Multi-Information Source HIN for Medical Concept Embedding

Yuwei Cao[1], Hao Peng[2,3], and Philip S. Yu[1]

[1] Department of Computer Science, University of Illinois at Chicago,
851 S. Morgan Street, Chicago, IL 60607-7053
{ycao43,psyu}@uic.edu
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing,
Beihang University, Beijing 100083, China
[3] School of Cyber Science and Technology, Beihang University, Beijing 100083, China
penghao@act.buaa.edu.cn

Abstract. Learning low-dimensional representations for medical concepts is of great importance in improving public healthcare applications such as computer-aided diagnosis systems. Existing methods rely on Electronic Health Records (EHR) as their only information source and do not make use of abundant available external medical knowledge, and therefore they ignore the correlations between medical concepts. To address this issue, we propose a novel multi-information source Heterogeneous Information Network (HIN) to model EHR while incorporating external medical knowledge including ICD-9-CM and MeSH for an enriched network schema. Our model is well aware of the structure of EHR as well as the correlations between medical concepts it refers to, and learns semantically reflective medical concept embeddings. In experiments, our model outperforms unsupervised baselines in a variety of medical data mining tasks.

Keywords: Heterogeneous Information Network · Medical Concept Embeddings · Electronic Health Records · Multi-Information Source.

1 Introduction

Analogous to how word embedding [17, 18] empowers natural language processing (NLP) [13], medical concepts embedding is indispensable for machine learning to show its enormous potential in healthcare [1]. Embeddings of medical concepts enable the studies of correlations between concepts, such as co-occurrence of diagnosis and symptoms, and they can also be used as features to predict future events of interest [3]. One such example is computer-aided diagnosis systems, which can liberate clinicians from analyzing complex, enormous information [10].

The abundant Electronic Health Records (EHR) datasets nowadays provide a great information source for medical embedding learning. EHR datasets are often organized by admissions, and contain detailed documentation of patients' diagnostic and treatment information, including demographic characteristics, symptoms, laboratory test results, diagnoses, and medications. EHR datasets also present unique challenges. On the one hand, missing values are commonly

seen [2]. On the other hand, EHR datasets are high-dimensional and have complex structure, which often involves tens of thousands of medical concepts.

Prior studies apply different methods for medical feature extraction. Handcrafted feature engineering approaches [25, 23] are labor-intensive and also require extensive clinical expertise. The performances of Knowledge Graph Embedding (KGE) based methods [27] are greatly limited by the unbalance and sparsity of EHR [14]. Homogeneous skip-gram based models [3–5] that consider co-occurrence of medical concepts, on the other hand, treat all types of medical concepts equally, and miss the structural information of EHR [10]. Heterogeneous Information Networks (HIN) [8, 24] based models such as HeteroMed [10], though introducing heterogeneity, contain insufficient correlations since they extract edges only from EHR. EHR datasets have limitations. For example, *Congestive heart failure* and *Systolic heart failure* are sub-types of *Heart failure*, which in turn belongs to the *Heart Disease* hierarchy, but EHR datasets do not contain these relations. Existing methods rely on EHR as their only information source, and thus are unaware of the correlations between medical concepts.

To supplement such shortage, we propose a novel multi-information source HIN to model EHR while incorporating external medical knowledge including The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) [19] and Medical Subject Headings (MeSH) [20]. We first preprocess data and extract medical concepts from EHR. These concepts, along with patients, are nodes in our HIN. We then add edges between patients and medical concepts based on their co-occurrences in EHR. Besides, we explore ICD-9-CM and MeSH for more edges. Both ICD-9-CM and MeSH contain valuable knowledge, understandings, and insights from medical experts, and reveal correlations between medical concepts. To be more specific, as *Congestive heart failure* and *Systolic heart failure* in the above example are closely correlated according to ICD-9-CM, we therefore append an edge between them to capture such correlation. Given the enriched HIN schema, we adopt the commonly-used HIN embedding technique [6] to learn medical concept embeddings.

Our work marks the following contributions:

- We propose a novel multi-information source HIN that incorporates EHR with abundant external medical knowledge including ICD-9-CM and MeSH. Our design simultaneously preserves structural information lies in EHR and correlations between medical concepts reflected by external medical databases. This work enables the learning of more semantically reflective embeddings, and eventually allows more efficient and effective medical concept analysis.
- We quantitatively show that the learned embeddings offer significant performance gains over mainstream unsupervised baselines in various medical data mining tasks, including diagnosis, procedure, symptom classification, and clustering.
- We qualitatively demonstrate by visualization the internal correlations between medical concepts of the same type, as well as across different types.

