ContextCare: Incorporating Contextual Information Networks to Representation Learning on Medical Forum Data[*]

Sendong Zhao[*] , Meng Jiang[‡], Quan Yuan[‡], Bing Qin[*†], Ting Liu[*], ChengXiang Zhai[‡]
[*] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[‡] Department of Computer Science, University of Illinois at Urbana-Champaign, USA
{sdzhao,bqin,tliu}@ir.hit.edu.cn, {mjiang89,qyuan,czhai}@illinois.edu

Abstract

Online users have generated a large amount of health-related data on medical forums and search engines. However, exploiting these rich data for orienting patient online and assisting medical checkup offline is nontrivial due to the sparseness of existing symptom-disease links, which caused by the natural and chatty expressions of symptoms. In this paper, we propose a novel and general representation learning method CONTEXTCARE for human generated health-related data, which learns latent relationship between symptoms and diseases from the *symptom-disease diagnosis network* for disease prediction, disease category prediction and disease clustering. To alleviate the network sparseness, CONTEXTCARE adopts regularizations from rich contextual information networks including a *symptom co-occurrence network* and a *disease evolution network*. Extensive experiments on medical forum data demonstrate that CONTEXTCARE outperforms the state-of-the-art methods in respects.

1 Introduction

With the prevailing of Web 2.0 applications, an increasing number of individuals are seeking diagnostic aids online. As reported, 33% American adults went online to figure out what medical conditions they might have [Fox and Duggan, 2013] and 5% google searches are healthcare related [Joshi, 2017]. To get online diagnostic aids, individuals can either search symptom related queries or ask questions on medical forums such as PatientsLikeMe and Haodf [Baike, 2017]. In the former scenario, part of users would click returned documents of diseases according to their symptoms, which generates links between diseases and symptoms. In the latter scenario, medical experts or patients with the same symptoms would help to explain what's going on with one's body on medical forums, which generate plenty of symptom-disease links as well.

Can we leverage these large amounts of medical forum data and heath related query logs to orient patient online and assist professional clinical checkup offline? Are these links between textual symptoms and textual diseases good enough to use directly? Unfortunately, symptoms in both queries and medical forum posts are usually informally expressed with narrative language. Especially on medical forums, symptoms and diseases, which are in rounds of QAs, are usually too natural and chatty instead of professional and brief. As a consequence, symptoms with similar literal meanings or medical implications are usually expressed in different narrative ways, leading to very sparse links between symptoms and diseases.

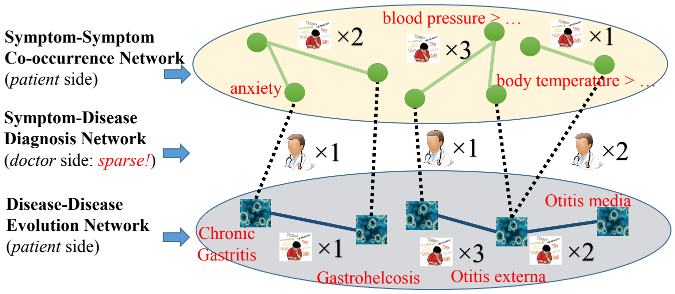

Symptom-Symptom Co-occurrence Network (*patient* side)
Symptom-Disease Diagnosis Network (*doctor* side: *sparse!*)
Disease-Disease Evolution Network (*patient* side)
×2
blood pressure > …
×3
×1
anxiety
body temperature > …
×1
×1
×2
Otitis media
Chronic Gastritis
×1
×3
×2
Gastrohelcosis
Otitis externa

Figure 1: *Three contextual information networks from online medical forum for disease prediction.* Each node of top layer is a symptom description and each node of bottom layer is a disease description. The symptom-disease diagnosis network seriously suffers from the sparseness. However, the symptom co-occurrence network and disease evolution network that carry rich contextual information from the patient side can be used to alleviate the sparsity.

In order to enhance the utility of the rich online medical forum data for disease prediction, as mentioned above, we have to face challenges of the serious issue of sparseness, that are, 1) the number of descriptions of symptoms and diseases is in thousand-level due to the variety of symptoms and diseases and the diversity of natural language expressions for symptoms, 2) the number of symptom-disease links is *relatively small*. For example, the sample on our data indicates that the density of the symptom-disease network is only 0.07%.

