Automated Medical Diagnosis by Ranking Clusters Across the Symptom-Disease Network

Jingchao Ni[1], Hongliang Fei[2], Wei Fan[2] and Xiang Zhang[1]
[1]College of Information Sciences and Technology, Pennsylvania State University, [2]Baidu Research Big Data Lab
[1]{jzn47, xzhang}@ist.psu.edu, [2]{hongliangfei, fanwei03}@baidu.com

***Abstract*—The rapid growth of medical recording data has increased the demand for automated analysis. An important problem in recent medical research is automated medical diagnosis, which is to infer likely diseases for the observed symptoms. Existing approaches typically perform the inference on a sparse bipartite graph with two sets of nodes representing diseases and symptoms, respectively. By using this graph, existing methods basically assume no direct dependency exists between diseases (or symptoms), which may not be true in practice. To address this limitation, we propose to integrate two domain networks encoding similarities between diseases and those between symptoms to avoid information loss as well as to alleviate the sparsity problem of the bipartite graph. Another limitation of the existing methods is that they usually output a ranked list of diseases mixed from very different etiologies which greatly limits their practical usefulness. An ideal method should allow a clustered structure in the disease ranking list so that both similar and different diseases can be easily identified. Therefore, we formulate automated diagnosis as a novel *cross-domain cluster ranking* problem, which identifies and ranks the disease clusters simultaneously in the symptom-disease network. Our formulation employs a joint learning scheme in which the dual procedures of cluster finding and cluster ranking are coupled and mutually reinforced. Experimental results on real-world datasets demonstrate the effectiveness of our method.**

I. Introduction

Recent advances in medical research have generated rich data about human symptoms and diseases [1], which has offered great opportunities for developing automated diagnosis methods to help people do self-prognosis to understand their health conditions. The general goal of automated medical diagnosis is to identify possible diseases for a given set of symptoms. Most existing methods are based on the Quick Medical Reference (QMR) graphical model [2], [3]. In this model, symptoms and diseases are regarded as two sets of nodes forming a bipartite graph. Each pair of associated symptom and disease are connected by an edge, with a weight indicating the correlation level between the symptom and disease [1]. The conditional probabilities of the diseases for the given symptoms can be inferred based on this bipartite graph. Then, the top ranked diseases will be examined by heath experts in further details. Despite their success, the existing methods are limited by the following two critical problems.

The first limitation is that by using the bipartite graph, existing methods assume diseases (and symptoms) are independent with each other, which is not precise in real-world applications. More practically, many diseases are related one another [1], such as disease "bronchitis" and disease "asthma". Sometimes, certain diseases can even cause other diseases to happen [2]. On the other side, related symptoms frequently occur together. For example, symptom "cough" often occur together with symptoms "expectoration" and "sore throat". Therefore, ignoring such relationships may result in severe information loss. Moreover, the current known associations between diseases and symptoms are far from being complete [1]. Thus using only the sparse associations in the bipartite graph for diagnosis is insufficient to obtain reliable results.

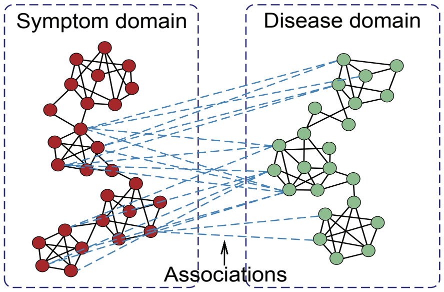


Fig. 1. An example of the symptom-disease network.

Another limitation of existing methods is that it is often difficult to interpret their outputs. For example, imagine a patient having symptoms "cough", "expectoration" and "sore throat", which can be caused by many kinds of diseases, such as "cold" and "asthma". Although they are relevant, these diseases are very different in their underlying etiologies [4]. Existing methods only account for the relevances of diseases for given symptoms, and often mix different diseases in the ranking list, such as {"cold", "asthma", "influenza", "bronchitis", ...}. Such a mixed ranking list can be confusing and even risky in practice since it may mislead the subsequent therapies, considering that etiologically different diseases may require different treatments [5].

To address the first limitation, we propose to integrate two *domain networks* that represent the relationships between symptoms and those between diseases. These domain networks are made available by the recent advancements in network medicine [4], [1]. Fig. 1 illustrates a symptom similarity network (left) and a disease similarity network (right). Each edge in the domain networks represents how similar the two connected symptoms or diseases are. We refer to the bipartite graph connecting symptoms and diseases as the cross-domain *association network*. The rich information encoded in the domain networks can supplement the association network and alleviate the sparsity problem of the existing methods. To our best knowledge, this is the first work to involve domain networks for automated medical diagnosis.