Our code is publicly available at *https://github.com/RingBDStack/MISMV/*.

2 Related Work

Medical Representation Learning. Pioneer works in medical representation learning that utilize handcrafted features [25, 23] can be traced back to the 2000s. Missing values are commonly seen in EHR, and such incompleteness is one of the leading issues [2]. Besides, feature design is laborious and requires medication expertise [10]. To deal with these, unsupervised approaches [3–5] that enlightened by word2vec [17, 18] concatenate medical concepts in admission records to form sequences, and then use the result as corpus. These studies improve and automate medical representation learning. However, they mainly explore co-occurrences and lack consideration of the complex structure of EHR [10]. By contrast, HIN based models [10] preserve the structure of EHR by modeling EHR into a HIN, and then apply heterogeneous skip-gram. Nevertheless, they are unaware of correlations between medical concepts that are absent from EHR.

Network Embedding. EHR datasets contain structured records that refer to a large set of medical concepts, and can be intuitively represented as networks. Network embedding methods [22, 26, 7] can thus be applied. These methods capture the semantics in the raw networks, and offer natural handling of missing values [10]. Compared to homogeneous ones [22, 26], HIN embedding techniques [6, 9] can jointly model structural and semantic information. This strength comes from the preservation of diverse node types and edge types. Random walks are guided by meaningful metapaths that differentiate nodes' neighbors by types so that a heterogeneous skip-gram model [6] can then be employed. HIN is therefore adopted by many recent studies [10], including our own. Enriched nodes and edges are essential. Efforts have been devoted to enriching the nodes. [10] properly explored raw text, numerical and categorical data in EHR and fully utilizes information in terms of node extraction. Its edges, however, come only from the EHR. In non-medical domains, it has been shown in [11, 21] that external information sources can reveal correlations between nodes and are worth integrated as edges to enrich the network. In this paper, as we incorporate external medical knowledge including ICD-9-CM and MeSH into network modeling, we extract edges from them for a more informative and semantically rich network.

3 The Proposed Framework

In this section, we propose the Multi-Information Source Medical Vectors (MISMV) model. We construct a multi-information source HIN, and learn medical concept embeddings from it. Figure 1 shows the MISMV framework.

3.1 Construction of Multi-Information Source HIN

As illustrated in Figure 1(a), we combine EHR and external knowledge databases, and model them into a HIN. A HIN is defined as a graph $G = (V, E)$ where V and E stand for collections of nodes (patients and medical concepts) and edges (relations) that are of various types [8]. We also construct a HIN schema, which can be viewed as a meta template of G.

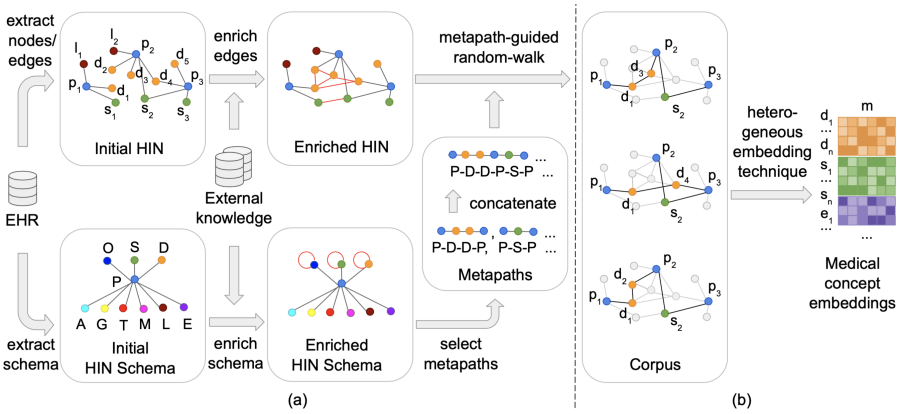

Fig. 1. The Multi-Information Source Medical Vectors (MISMV) framework. S, D, O, E, L, M, T, G, A and P are concept types, and they refer to *symptom*, *diagnosis*, *procedure*, *prescription*, *laboratory test*, *microbiology test*, *ethnicity*, *gender*, *age* and *patient*, respectively. Lowercase letters with subscripts are concepts, for example, d_1 stands for a specific diagnosis such as *Systolic heart failure*.

