Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(5): 5347–5363.
DOI: 10.3934/mbe.2021271
Received: 22 March 2021
Accepted: 07 June 2021
Published: 16 June 2021

Research article

SeekDoc: Seeking eligible doctors from electronic health record

Lu Jiang, Shasha Xie, Yuqi Wang, Xin Xu, Xiaosa Zhao, Ye Zhang, Jianan Wang*and Lihong Hu*

Northeast Normal University, Changchun 130117, China

* **Correspondence:** Email: lhhu@nenu.edu.cn; wangjn@nenu.edu.cn.

Abstract: With the development of online medical service platform, patients can find more medical information resources and obtain better medical treatment. However, it is difficult for patients to discover the most suitable doctors from the complex information resources. Therefore, the analysis and mining of Electronic Health Record(EHR) is very important for patients' timely and accurate treatment. Discovering the most suitable doctor is actually predicting the exact performance of the doctor for a specific disease. We believe that "a curative/bad treatment is likely to be caused by a good/bad doctor, and a good/bad doctor has a higher/lower evaluation by the patient(s)". In this paper, we propose a novel approach named *SeekDoc*, which is to seek the most effective doctor for a specific disease. Specifically, we build a doctor-disease heterogeneous information network and collect patients reviews and rating records for doctors. Then, we embed the comprehensive comment data for doctors and the constructed heterogeneous information network. Next, we use the autoencoder mechanism to learn the embedded features, which is an effective learning algorithm for constructing the latent feature representation in an unsupervised manner. After this learning, the latent features are input into the extreme gradient boosting (XGBoost) algorithm to improve their detection capabilities. Finally, extensive experiments show that our method can effectively and efficiently predict the doctor's experience score for specific diseases and has good performance compared with other algorithms.

Keywords: healthcare; electronic health record; heterogeneous information network; autoencoder; extreme gradient boosting(XGBoost)

1. Introduction

The number of users is growing every year on the Internet Medical Service Platform. For example, Guahao.COM, which is supported by the National Health and Family Planning Commission of China (NHFPC). As of July 2014, the Guahao website has been connected to the information systems of more than 900 key hospitals in 23 provinces across the country. It has more than 30 million real-name

registered users and more than 100 thousand key hospital experts. Guahao website is committed to connecting hospitals, doctors and patients via the Internet, and promoting the efficient sharing of information between them. The platform presents the network structure [1] naturally and is relatively clear, including data of multiple dimensions such as the number of people, hospitals, and satisfaction. Using new developments and effectively analyzing data, we can help patients understand the competence of each doctor and find a more suitable doctor for themselves to better meet their individual needs. It also improves the quality of the medical platform. But, It is difficult to find the most effective doctors for patients with specific diseases from these complex data.

In this research, we use the patient's review information to find the most eligible doctor for specific diseases. We believe that doctors with many positive comments from patients have better performance and vice versa. We build medical heterogeneous information network where we predict the real performance of the doctors' specific-disease, which is called the experience score. The higher the score, the better the doctor's performance. The medical heterogeneous information network [3] is shown in Figure 1, where the nodes with different types represent doctors and diseases, and the edges represent the treatment of doctors on diseases.

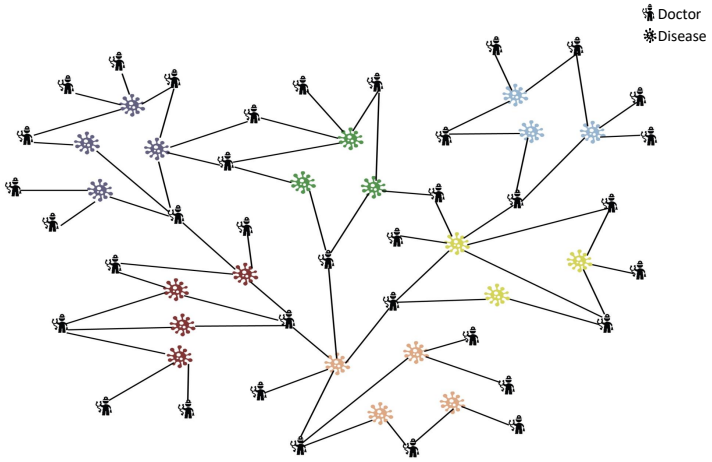


Figure 1. Doctor-disease heterogeneous information network.

Firstly, to obtain better performance of the doctors, we embed two aspects of features. (i) Embedding information: a patient publishes comments on the doctor's personal home page, displaying criticism, appreciation or suggestions based on his/her medical experience. (ii) Embedding network: extracted the doctor-patient-disease heterogeneous network [4] from millions of review records, each of which represents a patient visiting a doctor for a specific disease.

Secondly, in order to get more subtle and abstract features of doctors and diseases, we use the autoencoder [5] to learn the embedded representation of the network. Autoencoder is an important training model in deep learning [2, 6] that automatically learns data representations by attempting to reconstruct their inputs at the output layer. At the same time, when we embed textual features in the network, we use autoencoder processing technology to learn the general representation of textual information and extract the hidden features to avoid the high dimensional and sparse textual content.

Finally, considering that the doctor may be good at several specific diseases and the patient evaluate a limited number of doctors, we adopt the Extreme gradient boosting (XGBoost) [7, 8] for the sparse

heterogeneous network. The XGBoost algorithm has many advantages, which are needed in this paper, including the ability to build relatively quickly, to process continuous and categorical data naturally, to handle missing data naturally, to be robust to the outliers in the input, and to scale well on a large dataset.

