Diagnosis Prediction via Recurrent Neural Networks

Yangzi Mu, Mengxing Huang, Chunyang Ye, and Qingzhou Wu

***Abstract*—The prediction of patient's future health information from the historical electronic health records (EHR) forms the core of the development of personalized healthcare research tasks. Patient EHR data consists of sequences of visits over time, where each visit contains multiple medical codes, including diagnosis, medication, and patient profile. Using historical data from the EHR, we can predict medical conditions and medication uses. Existing works model EHR data by using recurrent neural networks (RNNs). However, RNN-based approaches have certain limitations: the performance of RNNs drops when the length of sequences is large and they ignore some of the characteristics of the patients themselves. We propose an application of using bidirectional RNNs to remember all the information of both the past and future visits and add some patient's characteristics as side information into this model. Experimental results on real world EHR datasets show that the proposed model can remarkably improve the prediction accuracy when compared with the diagnosis prediction approaches, and it can provide clinically meaningful interpretation.**

***Index Terms*—Component, electronic health records, bidirectional recurrent neural networks, side information**

I. Introduction

The common challenge in smart health is how to use the large amount of data in predicting visiting patients' diseases in a short period of time. Due to complicated processes, different symptoms, and pathological tests, making the correct diagnosis is a difficult task and causes delays in providing the proper treatment. Electronic health records (EHR) consisting of patient health data, including demographics, diagnoses, procedures, and medications, have been utilized successfully in several predictive modeling tasks in healthcare [1]-[3]. EHR data are temporally sequenced by patient medical visits that are represented by a set of high dimensional clinical variables (i.e., medical codes). While forecasting medical models have been developed to predict the expected demand, most of the existing works have focused on specialized forecasting models or a single target. In order to model the sequential EHR data, recurrent neural networks (RNNs) are used in the literature to obtain accurate and robust representations of patient visits in diagnostic predictive tasks [3], [5].

However, the predictive power of these models drops significantly when the length of the patient visit sequences is large. Further, these models usually ignore some of the characteristics of the patients themselves and others. While not so extreme, there are many diseases associated with gender, family history, region, season, and so on. Bidirectional recurrent neural networks (BRNNs) [6], which can be trained using all the available input information in the past and future, have been used to alleviate the problem of long sequences, thereby improving the predictive performance. Referring to the method of collaborative filtering (CF), we use the side information to reasonably interpret the importance of patients and medical codes in the prediction results. This side information can be obtained from the user profile and other information. Some side information has proven to be useful for heart disease decisions [7], [8]. Some hybrid CF methods have gained popularity in recent years [9], [10], where side information is integrated into matrix factorization to learn the effective latent factors.

We demonstrate that the proposed model achieves significantly higher prediction accuracy when compared to the other approaches in diagnosis prediction, using our datasets from Haikou People's Hospital. In summary, our main contributions are as follows:

- We propose a new, end-to-end, simple, and powerful model that can accurately predict future visits, without relying on any expert's medical knowledge.
- It models the patient's visit information in time- and reverse-time-ordered ways and employs side information as supplementary information.
- We show that the proposed new model outperforms existing methods in diagnosis prediction with regard to EHR datasets.

The rest of this paper is organized as follows: In Section II, we discuss the connection between the proposed approaches and related works. Section III details the proposed new model. The experimental results are given in Section IV. Section V concludes this paper.

II. Related Works

This part reviews the existing work for mining EHR data. In particular, it focuses on several state-of-the-art models on diagnosis prediction tasks. It also includes some works that use side information in CF-based methods.

A. EHR Data Mining

Mining EHR data is a popular topic in medical informatics. The investigated tasks include electronic genotyping and phenotyping [11], [12], disease progression [13], [14], diagnosis prediction [1], [2], [15], and so on. In most of these tasks, the machine learning model and depth neural network models can significantly improve the performance.

Diagnosis prediction is an important and difficult task in medical informatics. Machine learning can remarkably improve performance, such as using the SVM algorithm in

Manuscript received January 8, 2018; revised March 19, 2018.