To address the second limitation, we propose to output a list of ranked disease clusters instead of individual diseases. For example, the list {{1st class: "cold", "influenza"}, {2nd class: "bronchitis", "asthma"}, ...}, is more desirable than a mixed list of individual diseases. In our approach, similar diseases are grouped together, and the clusters are ranked by their relevance to the symptoms. This allows easy identification of disease categories and increases the interpretability of the results .

Motivated by the above discussions, we formulate automated diagnosis as the problem of inferring a probability distribution for the disease clusters given a symptom cluster based on the symptom-disease network. Note that symptoms are also considered at the cluster-level, since related symptoms often occur together. We refer to this problem as the *cross-network cluster ranking* problem.

The duality between the cluster finding and cluster ranking proposes new challenges that cannot be readily handled by the existing network analysis methods. For example, the network clustering algorithms [6], [7], [8] cannot handle the cluster ranking problem. The co-clustering algorithms [9], [10], [11] do not fully exploit the domain network structures. Therefore, we propose a novel approach, CROSSCR, to simultaneously cluster domain networks and infer the conditional probabilities of clusters in one domain for clusters in another. By leveraging the duality between clustering networks and ranking clusters across domains, both procedures are mutually reinforced in the learning process. Experimental results on real-life datasets demonstrate the effectiveness of our method.

II. PROBLEM DEFINITION

We present our method in a general form so that it is applicable to any number of domain networks, although our primary application has two domains.

We represent the i^{th} *domain network* by its adjacency matrix $\mathbf{A}^{(i)} \in \mathbb{R}_{+}^{n_i \times n_i}$, where n_i is the number of nodes in domain i. Each entry $\mathbf{A}_{xy}^{(i)}$ measures the similarity between nodes x and y in domain i. Suppose we have g domains, for any pair of $\mathbf{A}^{(i)}$ and $\mathbf{A}^{(j)}$, nodes in the two domains may be linked by an *association network* $\mathbf{B}^{(ij)} \in \mathbb{R}_{+}^{n_i \times n_j}$, with $\mathbf{B}_{xy}^{(ij)}$ measuring the weight between node x in $\mathbf{A}^{(i)}$ and node y in $\mathbf{A}^{(j)}$.

Our goal is to find clusters in each domain network, and for each cluster u in one domain, assign a relevance score to each cluster v in other domains. The score represents the relevance of cluster v to cluster u. More formally, suppose there are k_i clusters in domain network $\mathbf{A}^{(i)}$ $(1 \leq i \leq g)$, our goal is to (1) for each node x in $\mathbf{A}^{(i)}$, infer the cluster membership probabilities $P(u|x)$, which indicates the probability that node x belongs to cluster u $(1 \leq u \leq k_i)$; and (2) for any pair of domains i and j, infer the conditional probabilities $P(v|u)$ for cluster v of domain j given cluster u of domain i. Here, $P(u|x)$ is used for assigning nodes to clusters, while $P(v|u)$ is used for ranking clusters across domains.

III. THE CROSSCR ALGORITHM

Our method CROSSCR (Cross-network Clustering and Ranking) has two major components. For domain network clustering, we adopt a doubly stochastic matrix decomposition approach due to its superiority in clustering real-world sparse networks [8]. For cluster ranking, we develop a second-order random walk model to infer the cross-domain conditional probabilities of clusters. To leverage the complementary information in domain networks and association networks, we integrate the two procedures into a unified objective and optimize the two components jointly.

A. Domain Network Clustering

We employ the doubly stochastic matrix decomposition approach [8] as the basic method to cluster individual domain networks. This method has been shown to be more effective in clustering real-world sparse networks than many popular single-network clustering algorithms, such as spectral clustering and non-negative matrix factorization [8].

Suppose there are k_i clusters in domain network $\mathbf{A}^{(i)}$. Let $\mathbf{H}^{(i)} \in \mathbb{R}_{+}^{n_i \times k_i}$ be a *cluster membership matrix* with $\mathbf{H}_{xu}^{(i)} = P(u|x)$ indicating the probability that node x belongs to cluster u. A doubly stochastic approximation to the domain network $\mathbf{A}^{(i)}$ is defined by

$$\hat{\mathbf{A}}_{xy}^{(i)} = \sum_{u=1}^{k_i} \frac{\mathbf{H}_{xu}^{(i)} \mathbf{H}_{yu}^{(i)}}{\sum_{z=1}^{n_i} \mathbf{H}_{zu}^{(i)}} \tag{1}$$

where x, y and z are different node variables. Note $\hat{\mathbf{A}}^{(i)} \in \mathbb{R}_{+}^{n_i \times n_i}$ is symmetric and both of its columns and rows sum up to 1. Therefore, it is referred to as *doubly stochastic*.