In brief, we use a large number of patients' evaluation data for doctors to propose a new method named *SeekDoc*, which can predict the experience score of doctors for a specific disease, so as to find the most effective doctors for the disease. We summarize our contributions as follows:

- We build a doctor-disease heterogeneous network to explore the potential links between the disease and the doctor, and then collect reviews and rating records from patients and utilize network-embedding techniques to represent doctor-disease vector pairs.
- We use autoencoder to get the latent features. Then, XGBoost algorithm is adopted to predict the doctor's experience scores for special diseases.
- Experiment results with real-world large-scale datasets demonstrate the effectiveness of *SeekDoc*.

The rest of the paper is organized as follows. Preliminaries is presented in section 2, followed by section 3 that describe the proposed method *SeekDoc* in detail. In section 4, we describe the experiments and evaluations. Section 5 describes the related work. We conclude the paper and discuss the future work in section 6.

2. Preliminaries

We aim to predict the best doctor for a specific disease. Before presenting the proposed algorithm, we briefly show the required notations and definitions.

Definition 1. ***(Heterogeneous Medical Information Network, HMIN)*** *In the process of medical diagnosis, the network of participants connected by various relationships is heterogeneous medical information network. Define the node type as T, including the main medical participants. A represents the set of nodes, $A = \{A_t\}_{t=1}^{T}$, where A_t represents the set of nodes of type t. An undirected network is usually represented as a graph $G(V, E, W)$, when the set of node $V = A$, the set of edge E is a binary relation group on V, W is a weight mapping from edge E. When $T \geq 2$, information network is called heterogeneous information network.*

Definition 2. ***(Doctor-Disease Network, 2DN)*** *A $2DN$ is defined as a graph $G(V_{doctor}, V_{disease}, E, W)$, where V_{doctor} and $V_{disease}$ are two sets of nodes that represents doctors and diseases, E is the set of links, which connects a doctor and a disease, and W is the corresponding set of weights.*

Based on the two definitions, we further use the comments textual to quantify the weight between two adjacent nodes. Selecting this measure mainly has two reasons. First, comments textual provides an intuitive way to characterized the node relationship. Generally, the better doctor performs on a disease, the better comments the patient will evaluate the doctor. Secondly, the method of text-to-vector is very mature, and it is reasonable to convert numerical expression from text.

Definition 3. ***(Doctor-Disease Network Weight, 2DNW)*** *Given a pair of doctor and disease(i.e., the i^{th} doctor in V_{doctor} and the j^{th} disease in $V_{disease}$), SeekDoc collect the weight of textual, represented by w^i and w^j respectively, from textual comments.*

SeekDoc first aggregates all comments of the i^{th} doctor from the online content data, then it picks keywords, denoted as $word_1^i$, $word_2^i$, ..., $word_p^i$(suppose totally p keywords found), from the aggregated comments. Using Word2Vec [9] technique, these p keywords are then converted to vectors v_1^i,v_2^i, ..., v_p^i respectively. The weight w_1^i characterizing the comments that patients left to the i^{th} doctor is extracted as the center of keyword vectors, i.e..

$$w_1^i = \frac{1}{p}\sum_{k=1}^{p} v_k^i \tag{2.1}$$

In the similar way, *SeekDoc* aggregates the comments containing the name of disease, then extracts keywords from the comments. These keywords are converted to vector v_1^j, v_2^j,...,v_p^j through Word2Vec. Then the center of these keyword vectors is used to characterize the j^{th} disease. We denote this weight as w_1^j.

3. Methods

In this section, we describe the framework of *SeekDoc*. Afterward, we introduce two parts of the framework in detail.

3.1. SeekDoc model

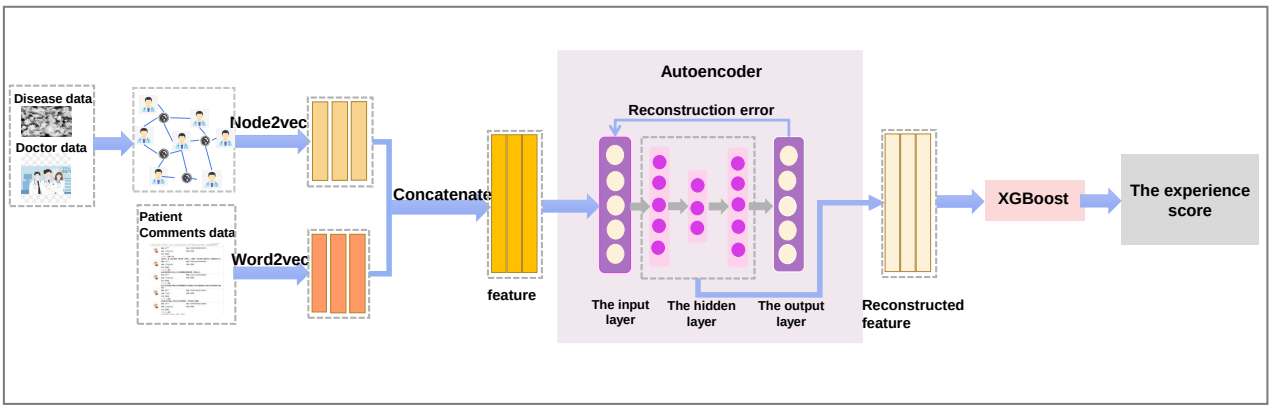


Figure 2. The framework of *SeekDoc*.

The framework of *SeekDoc* is shown in Figure 2. First, we build a hereogeneous information network with doctor and disease data and use Node2Vec * [10] technology to obtain the representation of the doctor and disease. At the same time, we use Word2Vec † [9] to transfer the comment data into the feature representation. Then we concatenate these two feature representation and apply autoencoder technology to learn them to get a reconstructed embedded feature. At last, we use Extreme Gradient Boosting to learn the reconstructed feature to improve the accuracy of the prediction and compute the experience score of doctors for a specific disease.

