An Embedding-Based Approach for Oral Disease Diagnosis Prediction from Electronic Medical Records

Guangkai Li
MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences
Beijing, China
liguangkai15@mails.ucas.ac.cn

Songmao Zhang
MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing, China
smzhang@math.ac.cn

Jie Liang
Department of Oral and Maxillofacial Surgery, Peking University School and Hospital of Stomatology
Beijing China
Liangjie.pkuss@gmail.com

Zhanqiang Cao
Information Center, Peking University School and Hospital of Stomatology
Beijing, China
caozhanqiang@pkuss.bjmu.edu.cn

Chuanbin Guo
Department of Oral and Maxillofacial Surgery, Peking University School and Hospital of Stomatology
Beijing, China
guodazuo@sina.com.cn

ABSTRACT

This paper reports a diagnosis prediction study from electronic medical records (EMRs) of oral diseases. We propose to learn continuous vector representations (embeddings) of symptoms and diagnoses through training neural networks. To the best of our knowledge, this is the first attempt to apply word embedding to predicting diagnoses from stomatologic EMR data. Evaluations on real-world EMR datasets from eleven departments in Peking University School and Hospital Stomatology demonstrate that our model has produced promising results in diagnosis prediction. Compared with classic machine learning algorithms, our model captures the correlation between symptoms and diagnoses, which leads to the best performance in terms of accuracy, precision, recall, F-score, and Cohen's kappa statistic. Visualizations illustrate the quality of the symptom and diagnosis embeddings generated by our approach.

CCS Concepts

•Computing methodologies→ Artificial intelligence •Applied computing→ Health care information systems.

Keywords

Diagnosis prediction; Word embedding; Electronic medical record

1. INTRODUCTION

Electronic medical records (EMRs) or electronic health records (EHRs) have been a valuable medical asset that can be used to facilitate clinical decision making by discovering hidden knowledge and predicting diagnoses [1]. Lots of research efforts have been focusing on EMR data, presenting algorithms and models targeting various diseases including cancers [2, 3, 4, 5], diabetes [6, 7] and heart failures [8]. These works use either pattern matching schemes (regular expressions) or language modeling-based methods to extract features from EMRs [9]. While features extracted in these ways can reduce the amount of data processing and produce meaningful results, they often fail to capture the correlations inherent in patient clinical data.

Established in 1941, Peking University School and Hospital of Stomatology (PKUSS) is one of the most prestigious hospitals in China for oral diseases. PKUSS has been enhancing the development of EMR information systems and over the years has accumulated a large amount of clinical data. Under the support of Beijing Municipal Science & Technology Commission Projects, we intend to exploit the PKUSS data by exploring state-of-the-art AI technologies that can capture correlations hidden in the EMRs. The stomatologic data is of high dimensionality and complexity, and there have been fewer studies on them compared with those on tumors and heart diseases.

In this paper, we propose a model of diagnosis prediction based on word embedding, where symptoms and diagnoses are represented as vectors in real space and trained using artificial neural networks. Our work is inspired by word embedding [10, 11], a method of representing words as continuous vectors, which has proven particularly useful in various natural language processing tasks. Usually, the vector representations of words are computed using their contexts so that words with similar meanings will have similar vector representations. Applying word embedding to the domain of biomedical informatics has mainly focused on medical concept representation learning [26, 27, 28, 29]. Unlike these works, our model learns symptom and diagnosis embeddings with a tailored neural network, where one symptom or diagnosis may consist of multiple medical concepts. Our model is specially designed for EMR diagnosis prediction, so that in the vector space mapped by the learned embeddings, medical symptoms/diagnoses with correlations are positioned closer than those not related. We evaluate the model on the real-world EMRs from eleven departments of PKUSS. The results demonstrate that the proposed approach outperforms classic machine learning algorithms including k-NN, logistic regression, C4.5, naïve Bayes and SVM in terms of classification accuracy, weighted precision, recall and f1-score, and Cohen's kappa statistic. To evaluate the quality of the embeddings learned, we further design visualizations incorporating TF-IDF to map embeddings into 2D

ICMHI 2018, June 8–10, 2018, Tsukuba, Japan

ACM ISBN 978-1-4503-6389-1/18/06...$15.00
https://doi.org/10.1145/3239438.3239451

space, and the results show that spatial distances among vectors are in accordance with medical explanations.

The main contribution of our work presented in this paper can be summarized as follows. 1) We propose to learn continuous vector representations (embeddings) of symptoms and diagnoses through training neural networks. To the best of our knowledge, this is the first attempt to apply word embedding to predicting diagnoses from stomatologic EMR data. 2) Evaluations on real-world EMRs demonstrate that our model has produced promising results in diagnosis prediction. Compared with classic machine learning algorithms, our model captures the correlation between symptoms and diagnoses, which leads to the best performance by diverse measures. Visualizations illustrate the quality of the symptom and diagnosis embeddings generated by our approach.

2. RELATED WORK

In this section, we briefly introduce the work on discovering hidden medical knowledge from EMRs and learning continuous representations or embeddings of words, which are relevant to our study in this paper.

2.1 Discovering Knowledge from EMRs

Discovering hidden knowledge from EMRs in biomedical informatics can be classified into two categories: the rule-based approaches and the machine learning-based approaches.

Typical rule-based systems deduce certain medical conclusions about patients by applying logical constraints to the symptoms extracted from EHRs, e.g., phenotypes of patients can be drawn based on conditions like *hemoglobin <10 AND age >60* [13]. The way rules are generated is the key factor affecting the performance of such systems. Most of them construct rules by adopting clinical judgments of physicians or expert opinions [2, 14], using guidelines or recommendations from health organizations [6, 15], or improving previous rules [16]. Diagnoses from rule-based systems are easy to interpret, but the rules can only cover limited, known conditions thus the diagnoses made are limited.

