.

This paper proposes a machine learning-based approach to risk prediction for aggravation of DN from EHR. The major features of our system are twofold. First, the system can predict the risk incorporating the long-term

[1] Corresponding Author: Takayuki Katsuki, IBM Research – Tokyo, Japan; E-mail: kats@jp.ibm.com.

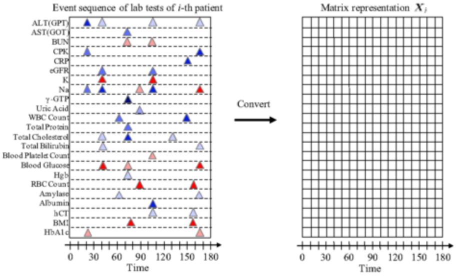

Figure 1. Event sequence of lab tests and its matrix representation. The horizontal and vertical axes respectively correspond to time stamps and different lab tests. Triangles indicate the values of the lab tests on the corresponding time; red and blue mean high and low values, respectively.

temporal information by using the features extracted from an event sequence of lab tests in EHR. Second, the extracted features are interpretable, which helps us to obtain knowledge about the characteristics and long-term course of DN.

To implement the above features, there were technical hurdles to overcome for each. In particular, our challenge for the first feature was how to handle the event sequence. As shown in Figure 1, since the events are recorded irregularly, when extracting features from that, we need to consider the time shift invariance and correlations between lab tests on both local and global time scales. This requires a hierarchical convolution and pooling mechanism across time in the feature extraction process.

The challenge for the second feature was how to make the extracted features interpretable while meeting the above requirement. It may be helpful if we can consider that the features are represented by affiliations to a small number of typical patterns of lab-tests sequence, i.e., some kind of filter for the sequence. This is because we can obtain these typical patterns through inverse analysis of the feature extractor and the extracted features can be interpreted as the similarities to these few patterns.

To meet these challenges, we propose a solution based on the convolutional autoencoder [5]. We will show that our approach performs better than baseline methods in experiments predicting aggravation on real-world EHRs. The extracted features reveal the typical sequences that are related to aggravation.

2. Methods

Our goal is to construct a prediction model for aggravation of DN from stage 1 to stage 2 after 180-days from the latest record in the input EHR, while keeping interpretability. We are given real-world EHRs obtained from 30,810 patients in a Japanese hospital. The aggravation label of the i-th input EHR is defined as $y_i \in \{0, 1\}$ such that $y_i = 0$ and $y_i = 1$ respectively represent that the patient stayed in stage 1 and that the one progressed to stage 2 after 180 days. The i-th input EHR is represented as a 180-day sequence of real-valued results of the lab tests and basic patient information, where we represent the sequence as a matrix $\boldsymbol{X}_i \in \mathbb{R}^{D \times T}$ whose horizontal dimension corresponds to the time stamp (time length $T = 18$ by taking the mean of each 10-day result in the 180 days) and whose vertical dimension corresponds to the lab tests having $D = 23$ attributes, as shown

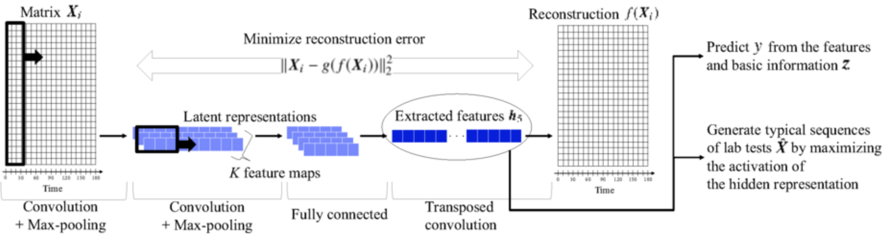
Matrix X_i
Reconstruction $f(X_i)$
Minimize reconstruction error
$\|X_i - g(f(X_i))\|_2^2$
Latent representations
Extracted features h_5
K feature maps
0 30 60 90 120 150 180
Time
0 30 60 90 120 150 180
Time
Convolution + Max-pooling
Convolution + Max-pooling
Fully connected
Transposed convolution
Predict y from the features and basic information z
Generate typical sequences of lab tests $\hat{X}$ by maximizing the activation of the hidden representation

Figure 2. Risk prediction framework via interpretable feature extraction from EHR.

in Figure 1, similar to [6]. The basic information is a vector, $\boldsymbol{z}_i$, consisting of age and sex, which is decoded as a combination of one hot vectors for age (every 10 years) and sex (man or woman). Using a total of N sets of the labels, matrices, and vectors, $\{y_i, \boldsymbol{X}_i, \boldsymbol{z}_i\}_{i=1}^{N}$, we learn the model for predicting y from $\{\boldsymbol{X}, \boldsymbol{z}\}$.