*https://github.com/eliorc/node2vec/
†https://github.com/3Top/word2vec-api/

3.2. Autoencoder

We design an autoencoder structure to learn embedded features. The basic framework of autoencoder is a neural network, which includes input layer, output layer and at least one hidden layer. An single hidden layer autoencoder can be described by Figure 3.

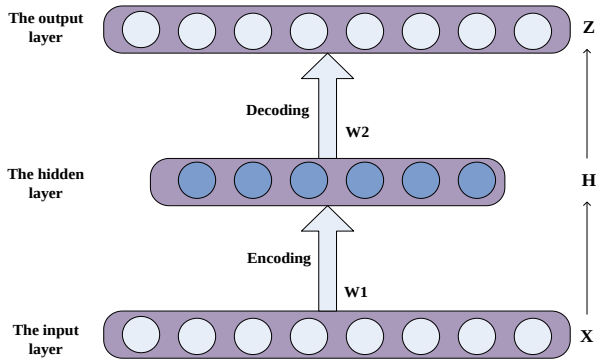


Figure 3. Single hidden layer autoencoder.

In Figure 3, for input X, assuming $X = (x_1, x_2, \cdots, x_d)$, and $x_i \in [0, 1]$, the autoencoder first maps the input X to an implicit layer, using the hidden layer to represent it as $H = (h_1, h_2 \cdots, h_{d'})$, and $h_i \in [0, 1]$, the process is called encode. The specific form of the output H of the hidden layer is

$$H = \sigma_1(W_1 X + b_1) \tag{3.1}$$

where W_1 represents the weight matrix between the input layer and the hidden layer. b_1 represents bias vector. σ_1 is a nonlinear mapping, usually an activation function. At this level, we use the Rectified Linear units (ReLU) [11]. The ReLU activation function is defined as follows:

$$\sigma_1(x) = max(0, x) \tag{3.2}$$

The output H is called the variable of the hidden layer, and the variable Z is reconstructed by the variable of the hidden layer. Here, the Z of the output layer has the same structure as the input layer X, and this process is called decode. The specific form of the output layer Z is

$$Z = \sigma_2(W_2 H + b_2) \tag{3.3}$$

where σ_2 is the Tanh activation function, which is defined as

$$\sigma_2(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.4}$$

The output of the output layer can be thought of as a prediction of the raw data X using the feature H. For the weight matrix W_2 in the decoding process, it can be regarded as the inverse process of the encoding process, that is, $W_2 = W_1^T$. b_2 represents bias vector.

In order to minimize the reconstruction error between the reconstructed Z and the original X, we define its loss function:

$$l = \| X - Z \|^2 \tag{3.5}$$

3.3. XGBoost

Extreme Gradient Boosting (XGBoost) is an integrated learning algorithm based on Gradient Boosting. Its principle is to achieve accurate classification by iterative calculation of weak classifier. The Gradient Boosting Machine algorithm uses the idea of gradient descent in generating each tree, based on all the trees generated in the previous step, moving in the direction of minimizing the given objective function. XGBoost is an implementation of the Gradient Boosting Machine that automatically leverages multithreading of the CPU for parallelism and improves the algorithm.

Then, we introduce the model construction of XGBoost.

Given a data set $D = (x_i, y_i)(|D| = n, x_i \in R^m, y_i \in R)$, the tree's integration model is represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3.6}$$

where $F = \{f(x) = w_{q(x)}\}(q : R^m \to T, w \in R^T)$ is the set space of the regression tree, x_i represents the feature vector of the i_{th} data point, k represents the number of trees, q represents the index of the result of each tree mapped to the leaf corresponding to the sample, and T represents the number of leaves, and each f_k corresponds to a separate tree structure q and the weight of the leaf w.

The original objective function form is as follows:

$$Obj(\Theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.7}$$

the first part is the training error between the predicted value $\hat{y}_i$ and the true value y_i , and the second part is the sum of the complexity of each tree. Complexity calculation formula:

$$\Omega(f) = \gamma \cdot T + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2 \tag{3.8}$$

Then we use the method of adding a new function on the basis of retaining the original model every time. The detailed process is as follows:

$$\hat{y}_i^{(0)} = 0 \tag{3.9}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{3.10}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{3.11}$$

$$\cdots$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3.12}$$

$\hat{y}_i^{(t)}$ is the model prediction value of the i^{th} sample in the t^{th} iteration, and it retains the model prediction value $\hat{y}_i^{(t-1)}$ of the iteration t-1, and a new function $f_t(x_i)$ is added. The choice to add a new function in each iteration is to minimize the objective function as much as possible. Thus, the rewrite target function is:

$$Obj^{(t)} = L^{(t)} = \sum_i^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) = \sum_i^n l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_t) + contanst \tag{3.13}$$

optimize this objective function with f_t. When the error function l is a square error, the objective function can be written as:

$$L^{(t)} = \sum_i^n [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + contanst \tag{3.14}$$

For the case of other error functions, Taylor expansion is used to approximate the objective function, details are as follows:

$$\begin{aligned} g_i &= \frac{\partial l(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}} \\ h_i &= \frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}} \\ \widetilde{L}^{(t)} &\cong \sum_i^n [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + contanst \end{aligned} \tag{3.15}$$

After removing the constant term, a relatively uniform objective function is obtained:

$$\widetilde{L}^{(t)} \cong \sum_i^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{3.16}$$

4. Experimental results

In this section, we first introduce statistical information about the real medical data set in this paper. Then, we compare *SeekDoc* with several baselines on this data set. The experimental results indicate that *SeekDoc* has better performance in many evaluation indicators.