Machine learning methodology is well suited for handling nonlinear correlations between feature vector components and metadata, such as patient subcategories [17]. The starting point for machine learning is a dataset of training samples, which in the context of EMR typically originate from individual patients. Each sample is represented by a feature vector, which can be any combination of data items that are stored in EMR systems. Typical machine learning models used in biomedical informatics include k-NN, logistic regression, C4.5, naïve Bayes, SVM, and so on. Many studies report the comparison of these models for prediction tasks [3, 7, 18, 19, 20]. Take Potter et al. [4] for example, which runs 56 classification algorithms on breast cancer datasets for comparison, and the result shows that there is no universal classification algorithm that stands out for all datasets.

2.2 Word Embedding

Word embedding is one of state-of-the-art technologies for natural language processing in recent years. It learns a continuous vector representation of words from context words so that the vectors can represent syntactic/semantic information of words. Bengio et al. [10] proposed a probabilistic neural network language model (NNLM) for learning word representations. Furthermore, Mikolov et al. [11] proposed a simpler and more effective neural network model for learning word representations, including the continuous skip-gram model and the continuous bag-of-word model (CBOW). Pennington et al. [21] proposed GloVe, a word representation algorithm based on global co-occurrence matrixes. In addition, there are works that apply the idea of word embedding to tasks like sentiment classification [22], information retrieval [23], machine translation [24], question answering [25], and others.

The neural network based representation learning approaches, although have shown promises in many domains, are relatively less studied in biomedical informatics. Antonio et al. [26] applied the word2vec model to processing PubMed medical texts. The obtained statistics are not satisfactory, showing that the ability of word2vec in retrieving significant words from restricted corpora is not suitable for high-precision tasks. Choi et al. [27, 28] proposed an effective method for medical concept representation learning by using real-world EHR datasets. Tran et al. [29] derived vector representations of medical objects with their restricted Boltzmann machines in the suicide risk stratification task. Unlike these works, our study aims to learn continuous vector representations (embeddings) of symptoms and diagnoses from EMRs, where one symptom may contain multiple medical concepts. The vectors learned in our approach target diagnosis prediction tasks, whereas others learn vectors based on skip-gram models and then apply them to medical applications.

3. MATERIALS AND METHODS

3.1 Data Description

We have obtained a total of 7208 PKUSS EMRs. As shown in Table 1, these records belong to the eleven departments of the hospital: Pediatric Dentistry, Oral and Maxillofacial Surgery, Laser Dentistry, Emergency, Oral Medicine, Prosthodontics, Geriatric Dentistry, General Dentistry, Orthodontics, Periodontology, and Implant Dentistry. Each record consists of six parts: chief complaints (C.C.), history of present illness (HPI), past medical history (PMH), family history (F.H.), physical examination (P.E.), and the diagnosis. Medical history information in C.C., HPI, PMH, and F.H. is mostly described in natural language, as exemplified by a HPI "The parent has found many caries, and because the child is little and does not cooperate with conventional treatments, requires applying general anesthesia for treatment". Conversely, physical examination information is mostly presented in medical terminologies, such as a P.E. "Complete eruption, percussion test(-), no mobility, normal gingival".

3.2 Feature Extraction and Representation

Raw EMRs are often noisy, sparse, and contain unstructured information (e.g., text) that are not directly computable. Extracting suitable features from EMRs decides the prediction performance for both rule-based systems and machine learning-based models. A diagnosis, in the sense of diagnostic procedure, can be regarded as an attempt to classify a patient's data into separate and distinct disease classes. Therefore, the prediction of diagnosis by EMR data can be regarded as performing a classification task of EMRs. The most important information in an EMR is the symptoms, i.e., clinical manifestations from both patient and physician. Symptoms are often nonspecific, and their combinations can be suggestive of certain diagnoses. Thus in our study symptoms are extracted from EMRs as features. Particularly, we extract symptoms based on bigram [30] from medical history including C.C., HPI, PMH, and F.H., and use regular expressions to match the symptoms from P.E. The number of obtained features of the eleven PKUSS datasets is listed in Table 1.

Table 1. EMR datasets of PKUSS

Dataset	Records	Features (symptoms)	Diagnosis classes
Pediatric Dentistry	669	609	6
Oral and Maxillofacial Surgery	728	211	8
Laser Dentistry	569	190	7
Emergency	371	242	6
Oral Medicine	425	253	4
Prosthodontics	1135	321	3
Geriatric Dentistry	320	286	2
General Dentistry	407	283	3
Orthodontics	671	886	4
Periodontology	1016	706	3
Implant Dentistry	897	111	6

Suppose the sets of symptoms are $\mathcal{S}_{C.C.}$, $\mathcal{S}_{HPI}$, $\mathcal{S}_{\mathrm{PMH}}$, $\mathcal{S}_{\mathrm{F.H.}}$, and $\mathcal{S}_{P.E.}$, respectively. For a given EMR, the feature vector of its C.C. can be represented as $x_{C.C.} = (x_{C.C.}^{(1)}, \ldots, x_{C.C.}^{(j)}, \ldots, x_{C.C.}^{(n1)})$, where the *j*-th element is defined as follows.

$$x_{C.C.}^{(j)} = \begin{cases} 1 & \text{C.C. of the EMR contains} \\ & \text{j-th symptom of } \mathrm{S}_{\mathrm{C.C.}} \\ 0 & \text{otherwise} \end{cases}$$

Similarly, feature vectors of HPI, PMH, F.H. and P.E. of the EMR can be represented as x_{HPI}, x_{PMH}, $x_{F.H.}$ and $x_{P.E.}$. Then, we concatenate these vectors as the feature vector of the EMR as follows.

$$x = (x_{C.C.}, x_{HPI}, x_{PMH}, x_{F.H.}, x_{P.E.})$$

Suppose we have m EMRs, from which n symptoms are extracted to form set $\mathcal{S} = \{symp_i\}_{i=1}^{n}$, and l diagnosis classes exist to form set $\mathcal{D} = \{diag_i\}_{i=1}^{l}$. An EMR dataset can be represented as $\{(x_i, y_i)\}_{i=1}^{m}$, where $x_i = (x_i^{(1)}, \ldots, x_i^{(n)})$ is the feature vector of the *i*-th record, and $y_i \in \{1, \ldots, l\}$ represents its diagnosis. The objective is to learn a model satisfying $f(x_i, y_i) = \hat{P}(y_i \mid x_i)$, i.e., given the feature vector of symptoms of a patient, its diagnosis can obtain the highest probability over other classes.