The clustering problem is to infer $\mathbf{H}^{(i)}$ by minimizing the approximation error of the KL-Divergence $\mathcal{D}_{KL}(\mathbf{A}^{(i)} || \hat{\mathbf{A}}^{(i)})$. After removing some constants, this is equivalent to minimize

$$-\sum_{(x,y) \in \mathcal{E}^{(i)}} \mathbf{A}_{xy}^{(i)} \log \hat{\mathbf{A}}_{xy}^{(i)} \tag{2}$$

where $\mathcal{E}^{(i)}$ represents the set of all edges in network $\mathbf{A}^{(i)}$.

To provide control of the sparsity of $\mathbf{H}^{(i)}$, a Dirichlet prior on $\mathbf{H}^{(i)}$ can be introduced [8], which gives the following objective function for individual domain network clustering

$$\begin{aligned} \mathcal{J}_A^{(i)} &= -\sum_{(x,y) \in \mathcal{E}^{(i)}} \mathbf{A}_{xy}^{(i)} \log \hat{\mathbf{A}}_{xy}^{(i)} - (\alpha - 1) \sum_{xu} \log \mathbf{H}_{xu}^{(i)} \\ &\text{s.t. } \mathbf{H}^{(i)} \geq 0, \ \mathbf{H}^{(i)} \mathbf{1}_{k_i} = \mathbf{1}_{n_i} \end{aligned} \tag{3}$$

where α $(\alpha \geq 1)$ is a parameter in the Dirichlet distribution, $\mathbf{1}_{k_i}$ is a column vector of length k_i with all 1's. The equality constraints preserves the probabilistic interpretation of $\mathbf{H}_{xu}^{(i)}$.

B. Cross-Network Cluster Ranking

Next, we propose a second-order random walk model [12] to infer cross-domain cluster ranking scores. We first consider the case when there are two domain networks. Then we generalize our method to multiple domain networks.

Given two domain networks $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, we first augment them by two sets of *latent nodes* $\mathcal{U} = \{u\}_{u=1}^{k_1}$ and

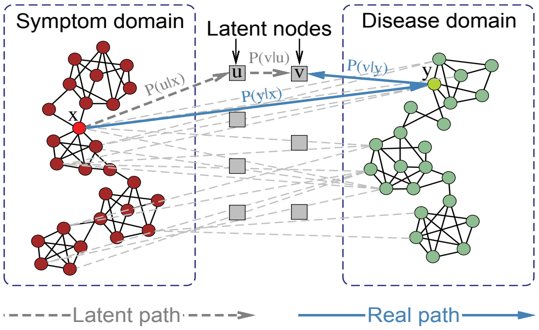


Fig. 2. An example of the augmented network.

$\mathcal{V} = \{v\}_{v=1}^{k_2}$. The latent nodes in $\mathcal{U}$ represent the hidden clusters in $\mathbf{A}^{(1)}$, and the latent nodes in $\mathcal{V}$ represent the hidden clusters in $\mathbf{A}^{(2)}$. Furthermore, every node in $\mathbf{A}^{(1)}$ is linked to every nodes in $\mathcal{U}$, every node in $\mathbf{A}^{(2)}$ is linked to every nodes in $\mathcal{V}$, and every node in $\mathcal{U}$ is linked to every nodes in $\mathcal{V}$. Thus the augmented network consists of $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$, and three complete bipartite graphs $\{\mathbf{A}^{(1)},\mathcal{U}\}$, $\{\mathbf{A}^{(2)},\mathcal{V}\}$ and $\{\mathcal{U},\mathcal{V}\}$.

Fig. 2 shows an example of the augmented network. The four squared latent nodes on the left, e.g., u, represent latent clusters of symptom domain, the three squared latent nodes on the right, e.g., v, represent latent clusters of disease domain. For clarity, the edges incident on latent nodes are omitted.

In the next, we present our method in one direction, i.e., given clusters in $\mathbf{A}^{(1)}$, ranking clusters in $\mathbf{A}^{(2)}$. The another direction can be obtained in a similar way.

Using the augmented network in Fig. 2, $P(u|x)$ can be regarded as a one-step transition probability for a random walker to jump from a node x in $\mathbf{A}^{(1)}$ to a latent node u, which can be obtained from $\mathbf{H}^{(1)}_{xu}$ by applying Eq. (3) on $\mathbf{A}^{(1)}$. Similarly, we have $P(v|y)$ from a node y to a latent cluster v in $\mathbf{A}^{(2)}$. Moreover, we have the cross-domain transition probability $P(y|x)$ from node x in $\mathbf{A}^{(1)}$ to node y in $\mathbf{A}^{(2)}$, which can be estimated by $P(y|x) = \mathbf{B}^{(12)}_{xy} / \sum_{z=1}^{n_2} \mathbf{B}^{(12)}_{xz}$.