We model EHR into an initial HIN. EHR datasets are patient-centered, i.e. each record is related to a patient, and refers to a medical concept accompanied by a value [10]. For example, a record in EHR may be abstracted as *Hematocrit 42.4%* or *yeast grew when tested*, where the former refers to the medical concept *Hematocrit* with a value of 42.4%, while the latter refers to medical concept *yeast culture* with a value of *culture-positive*. We extract medical concepts from EHR. First, for concepts of categorical values, we either directly grab their values or reduce them into smaller categories based on their similar or identical semantics. Genders are mapped into two nodes. Ages are split into groups using threshold 15, 30 and 64 as suggested in [15]. Ethnicities are reduced into five categories, with rarely seen ones combined as *other*. Prescriptions are reduced based on constituents, for example, *Aspirin* and *Aspirin (Buffered)* are mapped into one. Procedures and diagnoses are mapped into corresponding ICD-9-CM codes. Microbiology tests with culture-positive results are mapped into the names of organisms, for example, *yeast grew when tested* in the above example is mapped into *yeast*. Secondly, fields of continuous values, too, are reduced into categories. Laboratory tests are reduced to their codes combined with flags that indicate whether or not the results are within normal ranges, as *Hematocrit 42.4%* in the above example is mapped into *Hematocrit normal*. Finally, for raw-text fields, we extract nodes by phrase mining: we conduct phrase matching between notes and vocabularies in MeSH descriptor, and use matched terms as symptoms. We use patients and extracted medical concepts as nodes, and "*refer to*" relations between them as edges to build the initial HIN, as shown in the upper-left part of Figure 1(a). We also abstract the types of nodes and edges into an initial HIN schema, as shown in the bottom-left part of Figure 1(a).

We then enrich the initial HIN by exploring selected external knowledge databases for correlations between medical concepts, and integrate these cor-

relations as new edges. Procedures and diagnoses in our HIN are encoded by ICD-9-CM, and symptoms by MeSH. Both descriptors are ordered and of tree structures, which enable us to detect correlations revealed by codes. For each diagnosis, its ICD-9-CM code is comprised of three characters to the left of a decimal point, and one or two digits to its right, where the first three characters indicate which subclass this diagnosis belongs to. For example, *410.0 Acute myocardial infarction of anterolateral wall* and *410.2 Acute myocardial infarction of inferolateral wall* are both in category *410 Acute myocardial infarction*, which is a subclass of *390-459 Disease of the circulatory system*. As identical in the first three characters implies similarity, therefore, an edge can be added between them. In this way, we examine all pairs of diagnosis nodes in our HIN and append new edges. Procedure and symptom nodes are examined likewise, except correlations between procedure nodes are based on the identity of the first two digits of ICD-9-CM codes, while symptom nodes are decided by all digits up to the last decimal point in their MeSH codes. The appended edges are highlighted in red in the enriched HIN shown in Figure 1(a). We also append "*similar to*" as a new edge type onto the HIN schema. Figure 2 shows the enriched HIN schema, where the self-loops of *symptom*, *diagnosis*, and *procedure* are made possible by external knowledge extension.

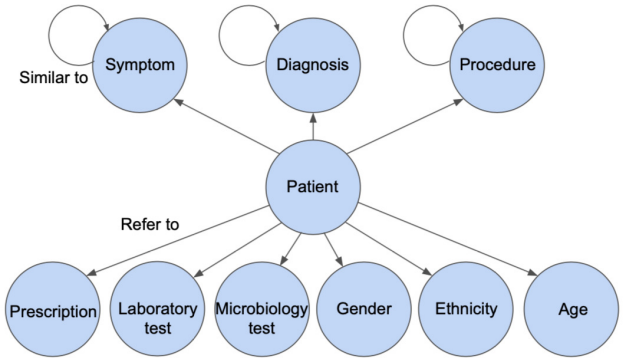

Fig. 2. Heterogeneous network schema

We derive semantically meaningful metapaths from the HIN schema. A metapath is a path on network schema that defines relations between node types [8], and it carries semantics. For example, *patient* $\rightarrow$ *diagnosis* $\leftarrow$ *patient* implies that two patients are similar because they have the same disease diagnosis. Table 1 lists all metapaths along with their semantics, where the metapaths of length 4 are enabled by exploring external knowledge, and they integrate correlations between nodes of the same type. We use these metapaths to guide heterogeneous random-walks [6], as discussed in detail in Section 3.2.

3.2 HIN Embedding

Figure 1(b) shows how we learn embeddings from the enriched HIN. We adopt a heterogeneous network embedding technique as proposed in [6]. Note that [6]

Table 1. Metapaths extracted from network schema.