The authors are with College of information Science &Technology, Hainan University, Haikou, China and State Key Laboratory of marine resource utilization in the South China Sea, Hainan University (e-mail: muyangzi521@163.com, huangmx09.com, cyye@ustc.com, wuqingzhou@21cn.com).

doi: 10.18178/ijmlc.2018.8.2.673

the diagnosis of heart disease [1] and combining SVM and identification set algorithm in the diagnosis of Alzheimer's dementia [2]. The application of a decision tree to multi-label classification can fully take into account the relevance of each classification label; each label can effectively tap the associated information. However, the decision tree for the continuity of the field is more difficult to predict. Further, it needs to perform a lot of data preprocessing; therefore, for a higher number of categories, the error rate increases rapidly. The most important decision tree is prone to the over-fitting phenomenon.

Lipton [3] employed LSTM RNNs for the multi-label classification of diagnoses for variable-length time series of clinical measurements. However, that approach does not associate both pre- and post-diagnostic information and the static information of patients. Amin [7] proposed a new hybrid model of neural networks and genetic algorithms to optimize the connection weights of artificial neural networks so as to improve their performance. They used some risk factors to predict the heart disease. Although it can prove that some of the risk factors can affect the occurrence of heart disease, it did not take into account the time sequence and disease history. Med2Vec [16] and Deep Patient [17] aimed to learn the representations of medical codes, which can be used to predict future visit information. This method ignores long-term dependencies of medical codes and some static factors.

B. CF

The most successful method in the CF-based approach is to learn effective latent factors directly from the user-item rating matrix factorization technique [18]. However, the rating matrix is often very sparse in the real world, causing CF-based methods to degrade significantly.

In order to overcome this problem, the CF-based method utilizes additional sources of information about users or items, also called side information. Therefore, this hybrid CF method has become popular in recent years [9], [10], [19], [20]. These methods are very effective in improving performance.

RESCAL [19] is a tensor factorization approach to relational learning. It is able to perform collective learning via the latent components of factorization.

Wang [10] proposed an algorithm for recommending scientific articles to users based on both content and other users' ratings. This approach combines the merits of traditional CF and probabilistic topic modeling, thereby providing an interpretable latent structure for users and items.

Recently, one of the powerful methods to learn effective representations is deep learning [25]. Thus, with large-scale ratings and rich additional side information, it is essential to integrate deep learning in recommender systems to learn latent factors. Therefore, some studies have directly used deep learning for CF tasks. Wang [4] utilizes SDAE or marginalized SDAE for CF, but this method requires the learning of a large number of manually adjusted hyper parameters.

In this paper, we address these challenges by designing a LSTM RNN, which can successfully learn complex sequential patterns.

III. Model

For clarity, we illustrate the idea assuming a patient state RNN. Assuming there are M patients and N diseases, we denote a binary vector $u_t \in \mathbb{R}^n$ as the diagnosis vector for a given patent at time *t*. Each diagnosis code can be mapped to a node of the International Classification of Diseases (ICD-9). For example, there are three diagnosis codes: 411.1 (unstable angina), 427.31 (unspecified atrial fibrillation), and 428 (heart failure, unspecified) in the whole dataset. If the patient has unspecified atrial fibrillation and heart failure at time t, then $u_t = [0,1,1]$, where $u_{tj} = 1$ if the patient has the disease j at time step t; $u_{tj} = 0$ otherwise. In addition, the data at time step *t* is denoted by c_t and μ =1 is used to express that the patient is new.

$$x_t := \mathrm{E}_{Embed}\,[u_t, \mu, c_t, c_{t-1}] \tag{1}$$

E_{Embed} is the transformation to be learned from the disease information into the embedding space. Further, x_t serves as the input to the BiLSTM at step *t*.