Our goal is to estimate $P(v|u)$, which represents the importance of a cluster v in $\mathbf{A}^{(2)}$ given a cluster u in $\mathbf{A}^{(1)}$.

We observe that from a node x to a latent node v, the random walk probabilities form two kinds of second-order random walk paths, as illustrated in Fig. 2:

$$\text{(1) Real path } (x \rightsquigarrow y \rightsquigarrow v)\text{: } P_r(v|x) = \sum_{y=1}^{n_2} P(v|y)P(y|x)$$

$$\text{(2) Latent path } (x \rightsquigarrow u \rightsquigarrow v)\text{: } P_l(v|x) = \sum_{u=1}^{k_1} \underbrace{P(v|u)}_{\text{unknown}} P(u|x)$$

We refer to $x \rightsquigarrow y \rightsquigarrow v$ as a *real path*, since the bridge node y is a real node, and $x \rightsquigarrow u \rightsquigarrow v$ as a *latent path*, since the bridge node u is a latent node.

To find $P(v|u)$, we can use the latent path transition probability P_l to approximate the real path transition probability P_r. The intuition is as follows. Suppose that node x belongs to cluster u (i.e., $P(u|x)$ is large), and node y belongs to cluster v (i.e., $P(v|y)$ is large). If cluster v is important to u (i.e., $P(v|u)$ is large), then it is more likely to generate a link from node x to node y, i.e., $P(y|x)$ is large; otherwise, $P(y|x)$ is small. This generative idea inspires us to use P_l to approximate P_r. More specifically, we want to minimize the approximation error $\mathcal{D}(P_r||P_l)$ for some divergence measure $\mathcal{D}(\cdot||\cdot)$. Since we are measuring the difference between probability distributions, a natural choice is KL-Divergence $\mathcal{D}_{KL}(\cdot||\cdot)$. This gives the following loss function for all pairs of (x, v):

$$-\sum_{x=1}^{n_1}\sum_{v=1}^{k_2} P_r(v|x) \log P_l(v|x) \tag{4}$$

Formally, we define a conditional probability matrix $\mathbf{S}^{(12)} \in \mathbb{R}_+^{k_1 \times k_2}$ with $\mathbf{S}^{(12)}_{uv} = P(v|u)$. Let $\tilde{\mathbf{B}}^{(12)}$ be the row normalized version of $\mathbf{B}^{(12)}$, i.e., $\tilde{\mathbf{B}}^{(12)}_{xy} = \mathbf{B}^{(12)}_{xy} / \sum_{z=1}^{n_2} \mathbf{B}^{(12)}_{xz}$. Then it is easy to verify

$$P_r(v|x) = (\tilde{\mathbf{B}}^{(12)}\mathbf{H}^{(2)})_{xv}, \qquad P_l(v|x) = (\mathbf{H}^{(1)}\mathbf{S}^{(12)})_{xv}$$

Therefore, by enforcing a stochastic constraint on $\mathbf{S}^{(12)}$, i.e., $\mathbf{S}^{(12)} \geq 0$ and $\mathbf{S}^{(12)}\mathbf{1}_{k_2} = \mathbf{1}_{k_1}$, Eq. (4) can be rewritten in a matrix form as

$$\mathcal{J}_R^{(12)} = -\sum_{xv} \underbrace{(\tilde{\mathbf{B}}^{(12)}\mathbf{H}^{(2)})_{xv}}_{\text{real paths}} \log \underbrace{(\mathbf{H}^{(1)}\mathbf{S}^{(12)})_{xv}}_{\text{latent paths}} \tag{5}$$

which is our loss function for cross-domain cluster ranking.

C. A Unified Objective Function

A principled way to infer the clustering of nodes and the ranking of clusters is to jointly train the objective functions in Eq. (3) and Eq. (5), which allows the mutual reinforcement of the two procedures. Suppose $\mathcal{I} = \{(i,j)\}$ represents the set of all domain pairs and $\{\mathbf{B}^{(ij)}\}_{(i,j)\in\mathcal{I}}$ represents the corresponding association networks. Then, by integrating Eq. (3) and Eq. (5), and generalizing the concept to any pair of domains in $\mathcal{I}$, we reach a joint optimization problem