3.3 Training a Neural Network to Learn Embeddings of Symptoms and Diagnoses

Our model is based on a neural network with the input layer, one hidden layer, and the output layer, as illustrated in Figure 1. The input layer consists of n nodes, and we define operation $\odot$ between x_i and C as:

$$x_i \odot C = (x_i^{(1)} C(symp_1), \ldots, x_i^{(n)} C(symp_n))^{\top}$$

Function $C(\cdot)$ maps symptom element i of $\mathcal{S}$ to real vector $C(i) \in \mathbb{R}^k$, where k represents the dimension of embedding representation. Correspondingly, we define function $V(\cdot)$ to map diagnosis element j of $\mathcal{D}$ to $\mathbb{R}^k$. Mappings $C(\cdot)$ and $V(\cdot)$ respectively represent embeddings associated with each symptom and diagnosis in dataset. In practice, they are represented by a $n \times k$ matrix and a $l \times k$ matrix of free parameter k.

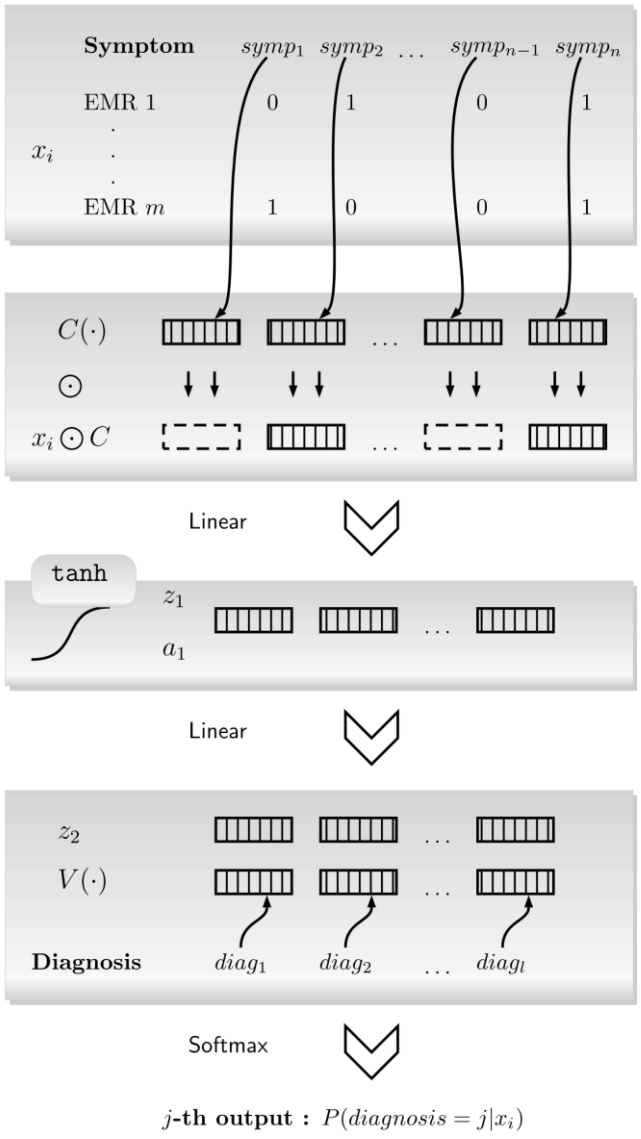


Figure 1. A neural network architecture for learning symptom and diagnosis embeddings.

In the hidden layer, as shown in Figure 1, the hyperbolic tangent function is employed as an activation function, and the mathematical formulation of the input and hidden layer is presented as follows.

$$z_1 = W_1(x_i \odot C) + B_1$$

$$a_1 = tanh(z_1)$$

$$z_2 = W_2 a_1 + B_2$$

In this formulation, W_1, W_2 and B_1, B_2 are the weight matrices and bias terms, respectively. In particular, B_1 and B_2 are in the form of a matrix, each row being a *k*-dimensional vector.

The output layer consists of l nodes, and based on z_2, we use a softmax function (normalized exponential function) to compute

the probability of each diagnosis, where the *j*-th node is calculated as follows.

$$\mathcal{O}_{ij} = \frac{e^{z_2[j]V[j]^\top - \mathcal{U}_{ij}}}{\sum_{k=1}^{l} e^{z_2[k]V[k]^\top - \mathcal{U}_{ik}}}$$

In this formula, $\mathcal{U}_{ij}$ is the 2-norm of the difference between the mean of symptom vector in the *i*-th EMR and the *j*-th diagnosis vector, as follows.

$$\mathcal{U}_{ij} = \| \frac{x_i C}{\| x_i \|_1} - V[j] \|_2$$

Training is achieved by looking for θ that maximizes log-likelihood loss function:

$$\mathcal{L}(y, \mathcal{O}; \theta) = \frac{1}{m} \sum_{i}^{m} \sum_{j}^{l} y_{ij} log \mathcal{O}_{ij}$$

The stochastic gradient algorithm is used for training and the updating rule of the gradient is defined as:

$$\theta \leftarrow \theta + \varepsilon \frac{\partial \mathcal{L}(y, \mathcal{O}; \theta)}{\partial \theta}$$

where $\theta = (C, V, W_1, W_2, B_1, B_2)$ is the parameter to be trained.