Semantics	Metapaths
two patients are related because they refer to a common medical concept	patient-age-patient, patient-gender-patient patient-ethnicity-patient, patient-symptom-patient, patient-lab test-patient, patient-micro test-patient, patient-procedure-patient, patient-diagnosis-patient, patient-prescription-patient
two patients are related because they refer to two similar medical concepts	patient-diagnosis-diagnosis-patient, patient-procedure-procedure-patient, patient-symptom-symptom-patient

uses a single metapath, in contrast, we incorporate rich semantics using multiple metapaths as listed in Table 1. In practice, since all our metapaths begin and end with *patient*, we concatenate them together repeatedly for the random-walks to keep going. Metapaths can have equal or different weights in the concatenation. Figure 1(a) shows an example where we assign both *P-D-D-P* and *P-S-P* a weight of 1, and get *P-D-D-P-S-P*.

Similar to homogeneous techniques [26, 22] inspired by word2vec [17], our embedding approach is based on local structure prediction, and aim to maximize the probability of seeing the local neighborhood of each node in the network. In addition, we further differentiate the types of neighbors by metapath-guided, heterogeneous random walks. Specifically, after a node is sampled, instead of randomly choosing the next node from its neighbors, we only choose from those of the type designated by the metapath. Figure 1(b) shows a concrete example of random walks guided by metapath *P-D-D-P-S-P*. Suppose we start from p_1, then we can only walk to d_1, as the metapath requires the type of the next node to be *D*. After then, we move on by randomly choosing one from d_2, d_3, and d_4, as they are neighbors of d_1, and also are of type D as required by the metapath. We continue in this manner, and eventually get a metapath instance such as $p_1 - d_1 - d_4 - p_3 - s_2 - p_2$, which incorporates the semantics of the metapath. Given an embedding function $f : C \mapsto \mathbb{R}^m$, where C denotes the set all medical concepts, the objective of the heterogeneous skip-gram can be formalized as:

$$\underset{f}{argmax} \sum_{c \in C} \sum_{t \in T} \sum_{n_t \in N_t(c)} \log P(n_t | f(c)), \tag{1}$$

where $f(c)$ is the embedding of medical concept c, T stands for the set of all node types, and $N_t(c)$ stands for c's neighbors of type t. $P(n_t|f(c))$ can be defined as a softmax function:

$$P(n_t | f(c)) = \frac{\exp(f(c) \cdot f(n_t))}{\sum_{v \in C} \exp(f(c) \cdot f(v))} \tag{2}$$

For efficient computation, we apply negative sampling [18], and (2) becomes:

$$P(n_t | f(c)) = \log \sigma(f(c) \cdot f(n_t)) + \sum_{m=1}^{M} \mathbb{E}_{v^m \sim P(v)} [\log \sigma(-f(c) \cdot f(v^m))], \tag{3}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$, and $P(v)$ is the pre-defined distribution from which we sample M negative nodes. In each training step, we update the embeddings of c, n_t and M sampled negative nodes by Stochastic Gradient Descent (SGD).

4 Experiments

In this section, we evaluate the medical concept embeddings learned from our multi-information source HIN. We first describe datasets and preprocessing strategies, then introduce evaluation tasks, results and analyses.

4.1 Dataset

We use patients and medical concepts contained in Medical Information Mart for Intensive Care III (MIMIC III) [12] as nodes. We use relations between patients and medical concepts in MIMIC III, as well as relations between medical concepts in ICD-9-CM and MeSH as edges. MIMIC III is a large, public EHR dataset that contains de-identified records of more than forty thousand patients. It includes patient-centered clinical records such as demographics, vital sign measurements, caregiver notes, laboratory test results, along with high-level dictionaries of codes and terminologies. Table 2 summarizes our usage of tables and fields in MIMIC III. ICD-9-CM is the official coding system of assigning codes to diagnoses and procedures used by hospitals in the United States [19], where it organizes over $14,000$ diagnoses and $3,900$ procedures into 19 and 18 clinically meaningful classes, respectively. MeSH classifies a comprehensive range of medical concepts into 16 top-level categories, and serves to facilitates article searching [20]. We utilize three top-level categories in the MeSH descriptor hierarchy, i.e. anatomy concepts, organisms, and diseases. The resulting HIN contains 64,740 nodes, including 50,865 patients, 2,007 symptoms, 990 laboratory tests, 309 microbiology tests, 1,952 prescriptions, 2,003 procedures and 6,604 diagnoses. The resulting HIN contains 7,655,615 edges. 7,575,015 are from the initial HIN, including 947,633 between patients and symptoms, 4,281,748 between patients and lab-tests, 46,477 between patients and micro-tests, 1,322,586 between patients and prescriptions, 219,829 between patients and procedures, and 604,147 between patients and diagnoses. In addition, there are 80,600 edges extracted from ICD-9-CM and MeSH, including 46,960 between diagnoses, 31,618 between procedures and 2,022 between symptoms.