4.1. Data analysis

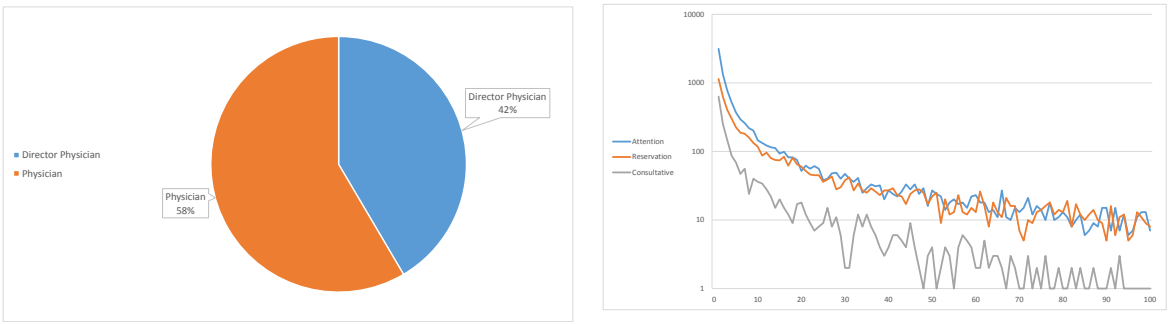
We use a real and comprehensive data set to perform the experiment with doctors and patients. The dataset come from an online clinical discussion platform serving 554 hospitals in China—Guahao.COM (https://www.guahao.com/) where the patient can make a rating for the doctor by making an appointment online or consulting a doctor. It contains all patient appointments, reviews and ratings covers three years from July 2012 to December 2015.

Table 1 gives some detailed statistics about the datasets.The dataset includes 28625 doctors which have 14284 ratings and comments from patients and 358 categories of disease. The patient's rating range for the doctor is [−2,3], with −2 representing the patient being very dissatisfied with the doctor and 3 representing the patient being satisfied with the doctor. Based on these data, the algorithm will always give an excellent recommend.

Table 1. The detail of doctor-disease network dataset.

Data	Attributes	Numbers
Doctors	Number of Doctors(NoD)	28625
	NoD with comments	2719
	NoD with appointments	1999
	NoD with consultations	8463
Patients	Number of Comments	14284
	Number of effective comments	10935
	Grade of Efficacy	[−2,3]
Disease	Number of Diesase	358

In addition to the introduction of the data in the above table, we also consider the properties of the data itself, as mentioned in the following figure. Figure 4(a) shows the distribution of the doctor whether he is the director. Then, we summarize the doctor's profile information including number of reservation, number of consultation and attention as shown in Figure 4(b). The x-axis indicate is the number of doctors with the number of concerns, appointments, and consultations related to the disease. The y-axis corresponds to the amount of attention, appointment and consultation.



(a) Distribution of the director doctor

(b) Comparison of doctors' profile information

Figure 4. Properties of the data.

Next, in order to display rating information more intuitively. We extracted the information that Figure 5 refers to the patient's rating of the doctor's treatment effect, it is easy to see from the figure that more patients are satisfied with the results of the doctor's treatment, and only a small number of patients think they have not received effective treatment.

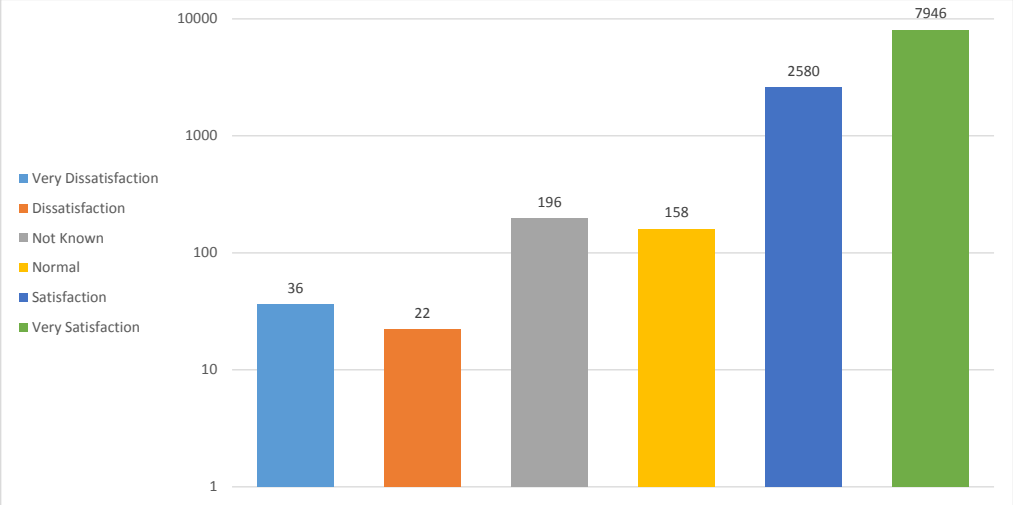


Figure 5. Statistics on the evaluation of doctors' treatment effect.

4.2. Baseline algorithms

In this section, we compare the proposed method with eleven baseline algorithms for predicting doctor-disease pairs score. Note that, the first five algorithms are used to compare the quality of the recommendation results. The sixth algorithms are used to compare the effects of different feature combination. They are as follows:

Multi-Layer Perception (MLP): the MLP [12] algorithm is one of the deep learning method. Learning through the Neural Network, include input layer, hidden layers, output layer to learning more mission-oriented features through hierarchical structure.