4. EVALUATION AND RESULTS

To evaluate our approach, we design comparative experiments to demonstrate the performance, and present visualizations to validate the numerical embeddings learned, described in two subsections as follows.

4.1 Prediction Performance

4.1.1 Experiments' Setup

We compare the performance of the proposed approach with five benchmark classification methods: k-NN, logistic regression, C4.5, naïve Bayes, and SVM. These models are frequently utilized in a wide range of domains and applications, and highly recognized for classification tasks. We implement them with Python Scikit-Learn 0.19.0.

For our approach, we select learning rate $\lambda = 0.05$ for stochastic gradient ascent, and embedding dimension $k = 20$ for each symptom and diagnosis. The number of nodes in the hidden layer of our neural network is determined by empirical formula $nh = \sqrt{ni + no} + o$, where ni is the number of nodes of the input layer, no the output layer, and o a constant between 1 and 10. We conduct 5-fold cross-validation and report on the average performance.

To measure the performance, we use the classification accuracy, weighted precision, weighted recall, weighted f1-score and Cohen's kappa statistic. Concretely, classification accuracy depends on the number of samples correctly classified and is computed by formula:

$$ACC = \frac{t}{N_{sum}} \times 100$$

where N_{sum} is the total number of sample cases and t those correctly classified.

For the binary classification problem, the precision, recall and F-score are defined as: $P = TP / (TP + FP)$, $R = TP / (TP + FN)$, and $F1 = (2 \times P \times R) / (P + R)$, where TP, FP and FN denote true positives, false positives, and false negatives, respectively. For the multiclass classification problem, though, weighted averages are used based on the corresponding measures of each class:

$$\text{weighted-}P = \sum_{i=1}^{l} \alpha_i P_i, \ \text{weighted-}R = \sum_{i=1}^{l} \alpha_i R_i$$

$$\text{weighted-}F1 = \sum_{i=1}^{l} \alpha_i F1_i$$

where $(P_1, R_1, F1_1), (P_2, R_2, F1_2), ..., (P_l, R_l, F1_l)$ denote the precision, recall and f1-score of each class, and the weighting α_i is calculated using the number of samples belonging to one class divided by the total number of samples in dataset.

Cohen's kappa [31] was first introduced as a metric to measure the degree of agreement or disagreement of two or more people observing the same phenomenon. It is a very useful, yet under-utilized, metric. Especially for multiclass classification problems, Cohen's kappa statistic can help provide a complete picture of the performance. It basically tells how better your classifier performs over a classifier that simply guesses randomly according to the frequency of each class. Cohen's kappa is defined as:

$$\mathcal{K} = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the observed agreement and p_e the expected agreement. Cohen's kappa is always less than or equal to 1. There is none standardized way to interpret its values, and we use the interpretation by Landis and Koch [32], as shown in Table 2.

Table 2. The interpretation of the kappa value [32].

Kappa value	Interpretation
< 0	no agreement
0.00 — 0.20	slight agreement
0.21 — 0.40	fair agreement
0.41 — 0.60	moderate agreement
0.61 — 0.80	substantial agreement
0.81 — 1	perfect agreement

4.1.2 Experimental Results and Analysis

We run the five benchmark classifiers and our approach on the eleven PKUSS datasets in Table 1, and the accuracy of diagnosis prediction is shown in Table 3. One can see that our approach has achieved the best performance on all datasets, showing the potential of the embedding-based models in diagnosis prediction of EMR data, whereas other classifiers perform variedly for different datasets. Moreover, we observe that there is a large performance disparity in different datasets for one classifier, e.g., the accuracy of our approach ranges from 70% to 96%, and SVM from 66% to 96%. This is not surprising due to the distinction of specialties in different departments. Doctors having personalized writing styles in EMRs may also cause varying degrees of noise to the datasets. In particular, naïve Bayes obtains the worst accuracy on all datasets, e.g., 19.8% on the Implant Dentistry dataset while all other classifiers achieved above 80%. Naïve Bayes classifiers assume that given the class variable, the values of features are independent with each other. For the PKUSS datasets, features are extracted from symptoms and symptoms are often correlated. For example, "no mobility" and "normal gingival" are related in EMR for which the diagnosis is deep caries. Such dependences among

features jeopardize classifiers like naïve Bayes whereas can be captured by our model that introduces the correlation between symptoms and diagnoses by embedding.

Table 3. The prediction accuracy of classifiers on eleven PKUSS datasets.

Datasets	k-NN	logistic regression	C4.5	naïve Bayes	SVM	Our approach
Pediatric Dentistry	70.5	75.0	75.0	33.8	76.4	**77.2**
Oral and Maxillofacial Surgery	70.9	81.7	68.9	58.7	79.7	**82.4**
Laser Dentistry	84.4	90.5	86.2	73.2	**91.3**	**91.3**
Emergency	72.3	71.0	71.0	64.4	72.3	**75.0**
Oral Medicine	89.5	83.7	90.6	80.2	86.0	**93.0**
Prosthodontics	**89.0**	87.7	85.1	70.3	86.4	**89.0**
Geriatric Dentistry	89.2	87.6	86.1	87.6	89.2	**92.3**
General Dentistry	75.9	**92.7**	75.9	61.4	90.3	**92.7**
Orthodontics	55.7	67.3	55.7	29.7	65.9	**70.2**
Periodontology	94.6	94.6	95.1	85.3	95.6	**96.0**
Implant Dentistry	80.1	80.1	80.6	19.8	80.6	**81.2**

Table 4. The weighted precision, recall and f1-score of classifiers on eleven PKUSS datasets.