Table 2. MIMIC III usage in our study.

Tables	Fields	Descriptions
patients, admissions	hadm_id, gender, admittime, dob, ethnicity	demographic information of patients
labevents	itemid, flag	laboratory test results along with flags (normal/abnormal)
noteevents	text, category (discharge summary)	raw text descriptions containing patients' symptoms
procedures_icd	icd9_code	procedures performed on patients, recorded in ICD-9-CM codes
microbiologyevents	org_itemid (not NULL)	microbiology tests with culture-positive results
prescriptions	drug	medications given to patients
diagnoses_icd	icd9_code	diagnoses recorded in ICD-9-CM codes

4.2 Experimental Setup

We evaluate our model and compare it to unsupervised baselines through classification, clustering, and visualization. All these are classic tasks that are commonly performed in representation learning studies [3, 7, 26]. All models are trained with window size set to 5, the number of negative samples to 20 and out dimension to 128. The models are as follows:

- **Med2Vec [3]**. A word2vec based multilayer neural network for medical concepts and admissions embedding.
- **Word2vec [17]**. We concatenate medical concepts referred to by each patient, and use the concatenations to train word2vec model. We experimented on two sets of word2vec embeddings: W2vRaw is trained with the entire corpus, while W2vFiltered ignores medical concepts with frequencies < 10.
- **HeteroMed [10]**. A HIN based model for medical concept embeddings.
- **MISMV (ours)**. We train our model in three variations. MISMV-D contains correlations between diagnoses, MISMV-DS further integrates correlations between symptoms, while MISMV-DSP contains correlations between diagnoses, symptoms, and procedures. We use equal weights for all metapaths as we found little difference in task results with different weights.

4.3 Medical Concept Classification

This section evaluates embeddings by multi-class classifications. We use ICD-9-CM [19] and MeSH [20] categories as the ground truths. There are 18 distinct classes for procedures, 46 for symptoms, and 19 for diagnoses. We observe the labels of certain proportions of all nodes, varying from $5-90\%$, and the task is to predict the labels of the rest nodes. We input embeddings to a *LogisticRegression* classifier, and report Macro-F1 and Micro-F1 scores.

Table 3. Multi-class procedure classification results.

Metric	Method	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Macro-F1	Med2vec	0.016	0.017	0.018	0.019	0.019	0.020	0.020	0.019	0.020	0.022
	HeteroMed	0.179	0.252	0.313	0.330	0.347	0.388	0.452	0.467	0.488	0.600
	W2vRaw	0.200	0.203	0.259	0.275	0.273	0.295	0.311	0.324	0.305	0.311
	W2vFiltered	0.204	0.220	0.267	0.334	0.352	0.410	0.444	0.426	0.374	0.417
	MISMV-D	0.269	0.335	0.352	0.449	0.452	0.479	0.487	0.495	0.488	0.520
	MISMV-DS	0.276	0.308	0.376	0.433	0.474	0.529	0.535	0.564	0.583	0.562
	MISMV-DSP	**0.672**	**0.817**	**0.903**	**0.945**	**0.976**	**0.969**	**0.977**	**0.970**	**0.980**	**0.976**
Micro-F1	Med2vec	0.170	0.186	0.184	0.187	0.193	0.188	0.183	0.176	0.190	0.199
	HeteroMed	0.390	0.471	0.518	0.532	0.555	0.558	0.560	0.570	0.588	0.601
	W2vRaw	0.486	0.488	0.518	0.530	0.518	0.527	0.544	0.551	0.546	0.537
	W2vFiltered	0.557	0.593	0.618	0.649	0.651	0.660	0.672	0.687	0.639	0.602
	MISMV-D	0.496	0.545	0.555	0.591	0.597	0.613	0.620	0.621	0.619	0.632
	MISMV-DS	0.491	0.516	0.550	0.568	0.609	0.624	0.627	0.629	0.638	0.662
	MISMV-DSP	**0.794**	**0.922**	**0.969**	**0.984**	**0.990**	**0.988**	**0.990**	**0.993**	**0.995**	**0.990**

Result analysis Tables 3, 4 and 5 show results for procedure, symptom, and diagnosis classification, respectively. Our models consistently outperform all baselines by large margins in all three categories. Take symptom classification for example, compared to the highest baseline (HeteroMed), MISMV-DSP shows 175%-313% improvements in Macro-F1 and 20%-98% gains in Micro-F1 regardless of the variation of training size. A comparison between variations

Table 4. Multi-class symptom classification results.