To address the sparsity problem, we propose a new idea of using the rich contextual information of diseases and symptoms to bridge the gap between disease candidates and symptoms, and detach it from the specific way of implementing the idea using network embedding. Specifically, we create a two-layer network which consists of a bipartite diagnosis network and the following two homogeneous networks. (see Figure 1).

[*] This work was supported by 863 Program via grant 2015AA015407 and NSFC via grants 61632011 and 61370164.

[†] Corresponding author

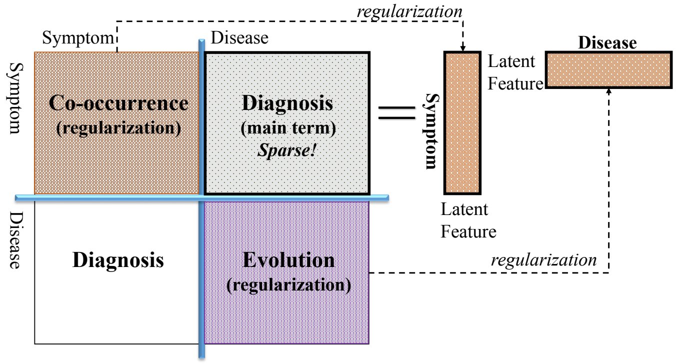


Figure 2: *Our* ContextCare *method that learns effective latent features of symptoms and diseases.* We adopt the diagnosis network to learn the representations as the main term, and we use the symptom co-occurrence network and disease evolution network to regularize the representations to alleviate the sparseness of the symptom-disease diagnosis network.

- *Symptom co-occurrence network.* A patient may have multiple symptoms. We spot co-occurrence of the symptoms, for instance, *headache* and *high blood pressure*.
- *Disease evolution network.* A patient may report the disease co-occurrence or evolution that they know from the experts online or clinicians offline. We observe the evolution of the diseases, for example, *tuberculosis* increases the risk of *lung cancer*.

We believe that these two networks can highly mitigate the sparsity of bipartite diagnosis network. To make use of those two networks as contextual information for accurate disease prediction, we propose a novel method, ContextCare, to *integrate the three networks of different aspects but rich information*. The basic ideas of ContextCare are 1) the linked disease and symptom in the bipartite diagnosis network should be closer in a certain way; 2) frequently co-occurred symptoms should be closer in representation; 3) evolutionary diseases should be closer in representation. Compacting symptoms in 2) and diseases in 3) both benefit alleviating the sparsity of the bipartite diagnosis network. By encoding all these aspects, our ContextCare utilizes information from the three networks to facilitate the predictive model. ContextCare takes the *sparse* symptom-disease diagnosis network as the main term in the objective function and adopts the *contextual information* including the symptom co-occurrence and disease evolution as regularization terms (see Figure 2).

It is worthwhile to highlight our contributions as follows.

- *Important problem and new idea.* We leverage the rich medical QA posts from online medical forums for disease prediction and patient guidance with informally expressed symptoms, which could benefit diagnostic aids online and offline. To deal with the sparseness of diagnosis network, we propose ContextCare to integrate the diagnosis network, symptom co-occurrence network and disease evolution network for learning the latent representations of symptoms and diseases.
- *Effectiveness in real data.* Experiments on medical forum data demonstrate that ContextCare outperforms the state-of-the-art methods in disease prediction on thousands of classes, with a 23.1% relative improvement.

2 Problem Statement

We define the networks get from online medical forums and the problem studied in this work.

Definition 1 (Symptom-Disease Diagnosis Network) *The network, denoted as $G^{SD} = (S \cup D, E)$, is a bipartite network that captures the relation between patient's symptoms S^p and the corresponding disease d. S is the set of symptoms and $S^p \in S$. D is the set of diseases and $d \in D$. $E \subset S \times D$ is the set of symptom-disease links extracted from the diagnosis by doctors and experts.*

The symptom-disease diagnosis network which consists of relationships between symptoms and disease is the essential resource for disease prediction. However, the extremely sparsity of paths in this bipartite network is an absolute disaster for any prediction models. In order to alleviate the sparseness of the symptom-disease diagnosis network, we introduce the following two networks.

Definition 2 (Symptom Co-occurrence Network) *This network is denoted by $G^{SS} = (S, E^{co})$, where S is the set of symptoms and $E^{co} \subset S \times S$ is the set of symptom-symptom co-occurrence links whose frequency is beyond a threshold τ. Each $(s_i, s_j) \in E^{co}$ indicates that symptom s_i and s_j co-occurred above τ times.*

Definition 3 (Disease Evolution Network) *This network is denoted by $G^{DD} = (D, E^{ev})$, where D is a set of diseases, and $E^{ev} \subset D \times D$ is the set of disease-disease evolution links. Each $(d_m, d_n) \in E^{ev}$ indicates that disease d_m and d_n are evolutionary diseases.*

Given these definitions, we can formally formulate the problem studied in this work.