Datasets	k-NN	logistic regression	C4.5	naïve Bayes	SVM	Our approach
Weighted-P						
Pediatric Dentistry	0.71	0.76	0.76	0.36	0.76	**0.77**
Oral and Maxillofacial Surgery	0.75	**0.84**	0.77	0.65	0.80	**0.84**
Laser Dentistry	0.85	**0.91**	0.86	0.77	**0.91**	**0.91**
Emergency	**0.80**	0.75	0.72	0.64	0.72	0.75
Oral Medicine	0.89	0.82	0.90	0.84	0.84	**0.93**
Prosthodontics	**0.89**	0.87	0.85	0.78	0.86	**0.89**
Geriatric Dentistry	0.89	0.87	0.86	0.88	0.89	**0.92**
General Dentistry	0.76	**0.93**	0.80	0.66	0.90	**0.93**
Orthodontics	0.55	0.69	0.60	0.30	0.68	**0.73**
Periodontology	0.92	0.92	0.93	0.92	0.93	**0.95**
Implant Dentistry	0.72	0.77	**0.78**	0.76	0.76	0.76
Weighted-R						
Pediatric Dentistry	0.70	0.75	0.75	0.33	0.76	**0.77**
Oral and Maxillofacial Surgery	0.70	0.81	0.71	0.58	0.79	**0.82**
Laser Dentistry	0.84	0.90	0.86	0.73	**0.91**	**0.91**
Emergency	0.72	0.71	0.71	0.64	0.72	**0.75**
Oral Medicine	0.89	0.83	0.90	0.80	0.86	**0.93**
Prosthodontics	**0.89**	0.87	0.85	0.70	0.86	**0.89**
Geriatric Dentistry	0.89	0.87	0.86	0.87	0.89	**0.92**
General Dentistry	0.75	**0.92**	0.79	0.61	0.90	**0.92**
Orthodontics	0.55	0.67	0.55	0.29	0.65	**0.70**
Periodontology	0.94	0.64	0.95	0.85	0.95	**0.96**
Implant Dentistry	0.80	0.80	0.80	0.19	0.80	**0.81**
Weighted-F1						
Pediatric Dentistry	0.69	0.74	0.74	0.31	0.76	**0.77**
Oral and Maxillofacial Surgery	0.71	0.81	0.73	0.57	0.79	**0.82**
Laser Dentistry	0.84	0.90	0.85	0.71	**0.91**	**0.91**
Emergency	0.71	0.72	0.71	0.63	0.72	**0.75**
Oral Medicine	0.89	0.81	0.90	0.81	0.83	**0.92**
Prosthodontics	**0.89**	0.87	0.85	0.70	0.86	**0.89**
Geriatric Dentistry	0.89	0.87	0.86	0.87	0.89	**0.92**
General Dentistry	0.74	**0.92**	0.79	0.61	0.90	**0.92**
Orthodontics	0.53	0.68	0.57	0.26	0.66	**0.71**
Periodontology	0.93	0.93	0.94	0.88	0.94	**0.95**
Implant Dentistry	0.74	0.77	**0.78**	0.20	0.77	0.77

In EMR data, the phenomenon of imbalance is ubiquitous, as the probability of occurrence of various diseases differs. For example, the medical records diagnosed with chronic periodontitis are far more than others in the Periodontology dataset, thus a classifier could achieve a high accuracy if all the medical records were classified as chronic periodontitis. This indicates the inadequacy of solely using accuracy to measure the performance of classifiers. We further compute the weighted precision, recall and f1-score of each classifier in Table 4. The weighted average is similar to an arithmetic mean, where instead of each contributing equally to the final average, some data points contribute more than others. In our experiment, the weight is proportional to the number of true samples for each label. This can result in an F-score that is not between the values of precision and recall. For example, the weighted precision of the k-NN algorithm on the Emergency dataset is 0.80, the weighted recall 0.72, and the weighted f1-score 0.71. The weighted precision of our approach is lower than k-NN on the Emergency dataset and lower than logistic regression and C4.5 on the Implant Dentistry dataset. Despite these, we have achieved the highest precision on all other datasets. In terms of the weighted recall and the weighted f1-score, our approach performs the best on all datasets, except the f1-score being slightly lower than C4.5 on the Implant Dentistry dataset. Overall, our approach demonstrates a clear advantage over other classifiers.

Cohen's kappa statistic can handle well both multiclass and imbalanced classification problems. The detailed results of the kappa statistic analysis of classifiers are presented in Table 5. Again, our approach outperforms others on most datasets, except being lower than logistic regression, C4.5, and SVM on the Implant Dentistry dataset, and lower than kNN on the Prosthodontics. According to the interpretation of the kappa statistic in Table 2, the performance of our approach is perfect on four datasets, substantial on five datasets, moderate on one dataset, and fair on one dataset. Particularly, we notice that all classifiers perform poorly on the Implant Dentistry dataset, with the maximum kappa statistic 0.43 and the lowest 0.08. This dataset has six diagnosis classes, and out of 897 EMRs, 712 are diagnosed as "dentition defect", 57 "mandibular dentition defect", 55 "residual root", 32 "maxillary dentition defect", 27 "chronic periapical periodontitis", and 21 "residual crown". Among the six diagnoses, "mandibular dentition defect" and "maxillary dentition defect" actually belong to "dentition defect", and most of their symptoms are the same. As a result, distinguishing among classes becomes vague and difficult.

Table 5. The Cohen's kappa statistic of classifiers on eleven PKUSS datasets.