Kernel Ridge (KR): the KR [13] algorithm combine Ridge Regression(Linear Regression) with kernel techniques. It will learn linear functions in space caused by individual kernels and data.

Gaussian Naive Bayes (GNB): the GNB [14]algorithm inheriting Naive Bayes, the feature possibility is assumed to be Gaussian.

Nonnegative Matrix Factorization (NMF): the NMF [15] algorithm makes all components after decomposition non-negative, and at the same time achieves a non-linear dimension reduction.

Dr.Right! (DR!): the DR! [16] algorithm develop a data analytical framework which incorporates the so-called network-textual embedding, together with data-imbalance-aware mixture multi-classification models to rate doctors per specific diseases.

Textual Features (*SeekDoc*-T): only textual features are embedded as inputs in this model.

Heterogeneous Network Features (*SeekDoc*-HN): only heterogeneous network features are embedded as inputs in this model.

Textual Features + Heterogeneous Network Features (*SeekDoc*-THN): textual features and heterogeneous network features are embedded as inputs in this model.

Textual Features + Autoencoder (*SeekDoc*-TE): adding autoencoder to textual features as inputs in this model.

Heterogeneous Network Features + Autoencoder (*SeekDoc*-HNE): adding autoencoder to heterogeneous network features as inputs in this model.

Textual Features + Heterogeneous Network Features + Autoencoder (*SeekDoc*-THNE):

adding autoencoder to textual features and heterogeneous network features are embedded as inputs in this model.

4.3. Evaluation metrics

Our experiments evaluate the proposed method from different perspectives, including its error, stability and comprehensiveness.

More concretely, we used MSE and RMSE metrics to calculate SDE for evaluating the stability of the proposed predicting framework. The description of the different metrics is as follows:

4.3.1. Mean square error

$$MSE = \frac{\sum_i^N \left|S_i - \hat{S}_i\right|^2}{N} \tag{4.1}$$

4.3.2. Root mean squared error

$$RMSE = \sqrt{\frac{\sum_i^N \left|S_i - \hat{S}_i\right|^2}{N}} \tag{4.2}$$

4.3.3. Standard deviation of error

$$SDE = \sqrt{\frac{\sum_i^N (Error_i - \bar{Error_i})^2}{N}} \tag{4.3}$$

where N represents the size of the data set, S_i represents the doctor-disease score observed in the data set, that is, the patient's score on the doctor's treatment effect, and $\hat{S}_i$ indicates the experimentally predicted doctor-disease score. $\left|S_i - \hat{S}_i\right|$ refers to the absolute error of the prediction, expressed as $Error_i$, and $\bar{Error_i}$ is used to represent the average error. The MSE is the average of the sum of squares of errors between the observed doctor-disease score and the predicting score. The smaller the value is, the lower error the prediction is. The RMSE is the square value of MSE, the smaller the value, the more accurate the prediction is. The SDE is the standard deviation between them. The smaller the SDE is, the more stable the performance of the model is.

In order to provide guidance for patients to find the right doctor, we use the softMax [17] function to classify their doctors, and use ACC to evaluate the quality of our classification.

4.3.4. Accuracy

$$ACC = \frac{\sum_i^N (S_i == \hat{S}_i)}{N} \tag{4.4}$$

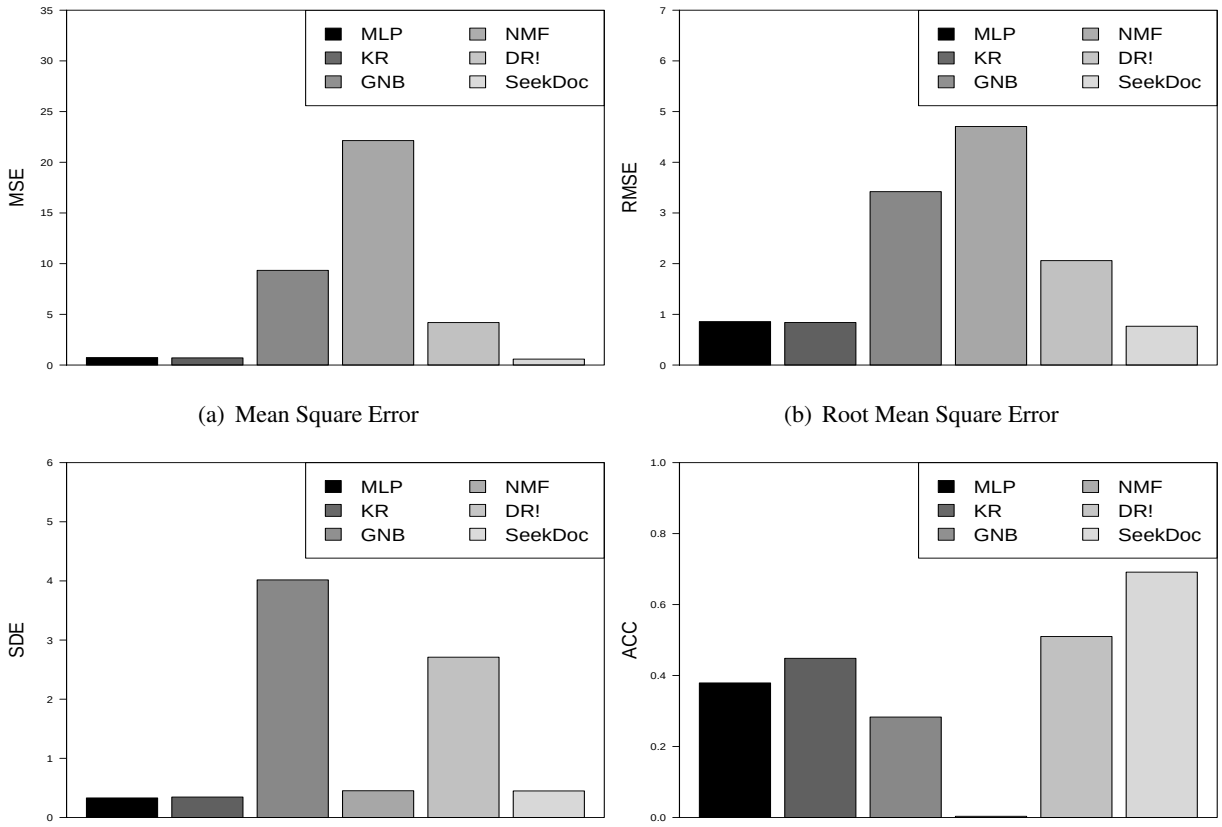
The ACC refers to the probability that the predictive score is the same as the observed doctor-disease score. The higher the value is, the higher the accuracy of the prediction is.