Problem 1 (Representation Learning) *Given the sparse network G^{SD} and rich contextual information from networks G^{SS} and G^{DD}, find the latent representation $\mathbf{s}$ of symptom s and $\mathbf{d}$ of disease d.*

It is a fundamental problem towards disease prediction, since the prediction performance highly depends on the quality of the latent features we learn.

3 The ContextCare Method

We introduce details of the proposed representation learning method (ContextCare) in this section. This model operates in bipartite symptom-disease diagnosis network learning and makes use of symptom-symptom co-occurrence network and disease-disease evolution network as the constraints of the bipartite network learning.

3.1 Bipartite Symptoms-Disease Network Learning

Given the bipartite symptom-disease network G^{SD}, it is quite appealing to learn latent representations of vertices via modeling symptom-disease paths. The underlying assumption of the embedding method is that the disease can be represented with its relationships with symptoms and vice versa, which derives from [Bordes *et al.*, 2013]. It is implemented by taking the symptom-disease relationship as a transition $\mathbf{t}$. Specifically, We define a simple energy function $f(S^p, d)$ on each (S^p, d) as follows.

$$f(S^p, d) = \left\| \frac{1}{|S^p|} \sum_{s \in S^p} \mathbf{s} + \mathbf{t} - \mathbf{d} \right\|_1 \tag{1}$$

We are using an ℓ_1 norm in the latent space, but other metrics could be used as well. We take the symptom-disease as a transition vector $\mathbf{t}$. True symptoms-disease pairs (S^p, d) are assumed to have low energies in the energy function. For each true symptoms-disease pair, we generate a contrast via sampling a negative pair with normal distribution and make any deviation from (S^p, d) as costly as possible. It is worth to mention that we also try another energy function, which considers the transition between symptoms and disease as multiplying a matrix, as a baseline model.

To learn the symptom embeddings $\{\mathbf{s}\}$, the disease embeddings $\{\mathbf{d}\}$ and the transition vector $\mathbf{t}$, we consider a ranking criterion. Intuitively, given a true pair (S^p, d), if the disease d is missing, we would like the model to be able to predict the correct disease. The objective of training is to learn the energy function f so that it can successfully rank the true pair (S^p, d) to be preceded all other possible pairs. Therefore, we define a loss to formalize this intuition:

$$L(G^{SD}) = \sum_{\substack{(S^p,d) \in \mathcal{M}^+ \\ (S^{p'},d') \in \mathcal{M}^-}} [\gamma + f(S^p, d) - f(S^{p'}, d^{'})]_+ \tag{2}$$

where $\mathcal{M}^+$ is the set of true (S^p, d) pairs found in the online medical QA pairs from medical forums, $\mathcal{M}^-$ contains corrupted pairs constructed by replacing the symptom set or the disease in the true (S^p, d), $\gamma > 0$ is a margin separating true symptom-disease pairs and corrupted pairs, and $[x]_+ = max(0, x)$ denotes the positive part of x.

3.2 Regularization of Symptom Co-occurrence

Beyond learning the bipartite symptom-disease diagnosis network, ContextCare also exploits symptom-symptom co-occurrence network G^{SS} as the constraints to alleviate the sparseness of the bipartite network for more accurate representations of symptoms, thus benefits disease prediction. Hence, we define the ℓ_1 penalty over frequently co-occurred symptoms, $\|\mathbf{s}_i - \mathbf{s}_j\|_1$, defines the symptom-symptom regularization. It incentivizes the vector representations of frequently co-occurred symptoms to be close, the larger w_{ij}, the greater penalty.

$$R_1(G^{SS}) = \sum_{(s_i,s_j) \in E^{co}} w_{ij} \|\mathbf{s}_i - \mathbf{s}_j\|_1 \tag{3}$$

where $w_{ij} = \frac{|\Gamma(s_i) \cap \Gamma(s_j)|}{|\Gamma(s_i) \cup \Gamma(s_j)|}$ is the Jaccard similarity between symptoms s_i and s_j, $\Gamma(s_i)$ and $\Gamma(s_j)$ denote the set of neighbors of symptom s_i and symptom s_j in the symptom-symptoms co-occurrence network with frequency respectively. Note that we consider the frequency of the edge between neighbors when counting $\Gamma(s)$, i.e., we take a symptom s_j as a neighbor of s_i when the co-occurrence frequency of s_i and s_j is beyond a threshold τ.