Datasets	k-NN	logistic regression	C4.5	naïve Bayes	SVM	Our approach
Pediatric Dentistry	0.61	0.68	0.67	0.22	0.70	**0.71**
Oral and Maxillofacial Surgery	0.66	0.78	0.67	0.52	0.75	**0.79**
Laser Dentistry	0.81	0.88	0.83	0.67	**0.89**	**0.89**
Emergency	0.65	0.64	0.64	0.55	0.65	**0.69**
Oral Medicine	0.84	0.76	0.86	0.73	0.79	**0.90**
Prosthodontics	**0.79**	0.76	0.71	0.46	0.73	0.78
Geriatric Dentistry	0.77	0.74	0.71	0.73	0.77	**0.83**
General Dentistry	0.61	**0.88**	0.68	0.43	0.84	**0.88**
Orthodontics	0.33	0.54	0.37	0.09	0.51	**0.58**
Periodontology	0.53	0.50	0.56	0.26	0.55	**0.65**
Implant Dentistry	0.29	0.38	**0.43**	0.08	0.38	0.37

4.2 Symptom and Diagnosis Visualization

Our approach learns the embeddings of symptoms and diagnoses so as to identify the correlation between them, resulting in both symptoms and diagnoses represented as vectors in a high-dimensional space. This says that each symptom or diagnosis corresponds to a point in the high-dimensional space, and the distance between them reflects their correlation. In order to evaluate the validity of our approach in revealing the correlation between symptoms and diagnoses, we visualize the high-dimensional vectors into 2D space by using a dimensionality reduction algorithm t-SNE [33]. There are no straightforward ways to quantify the quality of embeddings, and we propose to incorporate the Term Frequency Inverse Document Frequency (TF-IDF) to evaluate symptoms in each diagnosis. For symptom $symp_i$ in diagnosis $diag_j$, its TF-IDF value can be calculated as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad , \quad idf_i = log \frac{|D|}{1+|\{j : symp_i \in diag_j\}|}$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

In these formulas, $n_{i,j}$ is the number that the symptom $symp_i$ occurs in the EMRs diagnosed as $diag_j$, and the denominator is the sum of the number of all the symptoms that occur in the diagnosis $diag_j$; $|D|$ is the number of diagnoses, and $|\{j : symp_i \in diag_j\}|$ is the number of diagnoses that contain the

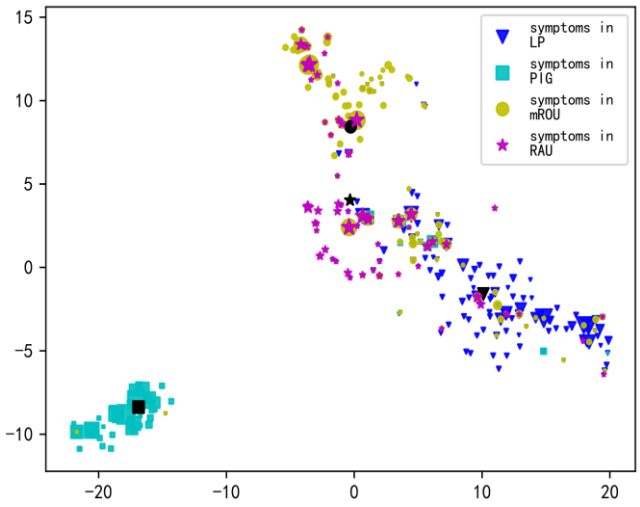
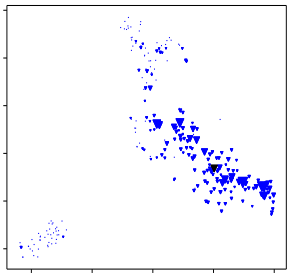
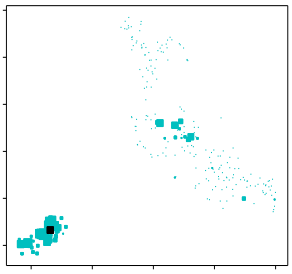


Figure 2. An overall visualization of the diagnoses and symptoms in the Oral Medicine dataset.

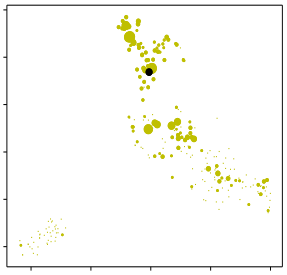
symptom $symp_i$.



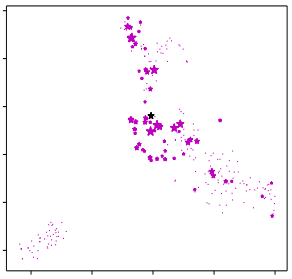
(a) Diagnosis class LP and its symptoms.



(b) Diagnosis class PIG and its symptoms.



(c) Diagnosis class mROU and its symptoms.



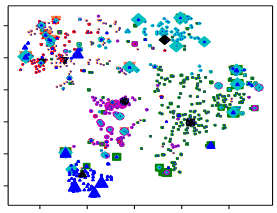
(d) Diagnosis class RAU and its symptoms.

Figure 3. Visualizations of the four diagnoses with their associated symptoms in the Oral Medicine dataset.

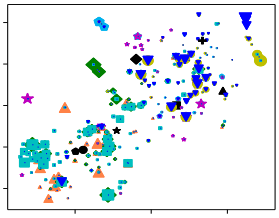
We use the Oral Medicine dataset to illustrate the visualization. There are 425 samples and 253 symptoms in this dataset, and the four diagnoses are lichen planus (LP), plaque-induced gingivitis (PIG), minor recurrent oral ulcers (mROU), and recurrent aphthous ulcer (RAU). Firstly, the TF-IDF value for each symptom in different diagnoses is calculated, the value indicating how important the symptom is to the diagnosis. Secondly, symptoms in different diagnoses are visualized by points in 2D space with different colors and shapes respectively. And thirdly, each symptom point is sized in proportional to its TF-IDF value. Figure 2 shows an overall visualization of the diagnoses and symptoms of the Oral Medicine dataset, and the zoomed-in effects of four diagnoses and their associated symptoms are presented in Figure 3. One can see that 1) each diagnosis is positioned by its related symptoms gathering around; 2) the greater the TF-IDF value of a symptom, the closer it is to its related diagnosis, and the farther from other diagnoses; and 3) the distance between every two diagnoses is different. Observations 1) and 2) indicate that the symptom and diagnosis embeddings generated by our approach learn the correlation, i.e., the closer the symptom is to the diagnosis, the more matters the symptom to the diagnosis. For 3), diagnoses mROU and RAU are close in the space, while PIG is far from the other three classes. To clarify, we sort the related symptoms according to their TF-IDF values, and then count the ratio of the top 10, 20, 30, 40 symptoms that every two diagnoses share. As shown in Table 6, mROU and RAU share the most symptoms, as 9 out of their top 10 symptoms are the same. Moreover, LP and mROU (RAU) partly share symptoms, whereas the symptoms in PIG and other three diagnoses are mostly different. This says that mROU and RAU are "similar", and PIG is "dissimilar" to other diagnoses, in accordance with the fact that mROU and RAU are different levels of a same disease, and LP, mROU and RAU occur primarily in oral mucosa whereas PIG in gums.