Figure 2 summarizes the concept of the proposed framework. For the matrices $\{\boldsymbol{X}_i\}_{i=1}^{N}$, we first learn a stacked convolutional autoencoder (SCAE) repeatedly applying one-dimensional convolution and pooling across time as a hierarchical feature extractor from the sequence. Then, we learn the prediction model for the aggravation y from the SCAE outputs with the basic information of patients $\boldsymbol{z}$. Finally, since the output of the SCAE becomes affiliations to a small number of typical patterns, we generate typical sequences of lab tests as those that maximally activate the SCAE outputs, which describe the features extracted with the SCAE.

First, for learning the SCAE, we minimize the following reconstruction error across the N lab-tests sequences $\{\boldsymbol{X}_i\}_{i=1}^{N}$:

$$\textstyle\sum_{i=1}^{N}\left\|\boldsymbol{X}_i - \boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{X}_i)\big)\right\|_2^2 \qquad (1)$$

where the function $\boldsymbol{f}(\cdot)$ is a feature extractor repeatedly mapping the input to the latent representations. We define the latent representation in the l-th layer as $\boldsymbol{h}_l$. Then, we use $\boldsymbol{f}(\boldsymbol{X}_i)$ to reconstruct the input $\boldsymbol{X}_i$ by performing a reverse mapping $\boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{X}_i)\big)$. As shown in Figure2, $\boldsymbol{f}(\cdot)$ has five hidden layers: 1) a convolutional layer with 64 $D \times 3$ filters; 2) a max-pooling layer of 1×2 filter; 3) a convolutional layer with 64 1×3 filters per map; 4) a max-pooling layer of 1×2 filter; 5) a fully connected layer of 128 hidden neurons. The k-th feature map in the l-th convolutional layer is a deterministic function, $\boldsymbol{h}_l^{(k)} \equiv \sigma\big(\boldsymbol{h}_{l-1} * \boldsymbol{W}_l^{(k)} + \boldsymbol{b}_l^{(k)}\big)$, where $\sigma(\cdot)$ is an activation function (we used a ReLU [7]) and the operator $*$ denotes a one-dimensional convolution across time with parameters $\boldsymbol{W}_l^{(k)}$ and $\boldsymbol{b}_l^{(k)}$. The max-pooling layers down-sample the latent representation by taking the maximum value over sub-temporal regions. The fully connected layer is also a deterministic function, $\boldsymbol{h}_5 \equiv \sigma(\boldsymbol{W}_5\boldsymbol{h}_4 + \boldsymbol{b}_5)$, where $\boldsymbol{W}_5$ and $\boldsymbol{b}_5$are parameters. The first convolutional layer receives its input from $\boldsymbol{X}_i$ as $\boldsymbol{h}_{l-1}$, and each of the other layers receives its input from the latent representation of the layer below. The reconstruction function $\boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{X}_i)\big)$ is the transposed convolution [8]. Through the unsupervised learning, the SCAE $\boldsymbol{f}(\cdot)$ works as a function that extracts features from an event sequence of lab tests without manually designing the function. The SCAE can capture local and global temporal information in early layers and later layers, respectively. As feature extraction results, we used 128-dimensional features $\boldsymbol{h}_5$ that were the output of the 5-th layer.

Next, we learn the prediction model for y from the joint feature vector $\boldsymbol{x} = \{\boldsymbol{h}_5, \boldsymbol{z}\}$ by solving an L1-regularized binary classification problem defined as

Table 1. Comparison of proposed and baseline methods in terms of mean AUC (higher is better).

Basic	Basic + Latest	Basic + Stats	SCAE	Proposed (Basic + SCAE)
0.60 ± 0.01	0.63 ± 0.01	0.62 ± 0.01	0.64 ± 0.01	0.66 ± 0.01

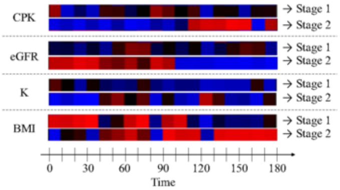

Figure 3. Comparison of sequences between patients who stayed in stage 1 and those who progressed to stage 2. Red and blue squares mean high test value and low-test value, respectively.