4.4. Experimental environment

In this work, the machine learning models include three parts. First, data preprocessing is trained on python 2.7.15 with several scientific computing libraries, such as Numpy 1.15.1, Xlrd 1.1.0, Networkx 2.1, Matplotlib 2.2.3 and Scipy 1.1.0. Next, in order to abstract high dimensional feature model from data, we use autoencoder method trained on python 3.6.8 with several scientific computing libraries, such as Sklearn 0.20.2, Numpy 1.15.4 and Pytorch 1.0.0. Finally, main program and baseline algorithms are run in python 2.7.15 with several scientific computing libraries, such as Numpy 1.15.1, Sklearn 0.19.2 and XGBoost 0.6.

Parametric Details: We set *learning_rate* = 0.1, *n_estimators* = 1000, *max_depth* = 30, *min_child_weight* = 1, *gamma* = 0.2, *subsample* = 0.8, *objective* = 'multi:softmax', *scale_pos_weight* = 1 for XGBoost.

4.5. Overall performance



(a) Mean Square Error

(b) Root Mean Square Error

(c) Standard Deviation of Error

(d) Accuracy

Figure 6. The comparison performance between five baseline algorithms.

In this section, to evaluate the performance of the proposed algorithm. The experimental results are shown in Figure 6 to indicate the state-of-the-art performance of *SeekDoc*. We used the MSE, RMSE, SDE and ACC to evaluate the proposed method. The average results of the proposed method and the other five baselines were obtained after 5-fold cross validation. From the experimental results, we can

draw the following conclusions:

(1) *SeekDoc* have the lowest MSE value(0.5857) than the MSE values of MLP(0.7330), KR(0.7043), GNB(9.3391), NMF(22.1384), DR!(4.1895), shown in Figure 6(a).

(2) *SeekDoc* have the lowest RMSE value(0.7653) than the RMSE values of MLP(0.8562), KR(0.8392), GNB(3.4204), NMF(4.7051), DR!(2.0598), shown in Figure 6(b).

(3) *SeekDoc* have lower SDE value(0.4487) than the SDE values of GNB(4.0169), NMF(0.4518), DR!(2.7102), but little higher than MLP(0.3299), KR(0.3454), shown in Figure 6(c).

(4) *SeekDoc* have the best ACC (0.6914) values are higher than the ACC values of MLP(0.3791), KR(0.4485), GNB(0.2830), NMF(0.0033), DR!(0.5099), shown in Figure 6(d).

SeekDoc has the best performence on MSE and RMSE, this represents the smallest error between our predicted Doctor's performance and reality. Although SDE value is not better than MLP and KR algorithm, the experimental results still show that our algorithm has better stability. We also classify doctors by using the software Max function and get the highest ACC value. It means that the prediction result is closest to the real situation. In a word, all experimental results demonstrate the priority of *SeekDoc*, compared with other baseline algorithms on the perspective of comprehensive evaluation.

4.6. Robustness check

In this section, in order to better understand the algorithms, we will introduce the robustness check following these two parts: (1) study of feature embedding; (2) study of timing consuming.

4.6.1. Study of feature embedding

In this section, we demonstrate the performance of embedding features on the MSE values. The result is shown in Table 2.

Table 2. MSE values for feature embedding.

MSE	*SeekDoc*-T	*SeekDoc*-HN	*SeekDoc*-THN	*SeekDoc*-TE	*SeekDoc*-HNE	*SeekDoc*-THNE
MLP	5.5678	12.7355	5.5378	0.7171	0.8140	0.7330
KR	**2.3868**	3.2435	3.1785	**0.6942**	**0.7726**	0.7043
GNB	13.0730	6.6758	11.1109	11.9811	9.1277	11.6990
NMF	20.1934	19.9607	20.1565	22.1384	22.1384	22.1384
DR!	5.0913	5.0739	5.0879	3.8736	3.2792	4.2428
SeekDoc	2.7067	**2.5982**	**2.6897**	0.8081	0.8479	**0.5857**

From the Table 2, we compared the MSE values with five different algorithms on six different features. It is obvious that the *SeekDoc* achieves the better performance on dataset. On the one hand,

SeekDoc have the best ACC value than other five baselines for DIHN, DITHN, DITHNE features respectively. On the other hand, when the method uses DITHNE features, the minimum MSE value is obtained. The second smaller value is DITE features, next, in order from smallest to largest, it is DIHNE features, DIHN features, DITHN features, DIT features. From the perspective of MSE values, the feature embedding method we used is reasonable and effective.