Table 6. The ratio of the top 10, 20, 30, 40 symptoms that every two diagnoses share in the Oral Medicine dataset.

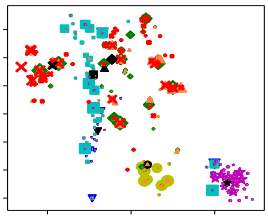
	LP, PIG	LP, mROU	LP, RAU	PIG, mROU	PIG, RAU	mROU, RAU
Top10	0	0.3	0.2	0	0	0.9
Top20	0.2	0.4	0.4	0.2	0.2	0.7
Top30	0.2	0.4	0.33	0.23	0.23	0.6
Top40	0.15	0.325	0.25	0.175	0.175	0.5



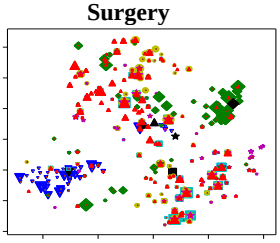
(a) Pediatric Dentistry



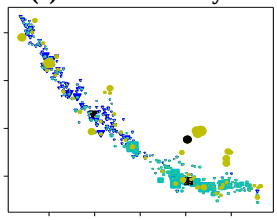
(b) Oral and Maxillofacial Surgery



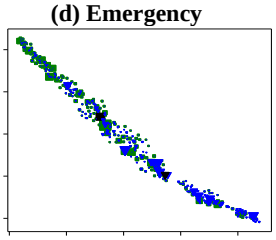
(c) Laser Dentistry



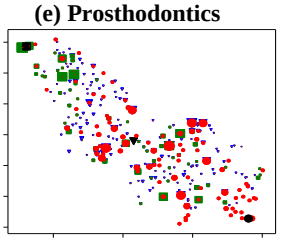
(d) Emergency



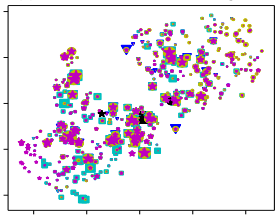
(e) Prosthodontics



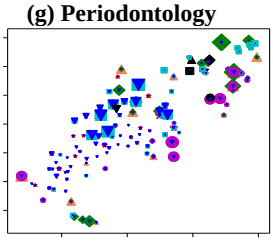
(f) Geriatric Dentistry



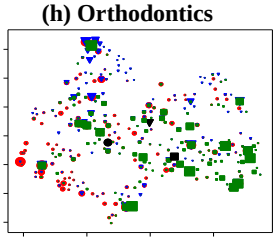
(g) Periodontology



(h) Orthodontics



(i) Implant Dentistry



(j) General Dentistry

Figure 4. Visualizations of symptoms and diagnoses in ten PKUSS datasets.

The visualization of other ten datasets is shown in Figure 4. For the Periodontology, Orthodontics, Implant Dentistry, and General Dentistry dataset, the clustering effect is not obvious, as the symptoms of different diagnoses in these departments are mostly the same, i.e., one symptom often belongs to multiple diagnoses.

5. CONCLUSIONS AND FUTURE WORKS

Large-scale adoptions of health information technology infrastructure in the form of electronic records have accumulated massive amounts of medical data. How to extract information from EMRs and discover the hidden knowledge so as to provide assistance to clinical decision making has been a challenging issue in biomedical informatics. In this paper, we propose to learn symptom and diagnosis embeddings for diagnosis prediction, in which symptoms and diagnoses are represented as vectors in real space. On the real-world EMR datasets provided by PKUSS, evaluations show that our approach outperforms the classic machine learning algorithms in terms of accuracy, precision, recall, F-score, and Cohen's kappa statistic. Visualizing embeddings in 2D space has validated the quality of embeddings generated by our approach in revealing the correlation between symptoms and diagnoses.

The key to our approach's advantage lies in expressing the unique correlations that exist in the medical records. The model we developed can be applied to EMR datasets of other diseases, with adaptations of ways for feature extraction. More broadly, for datasets in other domains, as long as there is correlation inherently between the features and labels, it is worth exploring embedding-based approaches like ours for the purpose of prediction or classification. Seeking alternative applications of symptom and diagnosis vectors generated by our approach will also be our future work.

6. ACKNOWLEDGMENTS

We This work has been supported by the National Key Research and Development Program of China under grant 2016YFB1000902, Projects of Beijing Municipal Science & Technology Commission, the Natural Science Foundation of China grant 61621003, the Knowledge Innovation Program of the Chinese Academy of Sciences (CAS), and Institute of Computing Technology of CAS.

7. REFERENCES

[1] Wu, J., Roy, J., and Stewart, W. F. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*. 48, 6 (June. 2010), 106-113. DOI= https://doi.org/10.1097/MLR.0b013e3181de9e17.

[2] Nguyen, A. N., Lawley, M. J., Hansen, D. P., Bowman, R. V., Clarke, B. E., Duhig, E. E., and Colquist, S. 2010. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*. 17, 4 (July. 2010), 440-445. DOI= https://doi.org/10.1136/jamia.2010.003707.

[3] Sesen, M. B., Kadir, T., Alcantara, R. B., Fox, J., and Brady, M. 2012. Survival prediction and treatment recommendation with Bayesian techniques in lung cancer. In *AMIA Annual Symposium Proceedings* (Chicago, Illinois, USA, November 03 - 07, 2012). 838–847.