Metric	Method	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Macro-F1	Med2vec	0.008	0.017	0.008	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	HeteroMed	0.004	0.068	0.081	0.108	0.108	0.093	0.117	0.130	0.131	0.138
	W2vRaw	0.024	0.028	0.038	0.036	0.036	0.047	0.061	0.057	0.081	0.052
	W2vFiltered	0.026	0.037	0.039	0.059	0.064	0.085	0.080	0.077	0.062	0.058
	MISMV-D	0.029	0.050	0.070	0.083	0.081	0.086	0.096	0.111	0.108	0.131
	MISMV-DS	0.104	0.159	0.211	0.259	0.317	0.378	0.412	**0.436**	**0.435**	0.419
	MISMV-DSP	**0.121**	**0.186**	**0.265**	**0.347**	**0.349**	**0.383**	**0.412**	0.393	0.433	**0.425**
Micro-F1	Med2vec	0.247	0.246	0.250	0.248	0.249	0.240	0.238	0.234	0.224	0.259
	HeteroMed	0.229	0.223	0.224	0.242	0.247	0.233	0.243	0.268	0.269	0.305
	W2vRaw	0.225	0.214	0.225	0.219	0.215	0.233	0.254	0.249	0.276	0.284
	W2vFiltered	0.243	0.270	0.268	0.277	0.273	0.297	0.296	0.306	0.288	0.273
	MISMV-D	0.232	0.232	0.228	0.223	0.228	0.230	0.235	0.245	0.227	0.250
	MISMV-DS	**0.302**	**0.330**	0.357	0.389	0.417	0.432	0.452	0.454	0.464	0.447
	MISMV-DSP	0.276	0.318	**0.373**	**0.414**	**0.423**	**0.452**	**0.480**	**0.478**	**0.487**	**0.492**

Table 5. Multi-class diagnosis classification results.

Metric	Method	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Macro-F1	Med2vec	0.018	0.018	0.018	0.019	0.019	0.020	0.020	0.021	0.022	0.022
	HeteroMed	0.219	0.254	0.309	0.330	0.348	0.348	0.385	0.394	0.408	0.376
	W2vRaw	0.244	0.308	0.347	0.367	0.371	0.384	0.389	0.396	0.390	0.377
	W2vFiltered	0.362	0.427	0.450	0.480	0.514	0.500	0.503	0.491	0.508	0.551
	MISMV-D	**0.572**	0.653	0.721	0.745	0.766	0.769	0.819	0.822	0.822	0.819
	MISMV-DS	0.558	**0.678**	**0.734**	**0.755**	**0.772**	**0.821**	0.829	0.839	0.844	0.836
	MISMV-DSP	0.523	0.644	0.726	0.747	0.755	0.819	**0.840**	**0.841**	**0.854**	**0.856**
Micro-F1	Med2vec	0.201	0.204	0.205	0.202	0.200	0.203	0.198	0.206	0.208	0.216
	HeteroMed	0.329	0.351	0.396	0.417	0.436	0.440	0.455	0.462	0.475	0.453
	W2vRaw	0.385	0.424	0.460	0.472	0.474	0.486	0.497	0.502	0.502	0.477
	W2vFiltered	0.482	0.518	0.548	0.573	0.594	0.586	0.591	0.576	0.593	0.604
	MISMV-D	**0.634**	0.703	0.768	0.789	0.808	0.809	0.815	0.822	0.821	0.817
	MISMV-DS	0.623	**0.725**	**0.772**	**0.797**	**0.808**	0.813	0.821	0.831	0.836	0.837
	MISMV-DSP	0.599	0.698	0.766	0.791	0.799	**0.821**	**0.838**	**0.843**	**0.849**	**0.846**

of our models also shows that adding correlations between medical concepts can help improving classification results, as MISMV-DSP, integrates correlations between procedures, shows a >50% higher Macro-F1 and Micro-F1 compared to MISMV-DS and MISMV-D in procedure classification. HeteroMed outperforms W2vRaw in procedure and symptom classifications when the training set becomes large enough (>=20%), which shows metapath-guided random walks essentially preserve more information about nodes' correlations. In diagnosis classification, however, HeteroMed does not perform as well. This is because diagnoses in MIMIC III are sparse, as 3,330 out of 6,604 diagnoses are referred by <=5 patients. For diagnosis nodes that are referred by very few patients, MIMIC III alone does not provide enough structural information to fully reveal their relations with other diagnoses in the network. Our models overcome this problem through enriching the structural information, as MISMV-D shows a >100% higher Macro-F1 and a >70% higher Micro-F1 compared to HeteroMed in diagnosis classification despite variation in training size. Moreover, MISMV-DSP shows an additional ∼5% improvement in both metrics when training size >=50%. This is because MISMV-DSP indirectly introduces more paths between diagnoses into the network by integrating symptoms and procedures correlations. The same thing is true for symptom classification: compared to MISMV-DS, MISMV-DSP on average gives a ∼10% higher Macro-F1 and a ∼3% higher

Micro-F1, because MISMV-DSP indirectly introduces more paths between symptoms by appending links between procedures. As expected, by getting rid of infrequent concepts, W2vFiltered shows better results compared to W2vRaw in all three categories. Med2vec embeddings are tuned for prediction purpose [3], and turned out are not suitable for classification tasks.