By minimizing $R_1(G^{SS})$, the representations of symptoms which encode contextual information lead to close representations of symptoms with similar meanings or implications.

3.3 Regularization of Disease Evolution

Our model considers the disease evolution as a very important contextual information of diseases for alleviating the sparseness of bipartite symptom-disease diagnosis network G^{SD}. In order to make full use of the explicit and implicit information involved in disease evolution, we define the ℓ_2 penalty over the difference on evolutionary diseases, $\|\mathbf{d}_m - \mathbf{d}_n\|_2$, defines the disease evolution regularization. It incentivizes the vector representations of diseases which are different stages of the same patient to be close, the larger v_{mn}, the greater penalty.

$$R_2(G^{DD}) = \sum_{(d_n,d_m) \in E^{ev}} v_{mn} \|\mathbf{d}_n - \mathbf{d}_m\|_2 \tag{4}$$

where $v_{mn} = \frac{|\Gamma(d_n) \cap \Gamma(d_m)|}{|\Gamma(d_n) \cup \Gamma(d_m)|}$ is again the the Jaccard similarity, $\Gamma(d_n)$ and $\Gamma(d_m)$ denote the set of neighbors of disease d_n and disease d_m in the disease evolution network respectively. Note that we make use of ℓ_2 on evolutionary diseases rather than ℓ_1 like the penalty on co-occurred symptoms. The reason is 1) we want treat these two kinds of relations differently and 2) put more emphasis on disease evolution because of more significant effect on compacting diseases.

The evolution of diseases is used as the constraint for learning more accurate representations of diseases, thus benefits disease prediction. Disease evolution is the knowledge of the progression and change directions of disease. By encoding disease progression and direction in them, disease representations are given the potential to predict different variations of the diseases and long-term diseases given initial symptoms of patients. The constraint of disease evolution is also good for compacting symptom representations because approximating those evolutionary diseases will indirectly lead to approximate those symptoms which should be compacted. Therefore, by minimizing $R_2(G^{DD})$, the representations of disease which encode contextual information lead to close representations of diseases which are similar in the medical sense and also benefit condensing similar symptoms.

3.4 Latent Representation Learning

Towards the goal of learning representations of symptoms and diseases, the final objective loss of our model combines symptom-disease bridging loss, symptom-symptom co-occurrence loss and disease-disease evolution loss 2, 3 and 4, which is minimized as follows:

$$\min_{\{\mathbf{s}\},\{\mathbf{d}\},\mathbf{t}} L(G^{SD}) + \alpha R_1(G^{SS}) + \beta R_2(G^{DD}) \tag{5}$$

where $\alpha > 0$ and $\beta > 0$ are both parameters which weight the regularizations. This proposed model directly bridges symptoms and diseases. Besides, it extends to incorporate symptom-symptom co-occurrence and disease-disease evolution as constraints for learning symptom and disease representations onto a latent space. This model can capture long-term influence from one disease to another by strengthening evolution between diseases. The frequently co-occurred symptoms are also grouped close to each other. All these two aspects benefit bridging symptoms with diseases. The optimization in Eq. (5) favors lower energies for true symptoms-disease pairs than corrupted pairs, and is thus a natural implementation of the intended criterion. The optimization is carried out by stochastic gradient descent in mini-batch mode. We enforce the constraints that the embedding of each node $\|\mathbf{s}\| = 1$ and $\|\mathbf{d}\| = 1$ to avoid overfitting. The computational complexity approximates to $\mathcal{O}(TNK)$, where N is the number of iterations, N the number of cases, K the number of symptoms in each case.

4 Experiments

4.1 Dataset

We conduct experiments on a real Chinese medical QA narrative data from an online medical forum, i.e., Haodf [Baike, 2017]. From the online medical QA posts of each patient, we have a set of symptoms in narratives, diagnosed diseases given by online experts. Except for the historical diseases proposed by the patient initially, the evolution of diseases can be easily obtained from multiple rounds of QA in each post. In our dataset, there are 17,803 medical QA posts with 18,899 symptoms and 1,066 diseases in total. We split all medical QA posts into a training/validation/test set randomly, with the ratio of 8:1:1. The first part is for training model, the second for hyper-parameter tuning, and the third for evaluation. According to ICD-10, the 1,066 diseases are assigned to 9 categories, which is to evaluate disease category prediction and clustering. In order to get symptom phrases, we leverage a medical dictionary and a unified pre-processing platform for Chinese [Che *et al.*, 2010] to recognize symptom related noun phrases and verb phrases.