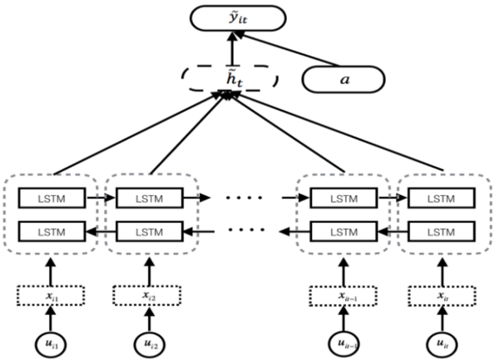


Fig. 1. The structure of the proposed model.

The LSTM model is the most common choice for processing timing sequences [22]. It was proven to have great performance with regard to timing sequences. The overall model is shown in Fig. 1. The state updates satisfy the following operations:

$$[f_t, i_t, o_t] = \sigma[W[h_{t-1}, x_t] + b] \tag{2}$$

$$\tilde{c}_t = \tanh[V[h_{t-1}, x_t] + d] \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{4}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{5}$$

where f_t, i_t, and o_t denote the forget, input, and output gates, respectively. They control the information flow through the sequence. Further, *W* and *V* are the weight matrices. For simplicity, we use

$$h_{it} := \mathrm{LSTM}(h_{it-1}, x_{it}). \tag{6}$$

We denote the additional *i* as the i$^{\text{th}}$ patient to distinguish between different patients. Then, we use the attention mechanism to capture relevant information to help predict the future visit y_t. In these tasks, the multiplicative attention mechanism [21], which more simply and quickly captures the relationship between h_t and h_j, is used to address this problem.

$$a_{it} = softmax\left(\sum h_t^T W_a h_j\right) \tag{7}$$

$$z_{it} = \sum_{j=1} a_{itj}\, h_j \tag{8}$$

$$\tilde{h}_{it} := \tanh(zi_t, h_{it}) \tag{9}$$

where a_t denotes the attention weight vector and z_t denotes the context vector. The context vector z_t is combined with the current hidden state h_t to generate the attentional hidden state $\tilde{h}_t$. Here, $\tilde{y}_t$ denotes the predicted disease for a given patient at time t. Even though the patient state may change over time, we speculate that there is still some fixed side information a_i that can help reasoning, such as the patient's gender, residence, etc. We use

$$\tilde{y}_{it} = \mathrm{Softmax}\left(W_y \tilde{h}_{it} + W_a a_{it} + b_y\right) \tag{10}$$

The optimization objective is to find predicted parameters that are close to the actual data. We employed cross entropy as a loss function:

$$\arg\min \frac{1}{M} \sum_{i=1}^{M} \frac{1}{T} \sum_{t=1}^{T-1} \{(y_{it}\log(\hat{y}_{it}) + (1-y_{it})\log(1-\hat{y}_{it}))\} + \lambda H(\theta) \tag{11}$$

where θ denotes all the parameters to be learned. λ is a regularization parameter and $H(\theta)$ denotes the regularization function.

IV. Experiments

A. Data Description

These experiments use a set of completely anonymous clinical events from the Hainan People's Hospital. The dataset contains the diagnosis ID, time, diagnostic code, order execution time, stop time, drugs code, patient profile, etc. We extracted the diagnostic code. Due to the complexity of data, data loss and other reasons, only 108 kinds of diseases were available after screening. Further, the code used is similar to ICD-9. Overall, the data involved 5391 patients and 21736 visits after screening for patients with less than 2 records. The patient time was grouped by weeks. The side information contains the patient's age, gender, and location. A total of 10 categories are encoded into a binary-valued vector. We randomly divide the dataset into the training set (70%), validation set (10%), and testing set (20%). The validation set is used to determine the best values of the parameters.