$$\begin{aligned} \min\ & \mathcal{J}(\{\mathbf{H}^{(i)}\},\{\mathbf{S}^{(ij)}\}) = \sum_{i=1}^{g} \mathcal{J}_A^{(i)} + \beta \sum_{(i,j)\in\mathcal{I}} \mathcal{J}_R^{(ij)} \\ \text{s.t. } & \mathbf{H}^{(i)} \geq 0,\ \mathbf{H}^{(i)}\mathbf{1}_{k_i} = \mathbf{1}_{n_i} \\ & \mathbf{S}^{(ij)} \geq 0,\ \mathbf{S}^{(ij)}\mathbf{1}_{k_j} = \mathbf{1}_{k_i},\ \forall 1 \leq i,j \leq g,\ i \neq j \end{aligned} \tag{6}$$

where β is a parameter to balance the importance between the network clustering and the cross-domain cluster ranking. When $\beta = 0$, Eq. (6) degenerates to g independent network clustering. Intuitively, the more reliable the association networks, the larger the value of β.

D. Prioritizing Nodes in Each Cluster

So far, we have considered how to order clusters in domain networks. Next, we derive a strategy to prioritize nodes within each cluster by their importances to that cluster.

Once we have obtained the cluster membership probabilities $P(u|x)$ (i.e., $\mathbf{H}^{(i)}_{xu}$) in domain i from Eq. (6), we can calculate the probability $P(x|u)$ by using the Bayes formula and expansion rule, which gives

$$P(x|u) = \frac{P(u|x)P(x)}{\sum_{z=1}^{n_i} P(u|z)P(z)} = \frac{P(u|x)}{\sum_{z=1}^{n_i} P(u|z)} \tag{7}$$

where the second equality comes from the uniform prior on nodes that is imposed by stochastic matrix decomposition.

Let $\mathbf{D}_H^{(i)}$ be a k_i-by-k_i diagonal matrix with $(\mathbf{D}_H^{(i)})_{uu} = \sum_{z=1}^{n_i} \mathbf{H}_{zu}^{(i)}$, the above equation can be rewritten by

$$P(x|u) = (\mathbf{H}^{(i)}(\mathbf{D}_H^{(i)})^{-1})_{xu} \tag{8}$$

Here, $P(x|u)$ indicates the importance of node x to cluster u in domain i. Thus, we can sort the entries in each column of $\mathbf{H}^{(i)}(\mathbf{D}_H^{(i)})^{-1}$ to obtain the most representative nodes in each cluster. In practice, showing several top ranked nodes in each cluster can help quick understanding of the category of the disease cluster and hence facilitate efficient retrieval.

E. Learning Algorithm

Since the objective function in Eq. (6) is not jointly convex in all variables, we take an alternating minimization framework that alternately solves $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{S}^{(ij)}\}$ until a stationary point is achieved. To solve $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{S}^{(ij)}\}$, we develop multiplicative updating rules with solid theories about the algorithmic convergence. For brevity, we omit the details here.

IV. Related Work

The existing computational methods for medical diagnosis are mostly based on the bipartite graph model [2], [13], [3]. As discussed before, these methods ignore the relationships between diseases and symptoms, and usually output a list of diseases mixed from different categories, which makes their results sub-optimal in practice.

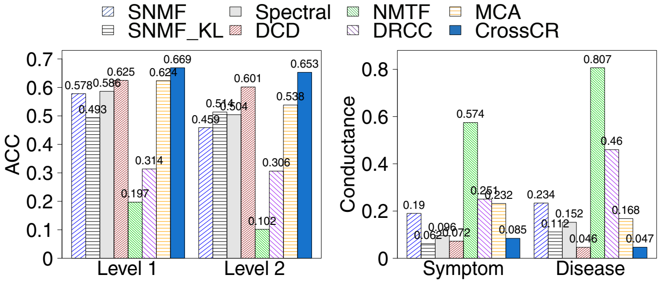
The proposed method is also related to multi-network clustering and co-clustering methods. The goal of most multi-network clustering approaches is to improve clustering accuracy by exploring the shared clustering structure in different networks [14], [15], [16]. However, these methods do not consider the relationships between clusters from different networks thus cannot handle the cluster ranking problem. Co-clustering methods [9], [10], [11] can be applied to partition rows and columns of the adjacency matrix of an association network. It has been shown that using domain networks to regularize the co-clustering can improve accuracy [11]. However, when the association network is sparse, the effectiveness of graph regularization becomes limited. In contrast, our method performs clustering directly on domain networks, and explicitly models the conditional probabilities between clusters from different domains, which is both intuitive and theoretically sound. Our experimental results also shows the clear advantage of our method over these methods.

V. Experiments

In this section, we perform extensive experiments to evaluate our method using real-world datasets.