4.4 Medical Concept Clustering

For medical concept clustering, we leverage the k-means algorithm and report normalized mutual information (NMI), purity score, and adjusted rand index (ARI) [16]. We also visualize the embeddings for a direct overview.

Table 6. Medical concept clustering results.

Node type	Metric	Med2vec	HeteroMed	W2vRaw	W2vFiltered	MISMV-D	MISMV-DS	MISMV-DSP
Procedure	NMI	0.050	0.228	0.251	0.394	0.405	0.405	**0.652**
	Purity	0.202	0.408	0.431	0.598	0.576	0.560	**0.697**
	ARI	0.000	0.087	0.055	0.187	0.206	0.227	**0.345**
Symptom	NMI	0.098	0.221	0.154	0.149	0.188	0.291	**0.300**
	Purity	0.265	0.294	0.248	0.305	0.279	0.338	**0.357**
	ARI	0.004	0.014	0.000	0.001	0.010	0.030	**0.037**
Diagnosis	NMI	0.031	0.205	0.221	0.307	0.380	0.409	**0.426**
	Purity	0.211	0.332	0.337	0.405	0.451	**0.474**	0.471
	ARI	0.003	0.106	0.091	0.149	0.170	0.175	**0.254**

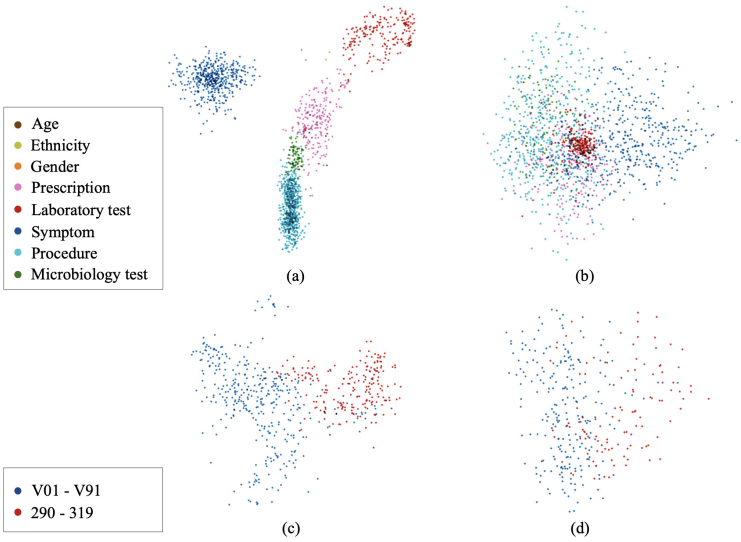

Fig. 3. Visualization of medical concept embeddings. (a) and (b) are 2D PCA projections of non-diagnosis medical concept embeddings learned by MISMV-DSP and W2vFiltered, respectively. (c) and (d) project the embeddings of all diagnoses in class *V01-V91 Mental disorders* and *290-319 Supplementary classification* learned by MISMV-DSP and W2vFiltered, respectively.

Result analysis Table 6 shows the results for procedure, symptom, and diagnosis clustering. Our models act significantly better than baselines. MISMV-DSP

shows >35%, >16% and >70% gains in NMI, purity, and ARI, respectively, compared to the best among baselines in all categories. This testifies that incorporating external knowledge can significantly improve clustering performances. A comparison between variations of our models further confirms the validity of our strategy. By integrating direct external knowledge of procedures correlations, MISMV-DSP performs >20% better in all three metrics compared to MISMV-DS and MISMV-D in procedure clustering. Indirect knowledge is also helpful. MISMV-DSP performs better than MISMV-DS in symptom clustering, because it appends edges between procedures, which indirectly creates more paths between symptom nodes. W2vFiltered outperforms other baselines because sparsity was removed. Med2vec embeddings, tailored for prediction tasks [3], did not perform as well in clustering tasks.