$$\min_{\boldsymbol{w}^{(\mathrm{pred})},b^{(\mathrm{pred})}} \sum_{i=1}^{N} L\left(y_i, \delta\left(\left[\boldsymbol{w}^{(\mathrm{pred})}\right]^{\mathrm{T}} \boldsymbol{x} + b^{(\mathrm{pred})}\right)\right) + \lambda \sum_{m=1}^{M} \left|w_m^{(\mathrm{pred})}\right| , \quad (2)$$

where $\boldsymbol{w}^{(\mathrm{pred})}$ and $b^{(\mathrm{pred})}$ are parameters, and the functions $L(\cdot)$ and $\delta(\cdot)$ are the cross entropy and sigmoid function, respectively. Using the learned parameters, $\hat{\boldsymbol{w}}^{(\mathrm{pred})}$ and $\hat{b}^{(\mathrm{pred})}$, we can predict the probability of the label for the new data as

$$P(y = 1) = \delta\left(\left[\hat{\boldsymbol{w}}^{(\mathrm{pred})}\right]^{\mathrm{T}} \boldsymbol{x} + \hat{b}^{(\mathrm{pred})}\right). \quad (3)$$

Finally, we output a typical sequence $\hat{\boldsymbol{X}}$ for the s-th feature, $[\boldsymbol{h}_5]_s$, through inverse analysis solving the following optimization problem:

$$\max_{\boldsymbol{X}} \, [\boldsymbol{h}_5]_s, \quad (4)$$

where we extract the pattern by maximizing the activation of $[\boldsymbol{h}_5]_s$. We can interpret the feature as the similarity to the pattern.

We used ADAM with the recommended hyper-parameters in [9] and mini-batches of 128 examples for the above optimization problems.

3. Results

Table 1 compares the results of the proposed method with those of the baseline methods that use the same prediction model as ours (Eq. (3)) but input different features from ours, i.e., only basic information (basic), basic information and the latest values of lab tests (basic + latest), basic information and basic statistics of the 180-day lab-tests sequence consisting of mean, standard deviation, and $\{0.05, 0.25, 0.5, 0.75, 0.95\}$ quantiles (basic + stats), and only SCAE outputs (SCAE). We evaluated the results with regard to the mean of the Area Under the Curve (AUC) in ten-fold cross validation using the given real-world EHRs and also computed the standard deviation of the AUC. Here, we set the regularization parameter candidates to $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ for all of the above methods and selected the optimal one by using five-fold cross validation in the training data of each cross-validation step. From Table 1, we can see that the mean AUC of the proposed method is better than those of the baselines. This shows that the SCAE can extract good features from the EHRs for the prediction.

Figure 3 is a comparison between a typical 180-day lab-tests sequence $\hat{\boldsymbol{X}}$ in Eq. (4) of patients who stayed in stage 1 and those who progressed to stage 2, and it shows features having notable differences between these groups. We determined $[\boldsymbol{h}_5]_s$ for the groups respectively as the latent representations having the lowest and highest values of the weights $\boldsymbol{w}^{(\mathrm{pred})}$. In the typical sequences of patients who progressed to stage 2; the values of the creatine phosphokinase (CPK) and body mass index (BMI) are increasing over time, the glomerular filtration rate (eGFR) value has opposite characteristic from them, and the potassium (K) value fluctuates through time. These results are mostly consistent with knowledge about DN. Additionally, if only the latest lab test values were analyzed, it would miss such temporal behaviors of the lab tests that we discovered. We can show typical sequence for each latent representation in each middle layer. Here, we omitted them due to space limitations.

4. Conclusion

We demonstrated that the proposed method can predict aggravation of diabetic nephropathy with higher accuracy than baseline methods by using features extracted from EHR by the SCAE. We could obtain the typical patterns for interpreting the extracted features. The next step is to combine our prediction model with other information, such as medication. Thanks to the unsupervised nature, the extracted features by the SCAE can be used flexibly for modeling with any features extracted by other methods. Considering what new information can be obtained from the extracted typical patterns is another interesting topic.

Acknowledgments

T. Katsuki was supported in part by JST CREST Grant Number JPMJCR1304, Japan.

References

[1] International Diabetes Federation, "IDF diabetes atlas eighth edition," http://www.diabetesatlas.org/resources/2017-atlas.html, 2017.

[2] M. Shimizu, K. Furuichi, T. Toyama, S. Kitajima, A. Hara, K. Kitagawa, Y. Iwata, N. Sakai, T. Takamura, M. Yoshimura *et al.*, "Long-term outcomes of japanese type 2 diabetic patients with biopsy-proven diabetic nephropathy," Diabetes Care, vol. 36, no. 11, pp. 3655–3662, 2013.

[3] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," J Am Med Inform Assoc, vol. 22, no. 4, pp. 872–880, 2015.

[4] J. B. Echouffo-Tcheugui and A. P. Kengne, "Risk models to predict chronic kidney disease and its progression: A systematic review," PLOS Medicine, vol. 9, no. 11, pp. 1–18, 11 2012.

[5] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," Artificial Neural Networks and Machine Learning, pp. 52–59, 2011.

[6] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 453–461.

[7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning, 2010, pp. 807–814.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[9] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proceedings of the 3rd International Conference for Learning Representations (ICLR2015), 2015.