A. The State-of-the-Art Methods

We compare CrossCR with the state-of-the-art single network clustering methods and (network-regularized) co-clustering methods. The single network clustering methods include (1) SNMF (symmetric non-negative matrix factorization using Euclidean distance [7]), (2) SNMF_KL (SNMF using KL-Divergence [17]), (3) Spectral (spectral clustering [6]), and (4) DCD (stochastic matrix decomposition approach [8]). For these methods, we adopt a two-step strategy to generate cluster ranking scores. Suppose that there are two domain networks $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. We first apply each method to cluster $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ individually. Then, we count the number of associations between every cluster in $\mathbf{A}^{(1)}$ and every cluster in $\mathbf{A}^{(2)}$, and use the row normalized counts as the ranking scores. Comparing with these single network clustering methods can demonstrate the importance of joint cluster finding and ranking.



(a) Level 1 and level 2 accuracy comparison of disease clusters

(b) Conductance comparison of disease clusters and symptom clusters

Fig. 3. The clustering performance comparison of different methods on the symptom-disease network.

The co-clustering methods include (1) NMTF [18], (2) DRCC [9], and (3) MCA [11]. NMTF is a basic matrix tri-factorization approach that can only be applied on the association network. The entries of its inferred middle factor matrix can be regarded as the relationship strengths between row and column clusters, but there is no clear theoretical basis for such meanings. DRCC and MCA are both graph regularized NMTF approaches. The major difference is that MCA generates non-negative cluster-level relationships while DRCC does not have this constraint.

B. Evaluation on Symptom-Disease Network

Dataset Description. The symptom-disease network dataset is collected from the largest medical website in China (http://www.xywy.com/). It contains a disease similarity network of $9,721$ disease nodes and $29,332$ edges, a symptom similarity network of $5,093$ symptom nodes and $22,548$ edges, as well as an association network with $5,337$ symptom-disease associations. The disease (symptom) similarity is calculated by the cosine similarity between a pair of vectorial representations of diseases (symptoms). The vectorial representations are trained by the w2v model [19] using 60 million medical Q&A descriptions about diseases and symptoms on the website. The association between a symptom and a disease is calculated based on their co-occurrence in the Q&A texts.

Clustering Evaluation. First, we evaluate the clustering performance of the selected methods. The ground truth class labels of the diseases are collected from WHO ICD-10 (In-

TABLE I

TOP RANKED DISEASE CLUSTERS GIVEN BY CROSSCR.

symptom cluster	1st disease cluster (probability)	2nd disease cluster (probability)	3rd disease cluster (probability)
(1) expectoration sore throat cough	cold varicose veins (0.6424) upper respiratory tract infection	bronchopneumonia bronchitis (0.2937) asthma	– (<0.1000)
(2) bloating burp stomachache	reflux esophagitis gastroesophageal reflux (0.7877) gastritis	duodenal inflammation antral erosion (0.1351) superficial gastritis	– (<0.1000)
(3) blurred vision dry eyes eyestrain	conjunctivitis keratitis (0.6153) pink eye	macular degeneration retinal detachment (0.1741) vitreous opacities	– (<0.1000)
(4) eye fissure photophobia pupillary block	macular degeneration retinal detachment (0.5002) vitreous opacities	amblyopia hyperopia (0.1569) esotropia	conjunctivitis keratitis (0.1330) pink eye
(5) cerebral hemorrhage intracranial hemorrhage increased intracranial pressure	cerebral infarction brainstem infarction (0.5400) stroke	skull fracture epidural hematoma (0.1449) brain contusion	diabetes hypertension (0.1161) dyslipidemia

ternational Classification of Diseases) data[1]. There is a two-level hierarchy of disease categories. Level-1 is more general than level-2. Level-1 has 17 categories and covers 1447 diseases (14.89%), Level-2 has 47 categories and covers 958 diseases (9.85%). There are no ground truth labels available for symptoms in this dataset.

Since only partial labels are available, we use purity accuracy (ACC) as the evaluation metric to evaluate the disease clustering results, instead of using normalized mutual information (NMI), which requires fully labeled data.

We also adopt the widely used conductance as a quality measure to evaluate both disease and symptom clusters. Conductance is independent of any ground truth and can be used to evaluate whether a set of nodes shows a cluster-like structure in a network. It is defined as

$$\mathrm{Cond}(\mathcal{C}) = \frac{|\partial(\mathcal{C})|}{\min\left(\mathrm{Vol}(\mathcal{C}), \mathrm{Vol}(\bar{\mathcal{C}})\right)} \tag{9}$$

where $\mathcal{C}$ is a set of nodes, $|\partial(\mathcal{C})|$ is the number of edges with one endpoint inside of $\mathcal{C}$ and another outside of $\mathcal{C}$, $\mathrm{Vol}(\mathcal{C})$ is the sum of node degrees in $\mathcal{C}$, and $\bar{\mathcal{C}}$ is the set of nodes outside $\mathcal{C}$. Usually, a lower conductance implies a better cluster structure.