4.2 Baseline Methods

Classification Models. Disease prediction by taking symptoms as binary features is a typical muti-label classification. Here, we take SVM [Chang and Lin, 2011], Decision Tree [Breiman *et al.*, 1984] and MaxEnt [Berger *et al.*, 1996] as baselines. However, these baselines just take the one-hot feature of symptoms thus suffer from the sparsity of symptoms.

Topic Modeling Method. The well-established Latent Dirichlet Allocation (LDA) model [Blei *et al.*, 2003] has been applied to get phenotypes for diagnosis code prediction [Perotte *et al.*, 2011]. Therefore, we take LDA [Blei *et al.*, 2003] to learn the symptom distribution and the disease distribution for disease prediction by regarding symptoms as words and diseases as documents.

Link Prediction Models. Researchers have proposed a few methods that can predict the links between symptoms and disease with the diagnosis network G^{SD}. P-PageRank prefers those popular diseases and therefore tends to take popular diseases as predictions. SimRank [Jeh and Widom, 2002] links a disease d and a symptom s by considering the links between the neighbor diseases of d, i.e., $\Gamma(d)$ and neighbor symptoms of s, i.e., $\Gamma(s)$. However, it is inflexible to link those diseases and symptoms with few neighbors and suffer from the sparseness of symptom-disease links. HeteSim [Shi *et al.*, 2014] learns links between diseases and symptoms by walking on the G^{SD} through the path [disease→disease→symptom→symptom]. However, it is insufficient to deal with latent relations thus suffers from the sparseness of symptom-disease diagnosis network. These methods are both designed for one-to-one link prediction instead of one-to-many relationship prediction as our task. In order to fit our many-to-one prediction task, we extend these methods to combine one-to-one links by voting.

Network Embedding Methods. Representing diseases and symptoms into latent spaces is a more flexible way to bridge the gap of disease and symptoms. LSHM [Jacob *et al.*, 2014] learns a classification function meanwhile takes into account that neighbor symptoms or diseases should be close. The idea behind is that two diseases or symptoms which are connected in G^{SD} will tend to share similar representations. However, it deals with evolutionary diseases and co-occurred symptoms in the same manner and treats every disease evolution and symptom co-occurrence with the same importance, which might limit the predictive performance. The LSHM is also designed for dealing with one-to-one similarity computation rather than many-to-one. Therefore, we leverage voting to meet with the many-to-one mode. ContextCare$_\times$ is a variation of our CONTEXTCARE by taking the symptom-disease relationship as a transition matrix $\mathbf{T}$ in ContextCare$_\times$ rather than a transition vector $\mathbf{t}$.

We considered the work on QA embedding by Bordes et al. [Bordes *et al.*, 2014], however, mapping entity names in questions to subgraphs of KBs is undoable in our narrative-data case. In their design, the question and the subgraph must share at least one entity. In this way, it is reasonable that questions and answers share the same embedding matrix. However, in our dataset, symptoms and the corresponding disease do not share even a word in most cases.

4.3 Implementation Details

For baselines, we tune hyper-parameters for best results. For our methods, we select the learning rate λ for stochastic gradient descent among $\{0.001, 0.01, 0.1\}$, the margin γ among $\{1, 2, 10\}$, and the latent dimension k among $\{10, 20, 30, 40, 50\}$ on the validation set. Optimal configurations for CONTEXTCARE are $k = 40$, $\lambda = 0.01$, $\gamma = 1$, $\alpha = 0.35$, $\beta = 0.55$. Training is limited to at most 1,000 epochs over the training set. The best models are selected by early stopping on the validation sets. We conducted multiple rounds of cross-validation to avoid the issue of time sensitivity.