Figure 3 shows 2D PCA projections of medical concept embeddings learned by our MISMV-DSP model and W2vFiltered, which is the most competitive baseline in the clustering tasks. In Figure 3(a), medical concepts of different types fall into clearly separated clusters. This suggests a good capture of the structural information in EHR by MISMV-DS. In Figure 3(b), however, all concepts are in one large cluster, as W2vFiltered embeddings do not contain the structural information. Figure 3(c) and (d) zoom in on diagnoses embeddings. Figure 3(c) shows that diagnoses from two different classes are clearly separated by MISMV-DSP. This proves that correlations between diagnoses are well preserved by our model. Compared to Figure 3(c), the separation in Figure 3(d) is not as clear, since W2vFiltered learns the correlations between medical concepts only from the co-occurrences of them in the EHR.

5 Conclusion

We propose a multi-information source HIN that cooperates EHR and external knowledge including ICD-9-CM and MeSH. By integrating various information sources to enrich heterogeneous network schema, our model is well aware of both the structure of EHR and the semantics of as well as the correlations between medical concepts it refers to. The embeddings learned by us are informative and semantically reflective. In experiments, our model significantly outperforms baselines in diagnosis, procedure, symptom classification, and clustering.

Acknowledgement

The corresponding author is Hao Peng. This work is supported by NSF under grants III-1526499, III-1763325, III-1909323, and CNS-1930941.

References

1. Beam, A.L., Kohane, I.S.: Big data and machine learning in health care. Jama **319**(13), 1317–1318 (2018)
2. Botsis, T., Hartvigsen, G., Chen, F., Weng, C.: Secondary use of ehr: data quality issues and informatics opportunities. AMIA Summit on TBI, 2010

3. Choi, E., Bahadori, M.T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., Sun, J.: Multi-layer representation learning for medical concepts. In: ACM SIGKDD, 2016
4. Choi, Y., Chiu, C.Y.I., Sontag, D.: Learning low-dimensional representations of medical concepts. AMIA Jt Summits Transl Sci Proc, 2016
5. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: ACM CIKM, 2014
6. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: Scalable representation learning for heterogeneous networks. In: ACM SIGKDD, 2017
7. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: ACM SIGKDD, 2016
8. Han, J., Sun, Y., Yan, X., Yu, P.S.: Mining knowledge from databases: an information network analysis approach. In: ACM SIGMOD, 2010
9. He, Y., Song, Y., Li, J., Ji, C., Peng, J., Peng, H.: Hetespaceywalk: A heterogeneous spacey random walk for heterogeneous information network embedding. In: ACM CIKM, 2019
10. Hosseini, A., Chen, T., Wu, W., Sun, Y., Sarrafzadeh, M.: Heteromed: Heterogeneous information network for medical diagnosis. In: ACM CIKM, 2018
11. Huang, X., Song, Q., Li, J., Hu, X.: Exploring expert cognition for attributed network embedding. In: ACM WSDM, 2018
12. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
13. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML, 2014. pp. 1188–1196 (2014)
14. Liang, X., Li, D., Song, M., Madden, A., Ding, Y., Bu, Y.: Predicting biomedical relationships using the knowledge and graph embedding cascade model. PloS one **14**(6) (2019)
15. Luo, J., Eldredge, C., Cho, C.C., Cisler, R.A.: Population analysis of adverse events in different age groups using big clinical trials data. JMIR medical informatics, 2016
16. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Natural Language Engineering **16**(1), 100–103 (2010)
17. Mikolov, T., Chen, K., Corrado, G., rey Dean, J.: Efficient estimation of word representations in vector space. In: ICLR, 2013
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, 2013
19. NCHS: International classification of diseases, ninth revision, clinical modification (icd-9-cm) (2015), https://www.cdc.gov/nchs/icd/icd9cm Last accessed 1 Sep 2019
20. NLM: Medical subject headings (mesh) fact sheet (2005), https://www.ncbi.nlm.nih.gov/mesh. Last accessed 1 Sep 2019
21. Peng, H., Li, J., Gong, Q., Song, Y., Ning, Y., Lai, K., Yu, P.S.: Fine-grained event categorization with heterogeneous graph convolutional networks. In: IJCAI, 2019
22. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: ACM SIGKDD, 2014
23. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare mimic datasets. Journal of Biomedical Informatics **83**, 112–134 (2018)
24. Shi, C., Li, Y., Zhang, J., Sun, Y., Philip, S.Y.: A survey of heterogeneous information network analysis. IEEE TKDE, 2016
25. Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. IJCA, 2011
26. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: WWW, 2015
27. Zhao, C., Jiang, J., Guan, Y., Guo, X., He, B.: Emr-based medical knowledge representation and inference via markov random fields and distributed representation learning. Artificial intelligence in medicine **87**, 49–59 (2018)