Fig. 3 shows the clustering accuracy comparison in terms of the level-1 and level-2 ground truths, as well as the conductance comparison. From Fig. 3(a), we can see that the methods integrating both domain networks and their associations, such as MCA and CrossCR, outperform their single network counterparts. This confirms the complementary interaction between the two domain networks. Moreover, the best accuracy of CrossCR demonstrates the importance to directly modeling the clustering structures of domain networks rather than using them as regularizers. It also validates the importance to correctly modeling the many-to-many and ordered cluster relationships across domains. Fig. 3(b) shows the conductances of the detected clusters. From the results, we can see that SNMF_KL, DCD and CrossCR achieve the best conductances. This indicates the clusters detected by them are more cluster-like than other methods. It should be noted that although SNMF_KL and DCD are comparable with CrossCR in terms of conductance, which are very small with little room to improve, their purity accuracies are lower than CrossCR. This means the clusters detected by CrossCR make more sense in both medical context and network topology than other competing approaches.

TABLE II

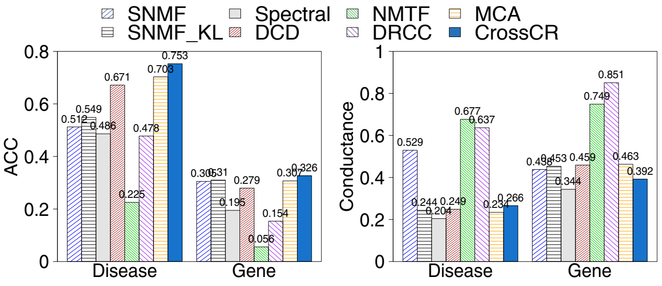
TOP RANKED DISEASES GIVEN BY A QMR-DT ALGORITHM.

Symptom #	Top ranked diseases
(1)	***tonsillitis***, ***cold***, ***asthma***, heart disease, fracture
(2)	***gastritis***, cold, heart disease, fracture, epilepsy
(3)	***dry eye***, ***conjunctivitis***, diabetes, cold, heart disease
(4)	***cataract***, ***uveitis***, ***ocular trauma***, ***keratitis***, ***pink eye***
(5)	***subarachnoid hemorrhage***, ***aneurysm***, ***hypertension***, cold

Ranking Evaluation. Next, we examine the disease cluster ranking lists generated by CrossCR. We select the symptoms and diseases that are relatively common in our population so that the general audience without background in medicine can still see their relationships. Table I shows the top 3 disease clusters for 5 symptom clusters. The disease clusters are sorted in descending order by their probabilities (i.e., $\mathbf{S}_{uv}$ in Eq. (6)), which are shown in the parentheses. Each symptom (or disease) cluster is represented by its top 3 representative symptoms (or diseases), as discussed in Sec. III-D. The disease clusters with probabilities less than 0.1 are filtered out. From the table, we can make several key observations. First, the diseases in the same cluster are from the same disease category. Second, the top ranked disease clusters are valid candidates for the symptoms. Third, the gap between the probability of the top-ranked disease cluster and those of the remaining ones is large. These properties are highly desirable and demonstrate the importance of simultaneous cluster finding and ranking.

For comparison of the ranking lists, we also apply a well-known QMR-DT based medical diagnosis algorithm, quickscore [2] on the same dataset. This algorithm only runs on the symptom-disease association network and returns a single ranking list of mixed diseases. Table II shows its top ranked diseases for the same set of symptoms as in Table I. We

[1] http://www.cdc.gov/nchs/icd/icd10cm.htm



(a) Accuracy comparison

(b) Conductance comparison

Fig. 4. The clustering performance comparison on the disease-gene network.

can see that the lists in Table II mix diseases from different categories, such as cold and asthma. This makes the results hard to interpret and limits their usefulness. Moreover, in Table II, we have highlighted relevant diseases in each line using the bold italics font. As can be seen, only a part of the diseases in each list are relevant to the corresponding symptoms. For example, heart disease and fracture are clearly irrelevant to the symptoms in (1) and (2). Such false inferences are caused by the limitation of only using the association network. This demonstrates the importance of integrating disease and symptom domain networks when performing diagnosis.

C. Additional Application of CrossCR

Although CrossCR is motivated from the application of automated medical diagnosis, it can also be applied to other problem settings. In this section, we further evaluate its performance on a disease-gene dataset [20].