4.4 Evaluation Criterion

For disease prediction and disease category prediction, we use accuracy to evaluate results. For those methods which can generate ranking lists, we use Precision@5 and Precision@10 to check the proportion of correct diseases located in the top 5 and 10. For each given set of symptoms, our model gives a ranking list of diseases. The prediction accuracy, that equals as precision@1, along with other precision@k metrics can well evaluate the performance. Recall (and f-score) is not sensitive in the task: one disease is often connected to only one symptom; so, if precision@k increases, recall increases as well. For disease category prediction, we re-train the classification models for 9 classes according to ICD-10. Rand index (RI) [Rand, 1971] is taken to evaluate the performance of

Method	Accuracy				Precision@5	Precision@10
	Diagnosis Network	+*Co*	+*Ev*	+*Co,Ev*		
SVM (linear)	16.02	-	-	-	-	-
SVM (RBF)	16.79	-	-	-	-	-
Decision Tree (C4.5)	17.31	-	-	-	-	-
MaxEnt	18.98	-	-	-	-	-
LDA	14.73	-	-	-	23.46	35.21
P-PageRank	17.22	19.71	17.25	19.74	43.17	62.16
SimRank	19.36	21.97	19.38	21.98	46.52	64.33
HeteSim	20.62	23.03	20.69	23.32	55.31	70.48
LSHM	21.38	25.87	22.55	25.77	65.74	82.45
ContextCare$_\times$	22.35	28.09	24.74	30.66	69.38	85.76
ContextCare	**23.57**	**30.32**	**27.26**	**31.73**	**73.21**	**87.36**

Table 1: Comparison with baseline methods in accuracy and precision@N (%).

disease clustering.

4.5 Disease Prediction

Results and Analysis. Table 1 shows the accuracy of the baseline methods as well as our ContextCare on disease prediction. Each row of Table 1 represents a model for disease prediction. Each column stands for a type of resource combination used to predict disease. For example, the column "*Diagnosis Network*" denotes the use of bipartite symptom-disease diagnosis network including a large number of symptom-disease paths and the column "+*Co*" and "+*Ev*" denotes that symptom co-occurrence and disease evolution are integrated separately besides symptom-disease paths.

From the second column of Table 1, we can see that network embedding methods outperform classification baseline models, LDA and link prediction models. Symptoms are represented as one-hot features which are extremely sparse in classification models, which greatly limits the performance of disease prediction. LDA is deficient for sparseness. Link prediction models utilize the explicit paths between symptoms and diseases, which is inflexible to link diseases and symptoms thus cannot help to solve the sparsity problem. Network embedding methods utilize symptom-disease relations to represent symptoms and diseases onto latent spaces. They are more flexible to capture latent links between symptoms and diseases thus benefit disease prediction.

From the last 6 rows of Table 1, we can see that taking into account of symptom co-occurrence and disease evolution separately or jointly can consistently improve the performance of disease prediction for both link prediction models and network embedding models. However, these resources produce greater improvement on disease prediction in our ContextCare. ContextCare increases accuracy by 6.75% from using symptom co-occurrence network to regularize the representations of symptoms, which indicates that compacting representations of symptoms benefits disease prediction. ContextCare increases accuracy by 3.68% from using disease evolution network to regularize the representations of diseases, which indicates that compacting representations of diseases improves disease prediction. When taking these two regularization terms, our ContextCare increases accuracy by 8.16% in total. Essentially, compacting symptoms and diseases are both ways to alleviate the sparsity of symptom-disease links. The results demonstrate that ContextCare successfully alleviate the issue of sparseness and improve the performance of prediction by integrating networks.

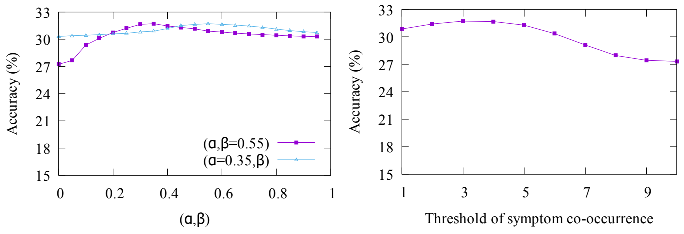
The above observation shows that: 1) our new idea of using additional homogeneous networks as contexts works well for all the ways to make predictions; 2) the particular embedding method ContextCare works better than other methods, showing that embedding-based approach is better because they are flexible to capture latent links in the diagnosis network; 3) our ContextCare is more effective in alleviating the sparseness of the diagnosis network by taking symptom co-occurrence network and disease evolution network to regularize the representations of diseases and symptoms, thus obtain the best results.

Effect of α and β in ContextCare. We tune the hyperparameter α and β of ContextCare on the validation set. We randomly assign a value from (0:0.5:1) to α and got the β which achieve the best accuracy, then conduct the same on β. Then we proceed to iteratively find best α and β. The Figure 3(a) shows the effect of α and β to the performance.