[4] Potter, R. 2007. Comparison of Classification Algorithms Applied to Breast Cancer Diagnosis and Prognosis. In *Industrial Conference on Data Mining-Posters and Workshops* (Leipzig, Germany, July 2007). 40-49.

[5] Cruz, J. A., and Wishart, D. S. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*. 2 (January, 2006), 59-77. DOI= https://doi.org/10.1177/117693510600200030.

[6] Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., et al. 2011. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*. 19, 2, (November. 2011), 212-218. DOI= https://doi.org/10.1136/amiajnl-2011-000439.

[7] Mani, S., Chen, Y., Elasy, T., Clayton, W., and Denny, J. 2012. Type 2 diabetes risk forecasting from EMR data using machine learning. In *AMIA annual symposium proceedings* (Chicago, Illinois, USA, November 03 - 07, 2012). 606-615.

[8] Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., and Lee, D. S. 2013. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*. 66, 4 (April 2013), 398-407. DOI= https://doi.org/10.1016/j.jclinepi.2012.11.008.

[9] Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J., and Shah, S. J. 2017. Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. *Journal of cardiovascular translational research*. 10, 3 (June 2017), 313-321. DOI= https://doi.org/10.1007/s12265-017-9752-2.

[10] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. 2003. A neural probabilistic language model. *Journal of machine learning research*. 3 (March 2003), 1137-1155. DOI= http://dx.doi.org/10.1162/153244303322533223.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (Lake Tahoe, Nevada, United States, December 05 – 08, 2013). NIPS'13. Curran Associates Inc, USA, 3111-3119. DOI= http://dl.acm.org/citation.cfm?id=2999792.2999959.

[12] Labutov, I., and Lipson, H. 2013. Re-embedding words. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Sofia, Bulgaria, August 04 – 09, 2013). 489–493.

[13] Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M. 2013. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*. 21, 2 (March. 2014), 221-230. DOI= https://doi.org/10.1136/amiajnl-2013-001935.

[14] Schmiedeskamp, M., Harpe, S., Polk, R., Oinonen, M., and Pakyz, A. 2009. Use of international classification of diseases, ninth revision clinical modification codes and medication use data to identify nosocomial clostridium difficile infection. *Infection Control & Hospital Epidemiology*. 30, 11 (January. 2015), 1070-1076. DOI= https://doi.org/10.1086/606164.

[15] Klompas, M., Haney, G., Church, D., Lazarus, R., Hou, X., and Platt, R. 2008. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *PLOS one*. 3, 7 (July. 2008), e2626. DOI= https://doi.org/10.1371/journal.pone.0002626.

[16] Wright, A., Pang, J., Feblowitz, J. C., et al. 2011. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *Journal of the American Medical Informatics Association*. 18, 6 (November 2011), 859-867. DOI= https://doi.org/10.1136/amiajnl-2011-000121.

[17] Jensen, P. B., Jensen, L. J., and Brunak, S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*. 13, 6 (May. 2012), 395-405. DOI= https://doi.org/10.1038/nrg3208.

[18] Mani, S., Chen, Y., Arlinghaus, L. R., et al. 2011. Early prediction of the response of breast tumors to neoadjuvant chemotherapy using quantitative MRI and machine learning. In *AMIA Annual Symposium Proceedings* (Washington, DC, USA, October 22 – 26, 2011). 868–877.

[19] Kawaler, E., Cobian, A., Peissig, P., Cross, D., Yale, S., and Craven, M. 2012. Learning to predict post-hospitalization VTE risk from EHR data. In *AMIA. Annual Symposium proceedings. AMIA Symposium* (Chicago, Illinois, USA, November 03 - 07, 2012). American Medical Informatics Association. 436–445.

[20] Kim, Y. J., Lee, Y. G., Kim, J. W., Park, J. J., Ryu, B., and Ha, J. W. 2017. Highrisk Prediction from Electronic Medical Records via Deep Attention Networks. *arXiv preprint arXiv:1712.00010*.

[21] Pennington, J., Socher, R., and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (Doha, Qatar, October 25-29, 2014). 1532-1543.

[22] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Baltimore, Maryland, USA, June 23-25, 2014). 1555-1565.

[23] Ganguly, D., Roy, D., Mitra, M., and Jones, G. J. 2015. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (Santiago, Chile, Aug 09 – 13, 2015). SIGIR '15. ACM, New York, NY, USA, 795-798. DOI= http://doi.acm.org/10.1145/2766462.2767780.

[24] Sutskever, I., Vinyals, O., and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (Montreal, Quebec, Canada, December 08-13, 2014). 3104-3112.

[25] Bordes, A., Usunier, N., Chopra, S., and Weston, J. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

[26] Minarro-Giménez, J. A., Marin-Alonso, O., and Samwald, M. 2014. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*. 205 (Jan. 2014), 584-588.

[27] Choi, E., Bahadori, M. T., Searles, E., et al. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA, August 13 – 17, 2016). KDD '16. ACM, New York, NY, USA, 1495--1504. DOI= http://doi.acm.org/10.1145/2939672.2939823.

[28] Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. 2016. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*.

[29] Tran, T., Nguyen, T. D., Phung, D., and Venkatesh, S. 2015. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of biomedical informatics*. 54 (April. 2015), 96-105. DOI= https://doi.org/10.1016/j.jbi.2015.01.012.

[30] Xiaomin, W. 2005. Dictionary-Free chinese words acquisition method based on bigram. *Computer Engineering and Applications*. 41, 22, 177-179.

[31] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. 20, 1 (April. 1960), 37-46. DOI= https://doi.org/10.1177/001316446002000104.

[32] Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*. 33, 1 (Mar. 1977), 159-174.

[33] Maaten, L. V. D., and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*. 9 (Nov. 2008), 2579-2605.