The disease-gene dataset consists of a disease network and a gene network. The disease network has $5,080$ nodes and $19,729$ edges. Each node represents a specific disease phenotype and an edge signifies the similarity between two diseases according to their co-occurrences in the clinical synopsis in OMIM records [20]. In the gene network, a node is a gene and an edge indicates a functional interaction between a pair of proteins transcribed from the genes. There are $8,503$ nodes and $32,189$ edges in this gene network. In addition, diseases and genes from the two domains are connected by $2,107$ disease-gene associations. In the dataset, there are 20 disease classes which cover 691 diseases (13.60%) in the disease domain. For the genes, there are 200 pathway labels (i.e., class labels), covering $2,615$ genes (30.75%) in the network.

Fig. 4 shows the clustering results on the disease-gene network dataset. We can see that all algorithms achieve higher accuracies in the disease domain than the gene domain. This is because there are less classes in disease domain than in gene domain, resulting in purer labeled clusters. Similar to the results on the symptom-disease network, methods integrating both domain networks and their associations outperform their single network counterparts. CrossCR achieves the highest accuracy among all methods, while obtains competitive conductances in both domains. Therefore, CrossCR is better than other compared methods considering the domain knowledge and network topology together.

VI. Conclusion

Developing computational methods for medical diagnosis is an important data mining problem. Traditional diagnosis algorithms often assume no direct dependency exists between diseases (or symptoms), making them suffer from severe information loss. To address this limitation, we introduce two domain networks to model the relationships between diseases and those between symptoms. To improve the interpretability of the diagnostic outcomes, we further study a novel cross-domain cluster ranking problem. In contrast to output a single ranking list of mixed diseases, we develop CrossCR that allows a clustered structure in the ranking list so that locating diseases can be effective. Our method employs a joint learning scheme to reinforce both procedures of cluster finding and cluster ranking. Experimental results on real-life datasets demonstrate the effectiveness of CrossCR.

Acknowledgment

This work was partially supported by the National Science Foundation grants IIS-1664629 and CAREER.

References

[1] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, “Human symptoms–disease network,” *Nat. Commun.*, vol. 5, 2014.

[2] D. Heckerman, “A tractable inference algorithm for diagnosing multiple diseases,” in *UAI*, 1989.

[3] D.-I. Curiac, G. Vasile, O. Banias, C. Volosencu, and A. Albu, “Bayesian network model for diagnosis of psychiatric diseases,” in *ITI*, 2009.

[4] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, “The human disease network,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 21, pp. 8685–8690, 2007.

[5] W. Wang, S. Yang, X. Zhang, and J. Li, “Drug repositioning by integrating target information through a heterogeneous network model,” *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[6] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[7] C. H. Ding, X. He, and H. D. Simon, “On the equivalence of nonnegative matrix factorization and spectral clustering.” in *SDM*, 2005.

[8] Z. Yang and E. Oja, “Clustering by low-rank doubly stochastic matrix decomposition,” in *ICML*, 2012.

[9] Q. Gu and J. Zhou, “Co-clustering on manifolds,” in *KDD*, 2009.

[10] L. Du and Y.-D. Shen, “Towards robust co-clustering.” in *IJCAI*, 2013.

[11] R. Liu, W. Cheng, H. Tong, W. Wang, and X. Zhang, “Robust multi-network clustering via joint cross-domain cluster alignment,” in *ICDM*, 2015.

[12] H. Tong, C. Faloutsos, and J. Y. Pan, “Fast random walk with restart and its applications,” in *ICDM*, 2006.

[13] T. S. Jaakkola and M. I. Jordan, “Variational probabilistic inference and the qmr-dt network,” *J. Artif. Intell. Res.*, pp. 291–322, 1999.

[14] A. Kumar, P. Rai, and H. Daume, “Co-regularized multi-view spectral clustering,” in *NIPS*, 2011.

[15] J. Ni, H. Tong, W. Fan, and X. Zhang, “Flexible and robust multi-network clustering,” in *KDD*, 2015.

[16] J. Ni, W. Cheng, W. Fan, and X. Zhang, “Self-grouping multi-network clustering,” in *ICDM*, 2016.

[17] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2001.

[18] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal nonnegative matrix t-factorizations for clustering,” in *KDD*, 2006.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013.

[20] T. Hwang, G. Atluri, M. Xie, S. Dey, C. Hong, V. Kumar, and R. Kuang, “Co-clustering phenome–genome for phenotype classification and disease gene discovery,” *Nucleic Acids Res.*, vol. 40, no. 19, pp. e146–e146, 2012.