Effect of threshold τ in $R_1(G^{SS})$. We investigate how the threshold of the frequency of the symptom-symptom co-occurrence affects the performance of ContextCare for disease prediction. We vary the threshold value of the symptom-symptom co-occurrence from 1 to 10, increased by 1. Results of the effect of threshold on disease prediction on our validation set are given in Figure 3(b). We can see that the best threshold configuration is 3 and greater or less than 3 will cause a worse performance.

4.6 Disease Category Prediction

Given symptoms, predicting specific disease from 1,066 diseases is a difficult task. The reason is that similar diseases tend to share similar symptoms, such as "gastric ulcer" and "gastritis". However, we will not fall into the unconsidered differences between "gastric ulcer" and "gastritis" in symptoms if we just predict the category of disease instead of specific disease because they are both diseases of the "digestive system" according to ICD-10. Predicting the category of disease is valuable because of the following two scenarios. First, it is necessary to early determine the department that a patient should make an appointment with. Second, doctors and experts can provide quick consulting if they have the correct category of the potential disease in advance. Figure 4 shows the results that LSHM is considerable for application but still not as good as our ContextCare which stands out from all baseline methods in disease category prediction.



(a) (b)

Figure 3: (a) Effect of α and β in CONTEXTCARE. (b) Effect of threshold τ in $R_1(G^{SS})$

4.7 Disease Clustering

This experiment is to test the effectiveness of the learned disease embeddings in clustering similar diseases. We utilize classical k-means [Hartigan and Wong, 1979] with random initialization to cluster the 1,066 diseases and evaluate the clustering performance with Rand index (RI) [Rand, 1971]. Note that we follow the disease categories in ICD-10, i.e., 1,066 diseases in our dataset are assigned to 9 categories. These labeled diseases are taken as the ground truth to verify whether diseases in the same category of ICD-10 are represented to be much closer than diseases belonging to different categories. From the experiment, LSHM gets 79.23%, ContextCare$_\times$ gets 88.09% and CONTEXTCARE gets 89.14%. It is clear that our proposed CONTEXTCARE performs the best. The effectiveness of our CONTEXTCARE in clustering diseases is a very important reason that we achieve the better disease prediction and disease category prediction. The reason is that CONTEXTCARE makes the representations of diseases with few supports of symptoms much closer to its similar diseases with rich supports of symptoms, and consequently improves the performance of predicting disease with few supports of symptoms.

5 Related Work

Disease prediction. There are several studies working on disease prediction based on rich electronic health record (EHR) with different methods, such as [Sun *et al.*, 2012; Wang *et al.*, 2014a; Choi *et al.*, 2015; Wang *et al.*, 2014b]. Disease prediction based on EHR requires relatively long records of a patient for generating good results. However, EHR is expensive for both patients and researchers. More importantly, there are large of amount of users want to check their medical condition with narrative symptoms online. Our research emphasizes on mitigating the sparsity between symptoms and diseases thus obtains better results on disease prediction with narrative symptoms only.

EHR-based phenotyping is a process to map raw EHR data into meaningful medical concepts, learning medically relevant characteristics of the data [Denny, 2012], and is important for supporting genome-wide association studies [Hripcsak and Albers, 2013]. Phenotyping can be viewed as a form of dimensionality reduction, where each phenotype forms a latent space. The well-established Latent Dirichlet Allocation (LDA) model [Blei *et al.*, 2003] has been applied to get phenotypes for diagnosis code prediction [Perotte *et al.*, 2011] and disease progression modeling [Wang *et al.*, 2014b]. However, our goal is to predict disease with informal symptoms for orienting patient who might want to get guidance from online medical QA systems, search engines and online medical forums. Therefore, the big challenge is the sparsity which comes from textual symptoms with informal expressions. Phenotyping uses clean structured EHR rather than the noisy online QA data from medical forums, thus much less sparsity than that in our task. LDA which essentially leverage the high frequent co-occurrence in data can hardly deal with the serious sparseness of narrative symptoms.

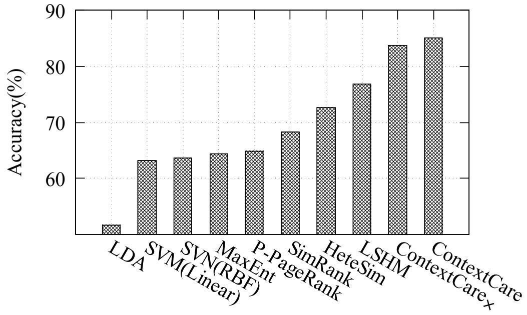


Figure 4: Comparison with baseline methods on disease category prediction in accuracy (%).