B. Baselines and Evaluation

We compared our proposed model with the following methods:

- **Item-KNN.** This is the standard item-based recommendation method. It uses the cosine similarity between the vectors of the patients as the items similarity, defined as the number of co-occurrences of two items in sessions divided by the square root of the product of the numbers of sessions in which the individual items occur.
- **SVM.** Traditional multi-classification support vector machine.
- **RNN.** The basic LSTM model is similar to Lipton's model [3].
- **$\mathbf{RNN_y}$.** This model is our proposed model.
- **$\mathbf{RNN_i}$.** This model uses side information as the input. The architecture is similar to Cristobal Esteban's model [23].

To evaluate the performance of predicting the next diagnosis for each approach, we use two measures: accuracy and accuracy@k. Accuracy is defined as

$$\frac{\text{the number of correct predicted medical codes}}{\text{total number of medical codes in (t+1)time}} * 100\% \tag{12}$$

Accuracy@k is defined as

$$\frac{\text{the number of correct predicted medical codes in top k}}{\text{total number of medical codes in (t+1)time}} * 100\% \tag{13}$$

After several experiments, the best parameters are further adjusted by experience and individually optimized for each parameter. We choose Adam [24] as the optimization method. In terms of deep network architecture, the number of layers is set to 2 in our experiments with 128 hidden units in each layer. The dropout rate is set to 0.4 and the regularization parameter λ, 0.01. Other parameters are uniformly initialized between [-0.03, 0.03].

C. Result

TABLE I: Recall@20 and MRR@20 for Different Best Performing Models in Our Experiments

Algorithms	accuracy	accuracy@5	accuracy@10
Item-KNN	10.61%	17.43%	21.19%
SVM	18.48%	25.81%	31.84%
RNN	38.53%	44.17%	48.10%
RNN_y	**47.16%**	**55.11%**	**64.75%**
RNN_i	43.26%	48.32%	52.87%

For each method, we perform 100 iterations and report the accuracy on the testing set for the model that gives the best results on the validation set. Table I shows the accuracy of the baselines method and the proposed model for the diagnosis prediction task.

In Table I, we show the performance achieved for each model. It can be seen that the accuracy of the proposed approach is higher than that of other baselines approaches. When comparing RNN and RNN_y, it is evident that side information can significantly improve a model's performance. In RNN_i, the side information is put into the first input unit, although the model could learn this side information. We speculate that even though it is connected to the input units, this data is different from the subsequent input data, which affects the judgment of the model.

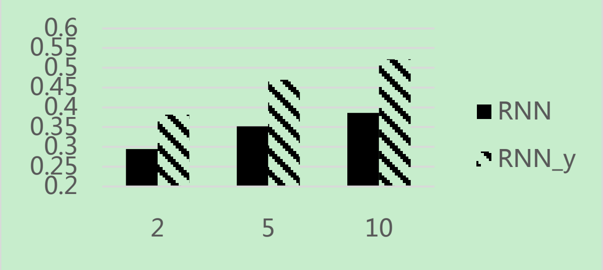


Fig. 2. Comparison of two models in different sequence lengths.

We compared the effects of different sequence lengths on the performance (Fig. 2). To study the cold-start problem, we set the sequence length to small values. When the sequence length is 2, our method is slightly better than RNN_i, but the improvement is not obvious. However, when the sequence length is greater than 2, the proposed model has a significant performance improvement as compared to $\mathrm{RNN_i}$ when the sequence length increases. It shows that the combination of biLSTM and attention mechanism can improve the characteristics' discovery of the patients. Therefore, our approach can better predict the diagnosis.

V. Conclusion

We proposed a novel model to effectively learn patient representations from a number of longitudinal patient records and combine the patients' static information to predict the patient's future visits. We demonstrate that this model is significantly better than other approaches.

We believe that the application of a deep neural network in diagnosis prediction is a novel way. It can not only imitate the ability of human doctors to predict but also provide the appropriate diagnostic results, which is very meaningful in clinical medicine. Although our experimental accuracy is not high, it can be seen that RNN utilized side information in a much better manner than other classification methods. Further, because our experimental data is not enough, our future works will involve collection of more data to verify and improve the effectiveness of our algorithm. At the same time, we will collect data from different hospitals to confirm our model's applicability. Furthermore, we can expand our experiments to involve further medical research, such as the analysis of medical images.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 61462022 and the International Cooperation Project of Hainan under Grant No. KJHZ2014-16.