4.6.2. Study of timing consuming

We further evaluate the time consumption of *SeekDoc* and other five algorithms measured in seconds. The running time is shown in Figure 7. It can be found DR! is a time-consumer among all algorithms, while *SeekDoc* takes a little longer than other algorithms. However, *SeekDoc* performs much better than other algorithms. It can be seen from the analysis that the effective and efficient of *SeekDoc*.

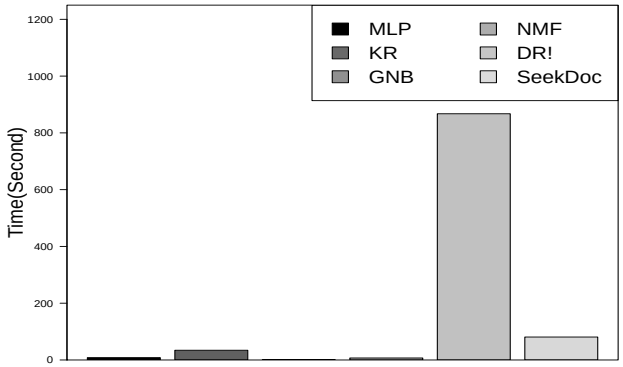


Figure 7. The comparison of seekDoc with different algorithm on time.

5. Related work

In our paper, we propose an advance approach named as *SeekDoc*, that help patients find the influential doctor for a given disease using online healthcare comments data. We summarize the most relevant studies as follows.

5.1. Healthcare field

Recently, it is very important for the patient to find a doctor who is suitable for the disease. More and more work has been done on healthcare. Edward et al. [18] proposed a GRaph-based Attention Model to solve the problem of insufficient data, which supplemented electronic health records(EHR) with hierarchical information inherent in medical ontology. Choi et al. [19] proposed Med2Vec, which not only learns the representations for both medical codes and visits from large EHR datasets with over million visits, but also allows us to interpret the learned representations confirmed positively by clinical experts. To exploit the potential information captured in EHRs, Ayoub et al. [20] proposed MI-BiLSTM, a multimodal bidirectional long short-term memory-based framework for cardiovascular risk prediction that integrates medical texts and structured clinical information. Suresh et al. [21] combined with multi-task learning model to help doctors and nurses give patients more appropriate treatment

through clinical prediction. Yin Zhang et al. [22] proposed a doctor recommendation based on hybrid matrix factorization. Jiang Ling et al. [23] investigated the approaches for measuring user similarity in online health social websites. Chang Xu et al. [24] proposed an online medical service recommendation scheme to protect privacy in the electronic health care system, which takes the doctor's reputation score and the similarity between user needs and doctor information as the basis for recommending medical services. Yan et al. [25] proposed a hybrid recommendation algorithm (PMF-CNN) based on deep learning is proposed for doctor recommendation, PMF-CNN model uses convolutional neural network to learn the context features of review information, so as to extract more accurate feature representation to realize the modeling of review information. Bo Jin et al. [26] proposed an effective and robust architecture for heart prediction. Ling Chen et al. [27] proposed a new network-based algorithm that ranks heterogeneous objects in a medical information network. Mateo et al. [28] showed that the healthcare expert system was implemented on the group cooperation model.

5.2. Autoencoder

As an important training model in deep learning, autoencoder has been paid more and more attention by researchers for its good performance in natural language processing. Jiawei Zhang et al. [29] introduced an embedded framework based on deep autoencoder, which aims to learn the embedded vector of users in emerging networks and reduce the degradation of embedding performance caused by network sparse structure. Bengio et al. [30] studied a hierarchical unsupervised learning algorithm empirically and explored variables to better understand its success and extended it to situations where the inputs are continuous or where the structure of the input distribution is not revealing enough about the variable to be predicted in a supervised task. Goodfellow et al. [31] proposed a number of empirical tests that directly measure the degree to which some learned features are invariant to different input transformations. These results further justify the use of "deep" vs. "shallower" representations, but suggest that mechanisms beyond merely stacking one autoencoder on top of another may be important for achieving invariance. Xiong Dapeng et al. [32] presented a computational DeepConPred, which includes a pipeline of two novel deep-learning-based methods (DeepCCon and DeepRCon) as well as a contact refinement step, to improve the prediction of long-rang residue contacts from primary sequences.

5.3. XGBoost classifier

Extreme gradient boosting(XGBoost) is a scalable machine learning system for tree xboosting that proposed by Chen in 2016. The advantage of the XGBoost algorithm is that it can automatically utilize the multi-threading of the CPU for parallelism [33], at the same time, it can improve the accuracy of the algorithm.

6. Conclusions and discussion

In this paper, we proposed a novel predicting model for patient-doctor score using XGBoost methods. Compared to popular methods such as MLP, GPC, KR, GNB, NMF and Dr. Right!, our methed exhibits superior performance in the prediction of influential doctor. Besides the popular methods, we also use parts of features (DIT, DIHN, DITHN, DITE, DIHNE, DITHNE) as inputs to demonstrate the

state-of-the-art of the feature-embedding algorithm. In the experimental data analysis and preprocessing, we used node embedding vectors to represent the doctor and disease with neighborhood network information, and used word embedding vectors to represent the patient comment in details. Further, we used autoencoder to extract latent features into the network. Then all the work is based the basic XGBoost model. Finally, a large number of experiments using the real-world datasets are used to demonstrate the effectiveness of *SeekDoc* in performance.