Network mining and analysis is also a related topic as we can formulate our problem as link prediction in disease-symptom networks. Most of the existing studies [Adamic and Adar, 2003; Jeh and Widom, 2003; Sun *et al.*, 2011; Shi *et al.*, 2014] predict links directly rely on existed links in networks, and these methods are not effective for our problem because disease-symptom networks are very sparse. As shown in experiments, our method outperforms state-of-the-art method in this line. In recent years, a number of embedding methods have been proposed, which learn distributed representations of nodes and can better handle sparsity problem. However, most of these studies focus on homogeneous networks, and they do not exploit contextual links among homogeneous nodes to alleviate sparsity problem, such as [Perozzi *et al.*, 2014], [Bordes *et al.*, 2013], [Tang *et al.*, 2015] and [Zhao *et al.*, 2017]. LSHM [Jacob *et al.*, 2014] is designed for heterogeneous networks, but it doesn't focus on addressing sparsity problem, and our method outperforms L-SHM significantly on disease prediction (see Table 1).

6 Conclusions

In order to leverage the large amounts of medical forum data and heath related query logs to orient patient online and assist professional clinical checkup offline, we proposed a general new idea of using the symptom-symptom and disease-disease networks to bridge the gap between disease and symptoms and then detach it from the specific way of implementing the idea using network embedding. Specifically, on the one hand, we treat the contextual information of symptom as a constraint on representations of symptoms. On the other hand, we treat the contextual information of disease as a constraint on representations of disease. By encoding the above contextual information, our CONTEXTCARE effectively alleviate the sparseness of the symptom-disease diagnosis network.

References

[Adamic and Adar, 2003] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[Baike, 2017] Baidu Baike. Haodf (haodf.com)— baike, the free encyclopedia in chinese, 2017. [Online; accessed 02-February-2017].

[Berger *et al.*, 1996] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[Bordes *et al.*, 2014] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *Proceedings of the EMNLP*, pages 615–620. ACL, October 2014.

[Breiman *et al.*, 1984] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Che *et al.*, 2010] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd COLING*, pages 13–16. ACL, 2010.

[Choi *et al.*, 2015] Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. Constructing disease network and temporal progression model via context-sensitive hawkes process. In *Proceedings of the IEEE ICDM*, pages 721–726. IEEE, 2015.

[Denny, 2012] Joshua C Denny. Mining electronic health records in the genomics era. *PLoS Comput Biol*, 8(12):e1002823, 2012.

[Fox and Duggan, 2013] Susannah Fox and Maeve Duggan. Health online 2013. *Health*, pages 1–55, 2013.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[Hripcsak and Albers, 2013] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[Jacob *et al.*, 2014] Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM WSDM*, pages 373–382. ACM, 2014.

[Jeh and Widom, 2002] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the 8th ACM KDD*, pages 538–543. ACM, 2002.

[Jeh and Widom, 2003] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th ACM WWW*, pages 271–279. ACM, 2003.

[Joshi, 2017] Anjali Joshi. Primum non nocere: Healthcare in the digital age. In *Proceedings of the 10th ACM WSDM*, pages 579–579. ACM, 2017.

[Perotte *et al.*, 2011] Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. Hierarchically supervised latent dirichlet allocation. In *NIPS*, pages 2609–2617, 2011.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM KDD*, pages 701–710. ACM, 2014.

[Rand, 1971] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[Shi *et al.*, 2014] Chuan Shi, Xiangnan Kong, Yue Huang, Philip S Yu, and Bin Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE actions on Knowledge and Data Engineering*, 26(10):2479–2492, 2014.

[Sun *et al.*, 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the 37th VLDB*, 2011.

[Sun *et al.*, 2012] Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Zahra Daar, and Walter F Stewart. Combining knowledge and data driven insights for identifying risk factors using electronic health records. 2012:901–10, 2012.

[Tang *et al.*, 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th ACM WWW*, pages 1067–1077. ACM, 2015.

[Wang *et al.*, 2014a] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM KDD*, pages 145–154. ACM, 2014.

[Wang *et al.*, 2014b] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM KDD*, pages 85–94. ACM, 2014.

[Zhao *et al.*, 2017] Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the 10th ACM WSDM*, pages 335–344. ACM, 2017.