References

[1] C. Marcin, "Ischemic heart disease detection using selected machine learning methods," *International Journal of Computer Mathematics*, vol. 90, no. 8, 2013, pp. 1734-1759.

[2] J. Ram ́ez *et al.*, "Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences*, vol. 237, no. 10, pp. 59-72, 2013.

[3] Z. C. Lipton, D. C. Kale, and R. C. Wetzel, "Phenotyping of clinical time series with LSTM recurrent neural networks," *Computer Science*, 2015.

[4] W. Hao, N. Wang, and D. Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1235-1244.

[5] E. Choi, *et al.*, "Doctor AI: Predicting clinical events via recurrent neural networks," *Computer Science*, 2015.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 2002.

[7] A. S. Umar, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," *Information & Communication Technologies*, pp. 1227-1231, 2013.

[8] A. Rani and K. Usha, *Analysis of Heart Diseases Dataset Using Neural Network Approach*, vol. 1, no. 5, 2011.

[9] S. A. Pratap, G. Kumar, and R. Gupta, "Relational learning via collective matrix factorization," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 650-658.

[10] W. Chong and D. M. Blei. "Collaborative topic modeling for recommending scientific articles," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, Ca, Usa, DBLP, August 2011, pp. 448-456.

[11] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, p. 395, 2012.

[12] C. R. Liu *et al.*, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 705-714.

[13] X. Wang, D. Sontag, and F. Wang, *Unsupervised Learning of Disease Progression Models*, pp. 85-94, 2014.

[14] J. Y. Zhou *et al.*, "A multi-task learning formulation for predicting disease progression," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, Ca, Usa, 2011, pp. 814-822.

[15] J. Y. Zhou *et al.*, "Patient risk prediction model via top-k stability selection," in *Proc. the 2013 SIAM International Conference on Data Mining*, 2012.

[16] C. Edward *et al.*, *Multi-layer Representation Learning for Medical Concepts*, 2016, pp. 1495-1504.

[17] R. Miotto *et al.*, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Scientific Reports, vol. 6, 2016.

[18] K. Yehuda, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30-37, 2009.

[19] N. Maximilian, V. Tresp, and H. P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. International Conference on International Conference on Machine Learning*, Omnipress, 2011, pp. 809-816.

[20] S. Yue, M. Larson, and A. Hanjalic, *Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges*, ACM, 2014.

[21] L. M. Thang, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Computer Science*, 2015.

[22] H. K. Moritz *et al.*, *Teaching Machines to Read and Comprehend*, 2015, pp. 1693-1701.

[23] E. Cristobal *et al.*, *Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks*, 2016, pp. 93-101.

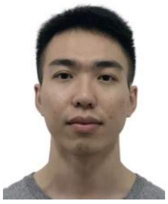
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, issue 5786, pp. 504-507.

Yangzi Mu was born in 1993 in Henan, China. He received the bachelor's degree in College of information Science & Technology, Zhengzhou University in 2015. He is currently a master student in Hainan University. His research interests include machine learning, data mining.

Mengxing Huang is currently a professor at College of information Science & Technology, Hainan University, Haikou, China. Prof. Huang received the Doctor degree in Northwestern Polytechnical University in 2007. His research interests include knowledge engineering, big data and cloud computing.

Chunyang Ye is currently a Researcher at College of information Science & Technology, Hainan University, Haikou, China. Dr. Ye received his bachelor degree and master degree in the Department of Computer Science, University of Science and Technology of China and then received the Doctor degree in Hong Kong University of Science and Technology in 2008. His research interests are software engineering issues on service-oriented computing.



Qingzhou Wu was born in Guangzhou, China He received the bachelor's degree in China University of Geosciences in 2016. He is currently a master student in Hainan University. His research interests include machine learning, data mining.