SeekDoc has a certain improvement in predicting the most effective doctors, but there are still some shortcomings, such as feature embedding based on the network structure, no analysis and mining of heterogeneous network itself. In the future work, we consider adding more case data to *SeekDoc* to extract more influential feature data from the network. In addition, the performance of the model can be improved by adjusting the parameters or combining with other algorithms to further optimize the algorithm.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities 2412018QD022, NSFC (under Grant No.61976050, 61972384, 21473025), Jilin Provincial Science and Technology Department under Grant No. 20190302109GX and Jilin Education Department No. JJKH20200791KJ. We are grateful to all anonymous reviewers whose insightful comments have helped us to improve the work.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. M. E. J. Newman, The structure and function of complex networks, *SIAM Rev.*, **45** (2003), 167–256.
2. L. Jiang, P. Wang, K. Cheng, K. Liu, M. Yin, B. Jin Y. Fu, et al., EduHawkes: A Neural Hawkes Process Approach for Online Study Behavior Modeling, *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 2021.
3. A. Hosseini, T. Chen, W. Wu, Y. Sun, M. Sarrafzadeh, Heteromed: Heterogeneous information network for medical diagnosis, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.
4. X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, et al., Personalized entity recommendation: A heterogeneous information network approach, *Proceedings of the 7th ACM international conference on Web search and data mining*, 2014.
5. M. Yousefi-Azar, V. Varadharajan, L. Hamey, U. Tupakula, Autoencoder-based feature learning for cyber security applications, *2017 International joint conference on neural networks (IJCNN)*, 2017.
6. Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436.

7. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.
8. T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, Xgboost: extreme gradient boosting, *R Package Version*, **1** (2015), 1–4.
9. Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, preprint, arXiv:1402.3722.
10. A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.
11. V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010.
12. Y. Li, Z. Zhang, Z. Teng, X. Liu, Predamyl-mlp: Prediction of amyloid proteins using multilayer perceptron, *Comput. Math. Methods Med.*, **2020** (2020).
13. I. Mihaylov, M. Nisheva, D. Vassilev, Application of machine learning models for survival prognosis in breast cancer studies, *Information*, **10** (2019), 93.
14. M. L. Gadebe, Smartphone nave bayes human activity recognition using personalized datasets, *J. Adv. Comput. Intell. Intell. Inf.*, **24** (2020), 685–702.
15. J. T. Chien, Nonnegative matrix factorization, *Source Sep. Mach. Learn.*, **2019** (2019), 161–229.
16. X. Xu, Y. Fu, H. Xiong, B. Jin, X. Li, S. Hu, et al., Dr. right!: Embedding-based adaptively-weighted mixture multi-classification model for finding right doctors with healthcare experience data, *2018 IEEE International Conference on Data Mining (ICDM)*, 2018.
17. M. Tokic, G. Palm, Value-difference based exploration: adaptive control between epsilon-greedy and softmax, in *Annual Conference on Artificial Intelligence*, Springer, 2011.
18. E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, Gram: graph-based attention model for healthcare representation learning, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
19. E. Choi, M. T. Bahadori, E. Searles, C. Coffey, J. Sun, Multi-layer representation learning for medical concepts, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
20. A. Bagheri, T. K. J. Groenhof, W. B. Veldhuis, P. A. de Jong, F. W. Asselbergs, D. L. Oberski, Multimodal learning for cardiovascular risk prediction using EHR data, preprint, arXiv: 2008.11979.
21. H. Suresh, J. J. Gong, J. V. Guttag, Learning tasks for multitask learning: Heterogenous patient populations in the icu, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.
22. Y. Zhang, M. Chen, D. Huang, D. Wu, Y. Li, idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization, *Future Gene. Comput. Syst.*, **66** (2017), 30–35.
23. J. Ling, C. C. Yang, User recommendation in healthcare social media by assessing user similarity in heterogeneous network, *Artif. Intell. Med.*, **81** (2017), S0933365717301185.

24. C. Xu, J. Wang, L. Zhu, C. Zhang, K. Sharif, PPMR: A privacy-preserving online medical service recommendation scheme in ehealthcare system, *IEEE Int. Things J.*, **6** (2019), 5665–5673.
25. Y. Yan, G. Yu, X. Yan, Online doctor recommendation with convolutional neural network and sparse inputs, *Comput. Intell. Neurosci.*, **2020** (2020).
26. B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, X. P. Wei, Predicting the risk of heart failure with ehr sequential data modeling, *IEEE Access*, **6** (2018), 9256–9261.
27. L. Chen, X. Li, J. Han, Medrank: discovering influential medical treatments from literature by information network analysis, *Proceedings of the Twenty-Fourth Australasian Database Conference*, 2013.
28. R. M. A. Mateo, B. D. Gerardo, J. Lee, Healthcare expert system based on the group cooperation model, *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*, 2007.
29. J. Zhang, C. Xia, C. Zhang, L. Cui, Y. Fu, S. Y. Philip, Bl-mne: emerging heterogeneous social network embedding through broad learning with aligned autoencoder, *2017 IEEE International Conference on Data Mining (ICDM)*, 2017.
30. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. Montreal, Greedy layer-wise training of deep networks, *Adv. Neural Inf. Proc. Syst.*, **19** (2007), 153–160.
31. I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, A. Y. Ng, Measuring invariances in deep networks, in *Adv. Neural Inf. Proc. Syst.*, **22** (2009), 646–654.
32. D. Xiong, J. Zeng, H. Gong, A deep learning framework for improving long-range residue-residue contact prediction using a hierarchical strategy, *Bioinformatics*, **33** (2017), 2675–2683.
33. M. Gumus, M. S. Kiran, Crude oil price forecasting using xgboost, *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017.

AIMS Press